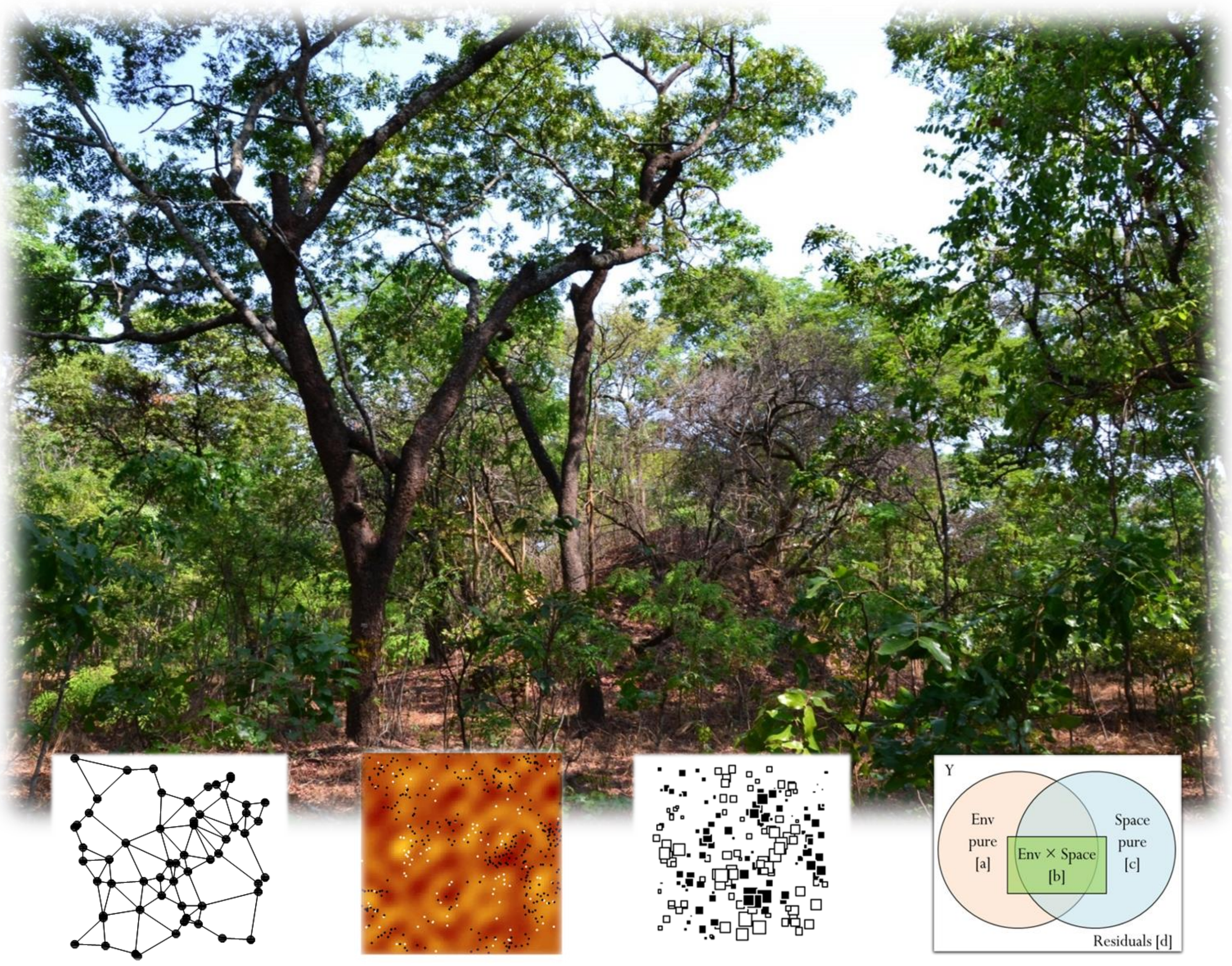
Analyses spatialement explicites des mécanismes de structuration des communautés d'arbres



Thèse de doctorat présentée en vue de l'obtention
du grade de Docteur en Sciences

David Bauman

Promoteur : Prof. Thomas Drouet
Département de Biologie des Organismes et Ecologie

Analyses spatialement explicites des mécanismes de structuration des communautés d'arbres

Thèse soumise à l'école doctorale de Biodiversité, Ecologie et Evolution en accord avec les règles du diplôme de

Docteur en Sciences de l'Université Libre de Bruxelles

par

David BAUMAN

Septembre 2018

Promoteur

Prof. Thomas DROUET
Laboratoire d'Ecologie végétale et Biogéochimie, Université Libre de Bruxelles – ULB

Composition du jury

Prof. Marius GILBERT
Service d'Epidémiologie Spatiale, Université Libre de Bruxelles – ULB

Prof. Patrick MARDULYN
Service d'Evolution Biologique et Ecologie, Université Libre de Bruxelles – ULB

Prof. Adeline FAYOLLE
TERRA Teaching and Research Center, Forest is Life, Gembloux Agro-Bio Tech (Université de Liège)

Prof. Marc DUFRÊNE
Biodiveristy and Landscape Unit, TERRA Teaching and Research Center, Gembloux Agro-Bio Tech (Université de Liège)

Cette thèse a été financée par le **F.R.S. – FNRS** (« Fonds de la Recherche Scientifique », Belgique) via un mandat d'Aspirant F.R.S. – FNRS.

La mission de terrain en République Démocratique du Congo (RDC ; septembre – novembre 2013) qui a permis la récolte d'une grande partie des données de la thèse a été financée par une bourse de *Wallonie-Brussels International* (WBI) et de l'ERAIFT (Ecole Régionale Post-Universitaire d'Aménagement et de Gestion Intégrés des Forêts et Territoires Tropicaux). C'est dans le cadre d'une collaboration entre le Laboratoire d'Ecologie végétale et Biogéochimie (ULB, Belgique) et la Faculté d'Agronomie de l'Université de Lubumbashi (Unilu, RDC) que la mission d'échantillonnage a été rendue possible. J'ai également bénéficié de données de grandes valeurs récoltées pendant plusieurs années par le Dr. Jonathan Ilunga Muledi, avec qui tant le travail de terrain que la collaboration scientifique ont toujours été un plaisir. La logistique du travail et du logement dans la forêt claire Katangaise a été assurée par Michel Anastassiou au travers de l'ASBL Mikembo. L'étude moléculaire de la communauté des champignons ectomycorhiziens échantillonnée en RDC a été rendue possible par une riche collaboration avec le Jardin Botanique de Meise, où le Dr. Olivier Raspé m'a tout particulièrement été d'une aide précieuse.
J'ai également bénéficié des infrastructures de la Gare de Treignes (ULB) pendant neuf semaines dans la forêt du Moessia.

L'aide et le soutien de ces différents acteurs ont été indispensables à la réalisation de cette thèse.

Préface

C'est passionné par ces processus cachés qui permettent l'équilibre fin entre compétition, coexistence et mécanismes d'interdépendances multiples qui caractérisent si bien les forêts que j'ai décidé de me lancer dans l'aventure qu'a été cette thèse. Rapidement est arrivée la prise de conscience qu'approcher un semblant de compréhension de ces processus requiererait inévitablement l'utilisation d'une boîte à outils méthodologique complexe. Je n'ai donc pas tardé à me donner une période de départ « méthodes » pour découvrir ces outils qui serviraient de porte d'entrée vers les processus d'assemblage des communautés des arbres.

C'était sans compter que je tomberais tout entier et pour plus longtemps que je ne l'aurais imaginé dans ce monde de statistiques en lien avec les grandes questions de l'écologie des communautés.

Ce travail souligne le rôle essentiel de l'approche méthodologique pour répondre aux questions fondamentales de l'écologie des communautés. Il met également en évidence l'importance de la communication entre écologues de terrain et statisticiens, modélisateurs et développeurs de méthodes. Le pont entre ces spécialistes n'est probablement pas encore suffisamment établi et un effort réel est encore nécessaire des deux côtés afin de faire avancer ce domaine d'étude passionnant qu'est l'écologie des communautés.

L'utilisation croissante du logiciel libre R a fortement augmenté le partage de connaissances et a rendu de plus en plus accessible des méthodes parfois très complexes grâce à la création de fonctions et packages divers. Cette plateforme recèle de nombreuses ressources, tutoriels, exemples et autres et est sans l'ombre d'un doute l'un des ponts les plus efficaces entre ces deux mondes. Toutefois, utiliser une méthode sans en comprendre les fondements et paramètres éventuels mène inévitablement à des erreurs et conclusions biaisées, lesquelles peuvent avoir un « effet boule de neige » et faire perdre au domaine scientifique concerné un temps considérable. Ma courte expérience, mes lectures, les discussions au fil des rencontres et mes propres erreurs m'ont donné la forte impression que *prendre le temps*, *se renseigner* sur les méthodes les plus à même de permettre d'aborder les questions d'intérêt, et *se donner la peine* de comprendre le fonctionnement des outils sélectionnés mènerait à une production plus lente, certes, mais de qualité et de profondeur supérieure.

Remerciements

Cette thèse n'est pas seulement le fruit de quatre années de recherche, elle est aussi (et bien heureusement) le résultat de nombreuses interactions biotiques positives (facilitation).

Avant tout, un nombre incalculable d'échanges, réflexions et (très) longues conversations avec mon promoteur de thèse. Thomas, tu t'es absolument toujours montré extrêmement disponible, tu as fait preuve d'une grande flexibilité quant à mes nouvelles idées ou redirections de recherche et tu as toujours trouvé un équilibre fin que j'ai beaucoup apprécié entre le respet d'idées nouvelles que je pouvais avoir et tes suggestions et opinions à leur propos. Nos nombreuses conversations et réflexions n'étaient pas seulement riches, elles ont rapidement été accompagnées d'une belle complicité (et de mauvais jeux de mots, en ce qui me concerne). Je te suis infiniment reconnaissant pour ces années intenses de partage et d'apprentissage.

Je voudrais également exprimer une reconnaissance toute particulière à Stéphane. Nos collaborations et « conversations spatiales » ont été des moments particulièrement riches et agréables de ma thèse, que j'espère sincèrement poursuivre (les collaborations, pas la thèse).

Olivier, si nos conversations ont été relativement irrégulières au cours de ces quatre années de thèse, elles se trouvent néanmoins parmi les moments les plus stimulants de réflexion que j'ai eus sur le monde de l'écologie des communautés, ses limites et défis à venir. Merci pour ces moments spéciaux.

Pierre… Professeur… Vraiment… Faire partie du labo a tout simplement été un plaisir du début à la fin. Merci pour le soutien et la confiance que tu m'as témoignés au cours de ces années. En tout cas…

Papa Guillaume (S*****), papa Arnaud (J****), mes très chers accolites… Que dire… Que cette thèse n'aurait pas été la moitié de ce qu'elle a été si vous aviez été à moitié moins là ? Que j'eus été de moitié plus las (et de moitié moins là) si vous n'eussiez été là ? Que grâce à vous, cette thèse a été au bonheur d'aller travailler ce que la graine digérée est à l'endozoochorie ?
Je pense pouvoir affirmer qu'aucune des trois hypothèses nulles associées aux hypothèses alternatives correspondantes et susmentionnées ne puissent ne pas être catégoriquement rejetée à un seuil de significativité de 0.05, et ce même après application d'une correction de Bonferroni pour tests multiples (c'est dire…). Merci pour ces innombrables moments de complicité !

Bast (Pat pour les intimes), ta présence au cours de ces quatre années a eu beau suivre à la perfection un pattern d'autocorrélation temporelle de large échelle, la *P*-value de la variable indépendante que tu as été sur la variable réponse qu'a été ma bonne humeur au boulot s'est révélée < 0.001 (avec un coefficient positif, *of course*). J'en profite pour remercier Monique, voisine en or et membre de la *dream team* du bureau, dont la bonne humeur légendaire et les mille-et-une merveilles culinaires faites maison ont accompagné la première moitié de nos thèses.

Jason, entre les *cenotes* du Yucatán, les nuages de moustiques des plages mexicaines, les heures interminables à se creuser la tête sur R à la recherche d'innombrables bugs tapis dans les forêts denses de nos scripts, ainsi que d'autres épisodes « juteux » (que la bienséance m'interdit de nommer ici), nos interactions n'ont jamais donné que d'agréables et enrichissantes collaborations scientifiques ou de beaux moments d'amitié. Il semble, néanmoins, que le plaisir de travailler ensemble ait été hautement positivement corrélé avec les rejets inopinnés d'articles en passe d'être acceptés (bien qu'un nombre de réplicas – heureusement – trop faible rende cette relation non significative).

Soph, merci pour ton amitié, ton écoute et ta franchise à toute épreuve tant dans les hauts que dans les bas, non seulement de cette thèse mais d'autres aspects de ma vie.

Kristel, Barbara et Mister Vanbaekel, un grand merci à tous les trois pour l'aide hautement appréciée que vous avez chacun apportée à plusieurs moments de ma thèse. Kristel, je pense qu'il va être temps qu'on retourne chercher nos *litter bags* oubliés…

Dulce linda, merci du fond du cœur pour le soutien et la patience dont tu as fait preuve au fil de toutes ces années ! Tes sourires et éclats de rire, ta présence, ton écoute et tes mots d'encouragement dans les moments (trop) intenses ont été une véritable source d'énergie (renouvelable qui plus est !). Merci, tout simplement, d'être la personne que tu es.

Alors que la fin de ces remerciements approche dangeureusement, croît inévitablement en moi une voix que je ne connais que trop bien et qui me dit (ou me répète ?..), avec un accent yiddish prononcé : « Et ta mère ?! Tu as pensé à ta mère ? Tu crois que ça a été facile de te mettre au monde ? Sept heures que j'étais occupée et tu n'arrivais toujours pas ! Et maintenant, ça y est, Docteur Bauman et tu ne penses plus à remercier ta mère ?! ».

Afin donc d'éviter à chacun un incident diplomatique le jour de la défense publique : merci, mère !

Résumé

La compréhension des processus écologiques qui sous-tendent l'assemblage des communautés végétales et la coexistence des espèces est un objectif central en écologie. Ces processus sont potentiellement nombreux et de natures contrastées. Ainsi, la composition d'une communauté de plantes dépend de processus déterministes liés aux conditions environnementales abiotiques (climat, conditions physiques et chimiques du sol, lumière) et d'interactions biotiques complexes, positives (facilitation, symbioses) comme négatives (compétition, prédation, pathogènes). En outre, les communautés sont influencées par des processus stochastiques (capacité de dispersion limitée, dérive écologique). Si les mécanismes à l'origine de ces processus sont très différents, ils ont néanmoins en commun la génération de motifs (*patterns*) spatiaux de distribution d'espèces dans les communautés. L'analyse de la structure spatiale des communautés permet ainsi une étude indirecte des processus régissant les communautés.

La nature complexe de ces patterns spatiaux a mené au développement de nombreuses méthodes statistiques de détection et de description de patterns. Les méthodes basées sur des vecteurs propres spatiaux sont parmi les plus puissantes et précises pour détecter des patterns complexes et multi-échelles. Ces vecteurs propres, utilisés comme prédicteurs spatiaux, peuvent être combinés à un ensemble de variables environnementales dans un cadre de partition de variation. Celui-ci permet, en théorie, de démêler les effets uniques et l'effet conjoint des variables environnementales et spatiales sur la variation de composition d'une communauté. Il mène ainsi à une quantification de l'action des processus déterministes et des processus stochastiques sur l'assemblage de la communauté.

Néanmoins, je montre dans cette thèse qu'un certain flou méthodologique concernant deux étapes déterminantes des analyses basées sur les vecteurs propres spatiaux a mené une proportion élevée d'études à utiliser ces méthodes de manière sous-optimale, voire fortement biaisée. Ceci compromet la fiabilité des patterns spatiaux détectés et des processus écologiques inférés. Une autre limitation de ce cadre d'analyse concerne la fraction de la partition de variation décrivant l'effet environnemental spatialement structurés qu'aucune méthode ne permet de tester.

Cette thèse présente des solutions non biaisées, puissantes et précises à ces différentes limitations méthodologiques et permet d'élargir le cadre de l'inférence de processus écologique à partir de patterns spatiaux de communautés. Les différentes étapes d'amélioration de ces méthodes ont également été illustrées dans la thèse au travers de trois cas d'études fournis par deux communautés d'arbres tropicale et tempérée et une communauté de champignons symbiotiques des arbres.

Summary

Understanding the ecological processes that underlie plant community assembly and species coexistence is a central goal in ecology. These processes are potentially numerous and of constrasting nature. Indeed, the species composition of a plant community depends upon deterministic processes related to abiotic environmental variables (climate, physical and chemical soil properties, light availability) and biotic interactions, both positive (facilitation, symbioses) and negative (competition, predation, pathogens). In addition, plant communities are driven by stochastic processes (limited dispersion ability, ecological drift). While these processes are very different and sometimes difficult or impossible to measure directly, they all have in common the generation of spatial patterns of species distribution in the community. Community spatial structures therefore allow indirectly studying the processes that drive community assembly.

The intrinsic complexity of these spatial patterns has motivated the development of numerous statistical methods aiming at the detection and description of spatial structures in communities. Spatial eigenvector-based methods are among the most powerful and accurate methods to detect complex multiscale spatial patterns. These eigenvectors can be used as spatial predictors in combination with a set of relevant environmental variables in a framework of variation partitioning. The latter theoretically allows disentangling the unique contributions and the shared contribution of environmental and spatial variables to the variation of community species composition. The variation partitioning therefore yields a quantification of the relative relevance of deterministic and stochastic processes on the community assembly.

In this thesis, I show that a methodological vagueness regarding the way two crucial steps of spatial eigenvector-based methods should be handled yielded a high proportion of studies to use these analyses either in an underpowered or in a highly biased manner. This has jeopardised the reliability of the spatial patterns detected and of the underlying processes inferred. Another limitation of this analytical framework concerns the absence of a statistical test for the fraction of the variation partitioning expected to describe the effect of a spatially structured deterministic process on the community. The lack of a test for this fraction has hindered a reliable interpretation of this fundamental process of community assembly.

This thesis presents unbiased, powerful, and accurate solutions to these different methodological limitations and allows expanding the framework using community spatial patterns as proxies of ecological processes. The different steps of methodological improvements were also illustrated in the thesis through three case studies provided by two tropical and temperate tree communities and one symbiotic fungal community associated to trees.

Table des matières

Introduction

I. L'écologie des communautés

L'écologie des communautés a pour objectif la compréhension systémique des processus écologiques sous-tendant l'**assemblage des communautés** et le maintien de la diversité du vivant. Le sens le plus commun du maintien de la diversité est la coexistence d'espèces présentant une écologie similaire (Chesson 2000). Ces espèces forment la « communauté ». Elles font partie d'un même niveau trophique et peuvent être définies comme appartenant à la même guilde (elles exploitent la même classe de ressources environnementales d'une façon similaire, Simberloff et Tamar 1991). Un deuxième aspect du maintien de la diversité fait référence au maintien d'une richesse ou équitabilité spécifique sur de longues échelles de temps. Cette définition du maintien de la diversité nécessite alors de considérer des processus particuliers tels que les taux de spéciation et d'extinction, par exemple (Hubbell 1997). C'est sur la première définition et donc sur l'assemblage de la communauté et la coexistence des espèces – principalement d'arbres – que le travail portera.

L'assemblage d'une communauté peut être appréhendé par le biais d'une métaphore de filtres (Fig. 1) (Kraft et al. 2015). Une communauté d'espèces en un endroit donné (pool local) serait le résultat d'une série de filtres ne permettant l'établissement et le maintien que des espèces présentant certains traits fonctionnels. Un **trait fonctionnel** (TF) est un trait morphologique (par ex. surface foliaire, masse d'une propagule), physiologique (par ex. composés de défense, type de photosynthèse), ou phénologique (par ex. période de floraison) à valeur adaptative et mesurable à l'échelle de l'individu (Violle et al. 2007). Ces TF se répartissent principalement sur trois axes complémentaires, à savoir, les traits de croissance, de résistance aux contraintes de l'habitat et de reproduction (Garnier et al. 2016). Ainsi, un premier filtre de dispersion sélectionne à partir du pool des espèces régionales le pool des espèces suffisamment proches ou pouvant se disperser suffisamment loin (migration pour les animaux, graines pour les plantes) que pour coloniser la zone considérée. Ensuite, un **filtre environnemental abiotique** (par ex. édaphique, climatique, régime de perturbation) empêche l'installation des espèces pour lesquelles l'environnement abiotique n'est pas viable. Finalement, un **filtre biotique** ne laisse persister que les espèces pour lesquelles le jeu des compétitions intra- et interspécifiques ainsi que d'autres types d'interactions biotiques (par ex. facilitation, symbioses) mène à une coexistence stable dans la communauté. Les filtres environnemental et biotique sont encore parfois repris ensemble sous le terme de **filtre de l'habitat** (Kraft et al. 2015, Muledi et al. 2017). L'étude de l'espace des TF d'une communauté locale est ainsi un moyen de tester l'effet des différents filtres, dès lors qu'ils auraient chacun une signature propre, le filtre abiotique menant à une convergence de certains TF (Weiher et Keddy 1995, Grime 2006) et le filtre biotique menant à une limitation de la similarité des TF, générant ainsi une distribution divergente (Fig. 1) (MacArthur et Levins 1967). L'interprétation de l'espace des TF n'est néanmoins pas aussi simple dans certains cas, dès lors qu'une compétition élevée peut être telle que certaines espèces sont exclues et que les TF des espèces restantes convergent (égalisation de la valeur sélective) (Chesson 2000, Grime 2006). D'autre part, il a été montré que de fortes contraintes abiotiques pouvaient mener à une divergence fonctionnelle (Cornwell et Ackerly 2009). Les TF de la communauté seront abordés dans les chapitres II et VI.

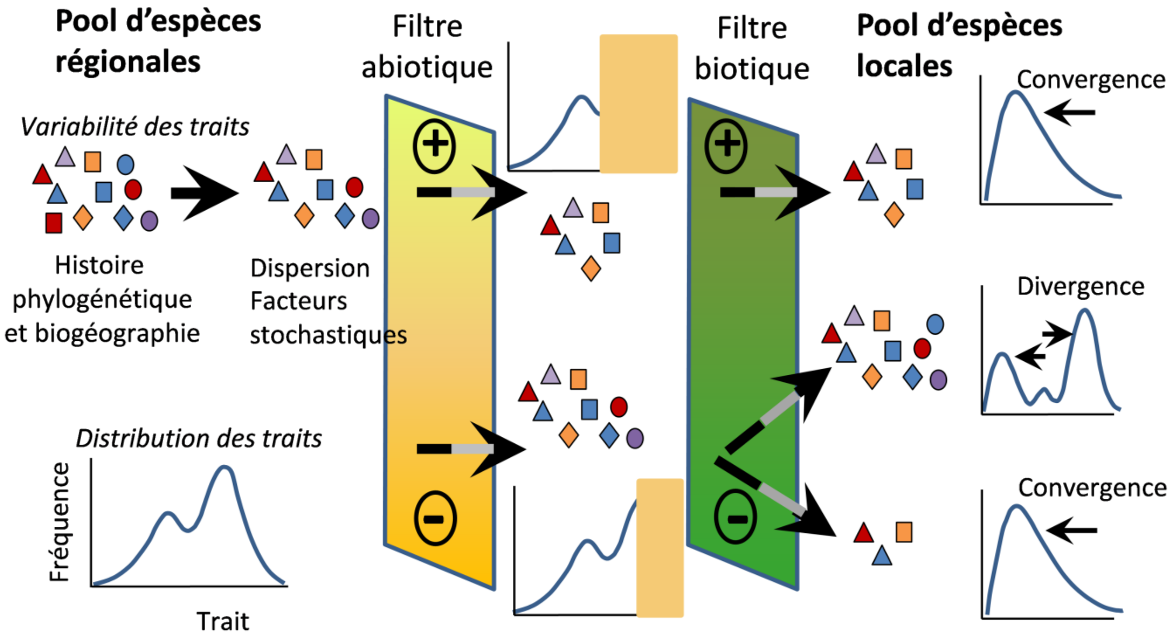


Figure 1 : Formation d'une communauté locale à partir d'un pool régional d'espèces par le passage au travers des filtres de dispersion, abiotique et biotique. Les filtres sont forts (+) ou faibles (-) et agissent sur les TF des espèces. Les rectangles orange masquent l'espace du trait filtré par les facteurs abiotiques. Les formes des symboles : valeurs de traits ; couleurs : espèces. Seules les espèces présentant des TF adaptés aux différents filtres composent la communauté locale.

Différentes théories écologiques ont été avancées pour expliquer la coexistence des espèces au sein d'une communauté. La **théorie de la niche** (Whittaker 1956, Bray et Curtis 1957, Hutchinson 1959) avance que ce sont les conditions physico-chimiques de l'environnement qui déterminent la distribution des espèces. Cette influence extérieure peut s'exprimer de façon directe, c'est-à-dire en agissant sur la niche fondamentale des espèces (l'espace des conditions environnementales physiologiquement viables), ou de façon indirecte, c'est-à-dire en agissant sur leur niche réalisée (le sous-espace de la niche fondamentale réellement occupé en conséquence des interactions biotiques). Ces deux aspects correspondent aux filtres environnemental et biotique évoqués ci-dessus (Fig. 1). Les processus écologiques associés à la théorie de la niche sont aussi appelés **processus déterministes**.

Suite à une étude spatio-temporelle de la distribution des espèces d'arbres d'un dispositif permanent de forêt de 50 ha à *Barro Colorado Island* (BCI, Panama), Stephen P. Hubbell détermina qu'une variation aléatoire des abondances des espèces (stochasticité démographique) dans le temps et l'espace associée à une dispersion aléatoire suffisaient à rendre compte de la diversité de la communauté végétale ligneuse. Il publia alors sa **théorie neutre de la biodiversité et de la biogéographie** (Hubbell 2001). La théorie neutre considère que tous les individus des différentes espèces de la communauté sont fonctionnellement équivalents du point de vue de leurs taux démographiques (mortalité, recrutement, spéciation, etc). Les abondances des espèces de la communauté seraient le résultat de deux processus : une dérive aléatoire des abondances des espèces dans l'espace et le temps, appelée **dérive écologique** (Fig. 2a), en référence à la dérive génique de la génétique des populations, et une **dispersion limitée** des individus (Fig. 2c). S'il est évident que supposer une équivalence fonctionnelle et démographique *per capita* entre les espèces n'est qu'une approximation, l'intérêt de cette théorie est de vérifier jusqu'à quel point cette hypothèse permet de rendre compte de la diversité observée (Hubbel 2006).

Actuellement, il est communément accepté que les processus neutres et de niches agissent de concert à des échelles spatiales et temporelles souvent complémentaires dans l'assemblage et le maintien de la diversité des communautés (Fig. 2d) (Cottenie 2005, Leibold et McPeek 2006, Chase et Myers 2011, Chase 2014, Brown et al. 2016). Ainsi, Garzon-Lopez et al. (2014) ont mis en évidence que la **perception de l'importance relative des processus** neutres et de niche variait fortement en fonction de la fenêtre spatiale d'observation de la communauté (Fig. 3). Ce qui, à une échelle donnée, peut être perçu comme étant principalement le fruit de processus stochastiques (part droite de Fig. 3), apparait soudain dominé par des différences d'habitat lorsque plus d'hétérogénéité environnementale est comprise dans l'étendue de l'étude (part gauche de Fig. 3). Cette prise de conscience transforma le débat opposant ces classes de processus, dès lors que leur perception dépend grandement de l'échelle spatiale d'observation. Ceci explique également la prédominance des processus neutres détectés par Hubbell et à la base de sa théorie neutre de la biodiversité (Chase 2014) : si les processus neutres semblaient avoir le dessus à l'échelle des 50 ha du dispositif permanent (relativement homogènes du point de vue des conditions de l'habitat), Garzon-Lopez et al. (2014) ont montré qu'à l'échelle de l'île entière, et donc pour une hétérogénéité environnementale supérieure, ce sont bien les différences d'habitat qui sont perçues comme dominantes pour expliquer la distribution des espèces dominantes d'arbres de canopée.

Vellend (2010) résume dans sa synthèse conceptuelle de l'écologie des communautés que les nombreux mécanismes responsables de l'assemblage des communautés présentés au cours des dernières décennies peuvent tous être résumés en quatre catégories de processus : la **sélection**, la **dérive**, la **spéciation** et la **dispersion**. Alors que la sélection concerne les interactions déterministes entre espèces ainsi qu'entre espèces et environnement abiotique, la spéciation souligne l'importance de processus agissant à des échelles spatiales et temporelles beaucoup plus larges pour comprendre les structures locales de diversité. Ceci revient à reconnaître l'importance du pool régional d'espèce, lequel dépend du processus de spéciations. La dérive fait référence à la variation aléatoire des abondances des espèces en lien avec des évènements imprévisibles (par ex. perturbations, mort d'individus), c'est-à-dire à la dérive écologique. La dispersion limitée des organismes ou de leur descendance est également un processus essentiel de coexistence des espèces et de différentiation de communautés locales dès lors que toutes les espèces ne peuvent être ou s'installer partout (Seidler et Plotkin 2006).

La perception de l'importance relative de la théorie de la niche et de la théorie neutre, à une échelle donnée, dépend néanmoins encore d'un facteur essentiel, à savoir la qualité et la pertinence du choix des paramètres environnementaux mesurés (Jones et al. 2008, Chang et al. 2013). Ainsi, alors que la plupart des études d'écologie des communautés d'arbres se sont concentrées au départ sur des paramètres topographiques facilement mesurables (par ex. pente, altitude, convexité) pour aborder l'importance des effets de niche sur l'assemblage des espèces de la communauté (par ex. Legendre et al. 2009, De Cáceres et al. 2012), Chang et al. (2013) ont montré que la prise en compte de variables physico-chimiques de sol précises modifiait grandement la perception obtenue uniquement avec les variables topographiques, au point d'inverser l'importance relative des effets neutres et de niche.

II. Ecologie spatiale : des patterns aux processus

La complexité des interactions du vivant ainsi que la nature contrastée des processus écologiques rendent la tâche de l'écologie des communautés particulièrement ardue. Néanmoins, aussi différents soient-ils, la grande majorité des processus écologiques ont en commun le fait qu'ils imposent une signature spatiale dans les communautés vivantes (McIntire et Fajardo 2009, Legendre et Legendre 2012, Fortin et Dale 2014).

En effet, l'environnement peut être appréhendé comme premièrement structuré par des processus physiques de large échelles – orogéniques, géomorphologiques sur les terres, liés aux vents et courants en milieu aquatique – qui par transfert d'énergie génèrent des gradients et structures agrégées dans l'environnement. Ces structures de larges échelles sont reflétées dans les systèmes biologiques sous forme de biomes et écosystèmes (Legendre et Legendre 2012). La structuration spatiale de l'environnement abiotique peut également générer, selon le même principe, des motifs de distribution agrégés à des échelles plus locales (par ex. Legendre et al. 2009, Garzon-Lopez et al. 2014, Bauman et al. 2016, Muledi et al. 2017, Vleminckx et al. 2017). D'autre part, localement, les interactions biotiques mènent certaines espèces à être systématiquement éloignées les unes des autres (par ex. inhibition par compétition ; voir autocorrélation spatiale négative dans la boîte 1), alors que d'autres tendent à être souvent plus proches que par hasard (par ex. relations de facilitation, prédateurs-proies, hôtes-symbiotes, partitionnement des ressources ; voir autocorrélation spatiale positive dans la boîte 1) (Hyatt et al. 2003, Yamazaki et al. 2009, HilleRisLambers et al. 2012, Wiegand et Moloney 2014, Velázquez et al. 2016). La dispersion limitée des organismes est également à la base de structures spatiales de type agrégées dans les communautés (par ex. barrière physique empêchant la dispersion d'une espèce d'une communauté à une autre, *kernel* de dispersion de graines) et est un processus important du maintien de la diversité d'une communauté (Stoll et Prati 2001, Bell et al. 2006, Seidler et Plotkin 2006, Garzon-Lopez et al. 2014, Lowe et McPeek 2014).

Hormis la dérive écologique, ces **processus** laissent tous une **empreinte spatiale** propre dans les **communautés** vivantes (Fig. 2b, c, d pour des exemples). La dimension spatiale des jeux de données écologiques est donc intrinsèquement fonctionnelle, dès lors que l'écosystème ne peut fonctionner qu'à la condition de la présence de cette organisation spatiale multi-échelle (Legendre et Legendre 2012, Fortin et Dale 2014).

Le reflet de la structuration spatiale d'un paramètre abiotique dans la distribution d'une espèce est appelé **dépendance spatiale induite** (*induced spatial dependence* ; ISD) et n'est qu'une reformulation dans le domaine spatial de la théorie de la niche ou du filtre environnemental mentionnés précédemment (Fig. 2b, d). La structuration spatiale d'une communauté causée par les processus intrinsèques à celle-ci (par ex. **limitation de dispersion**) correspond à ce que certains appellent autocorrélation spatiale *sensu stricto* ou autocorrélation spatiale véritable (Fig. 2c, d) (Legendre et Legendre 2012). En statistique, l'autocorrélation spatiale véritable est la dépendance spatiale restant dans le terme d'erreur d'une variable réponse observée dans l'espace une fois que toutes les variables explicatives spatialement structurées sont considérées dans le modèle. La somme de l'autocorrélation spatiale vraie et de l'ISD peut être appelée autocorrélation spatiale *sensu lato*, ou simplement corrélation spatiale (Legendre and Legendre 2012, voir boîte 1). Dans ce travail, le terme autocorrélation spatiale sera, sauf indication contraire, utilisé au sens large de corrélation spatiale, ne portant donc pas de connotation implicite quant à un processus sous-jacent.

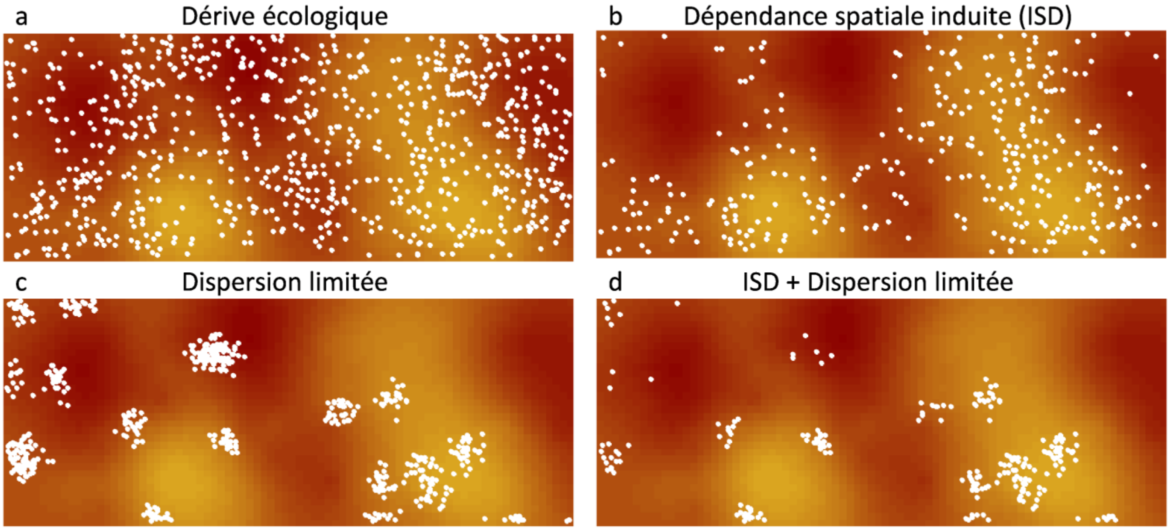


Figure 2 : Illustration de liens entre processus écologiques et patterns spatiaux. La carte de fond est une carte environnementale. La distribution des points blancs (individus) est régie par le/les processus indiqué(s). (a) Processus stochastique non structurant (dérive écologique) ; espèce généraliste. (b) Dépendance spatiale induite par un environnement spatialement structuré ; espèce spécialiste. (c) et (d) : espèce généraliste et spécialiste, respectivement, auxquelles s'ajoute une dispersion limitée.

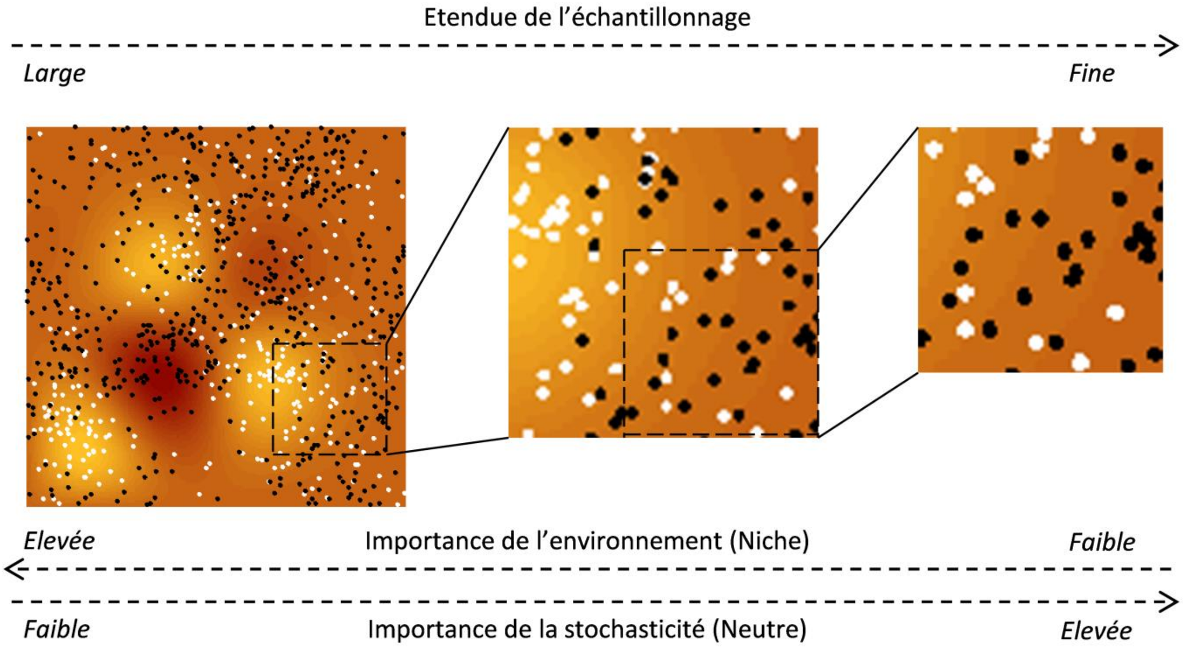

Etendue de l'échantillonnage
Large
Fine
Elevée
Importance de l'environnement (Niche)
Faible
Faible
Importance de la stochasticité (Neutre)
Elevée

Figure 3 : Perception relative des processus de niche et neutres en fonction de l'étendue de l'échantillonnage ou de l'échelle d'observation au sein de la zone échantillonnée. La carte de fond est un environnement hétérogène. Les points blancs et noirs sont deux espèces à préférence écologique opposée. L'importance relative de la niche diminue au profit des processus neutres lorsque moins de variabilité environnementale est capturée par l'échantillonnage.

Plusieurs classes d'approches ont ainsi été développées au fil des trois dernières décennies pour étudier les liens entre les patterns spatiaux au sein des communautés et les processus les sous-tendant (McIntire et Fajardo 2009).

Une première approche est l'**expérimentation**, qui consiste à tester l'influence d'un processus donné sur une ou plusieurs espèces en contrôlant tous les éléments externes au processus d'intérêt (par ex. expérience en jardin commun, Stoll et Prati 2001, Peay et al. 2015) ou en excluant un processus sur le terrain (par ex. exclusion des herbivores, McIntire et Hik 2005). Le but de cette approche est de tester assez directement l'effet d'un processus sur la réponse mesurée (par ex. croissance ou abondance d'une ou plusieurs espèces). Néanmoins, il est rare que la manipulation d'un facteur d'assemblage de communauté n'en affecte pas d'autres, dès lors qu'il est commun qu'un même facteur écologique ait un effet sur différentes composantes de l'écosystème qui elles-même influencent l'assemblage de la communauté (voir détails dans HilleRisLambers et al. 2012).

Une seconde approche consiste à définir les valeurs de paramètres de **modèles mécanistes** spatiaux prédictifs à partir de données de terrain. Cette approche vise à générer un processus d'intérêt de manière réaliste sur base de l'estimation de paramètres à partir de données réelles (Schultz et Crone 2001, Clark et al. 2004, Ogle et al. 2004, Mcintire et al. 2007, Schurr et al. 2012, Cabral et al. 2017), par exemple au moyen de modèles Bayésiens (Ellison 2004, Munoz et al. 2018). Le but est alors par exemple de vérifier si les patterns observés peuvent être reproduits par des processus donnés (par ex. Clark et al. 2004), ou de prédire des patterns à des échelles plus larges à partir de paramètre mesurés à des échelles plus fines (par ex. Schultz et Crone 2001, voir Dormann et al. 2012 et Cabral et al. 2017 pour une synthèse). Si cette approche a l'avantage de contrôler les processus à l'œuvre dans le modèle, elle est néanmoins limitée par la quantité de paramètres et la difficulté d'estimer correctement certains processus complexes ou la somme et l'interaction de plusieurs processus (Cabral et al. 2017).

La dernière approche – l'**approche corrélative** (Dormann et al. 2012) – consiste à étudier directement les patterns spatiaux des communautés afin de pouvoir en déduire les processus sous-jacents. Il s'agit de retracer l'action de processus à partir des empreintes spatiales qu'ils ont laissées (McIntire et Fajardo 2009, Dray et al. 2012, Legendre et Legendre 2012, Wiegand et Moloney 2014, Velázquez et al. 2016). Cette approche des « **patterns spatiaux comme substituts des processus** » a l'avantage de partir de données écologiques réelles portant en elles la somme de tous les processus ayant contribué à les générer. Cette approche a néanmoins été critiquée sur base d'un certain nombre de phénomènes biologiques rendant le lien entre pattern et processus complexe dans certains cas (McIntire et Fajardo 2009). Par exemple, plusieurs processus peuvent générer un même pattern (Cale et al. 1989, Perry et al. 2006), les liens de causalité peuvent ne pas être directes et évidents (Rees et al. 1996, Turner et al. 2001) et certains processus peuvent être le résultat de patterns spatiaux, plutôt que le contraire (Stoll et Prati 2001).

Néanmoins, McIntire et Fajardo (2009) avancent que dans de nombreux autres cas, un processus peut générer un pattern unique, un lien de causalité simple peut exister entre processus et pattern, et l'influence d'un pattern sur un processus peut ne pas avoir lieu à la même échelle que l'influence du processus sur le pattern, de sorte que les deux peuvent être discernés.

En outre, ces limitations de l'approche corrélative auraient une importante composante d'ordre analytique plutôt qu'intrinsèquement biologique. En effet, ces critiques diverses datent d'une époque à laquelle les outils d'analyses spatiales étaient bien moins avancés qu'actuellement, de sorte que le degré de caractérisation de patterns spatiaux atteignable aujourd'hui permet d'affiner fortement l'inférence des processus sous-jacents (Dray et al. 2012, Legendre et Legendre 2012, Velázquez et al. 2016), minimisant ainsi l'ampleur des limitations décrites ci-dessus. McIntire et Fajardo (2009) proposent ainsi l'utilisation des patterns spatiaux comme proxys pour découvrir l'effet de processus difficiles à mesurer ou non mesurés. Ils affirment et illustrent que l'analyses de patterns spatiaux et de leurs propriétés (par ex. échelle, intensité, *patchiness*) peut aider l'inférence des processus sous-jacents à condition que des hypothèses alternatives bien définies soient proposées a priori et que des modèles statistiques adaptés soient utilisés pour sélectionner l'hypothèse la plus probable. Néanmoins, comme nous le verrons plus loin, un affinement des méthodes d'analyses de patterns spatiaux ainsi que le développement de nouveaux outils méthodologiques sont encore nécessaires pour atteindre cet objectif.

Dans cette thèse, c'est cette dernière approche de l'utilisation des patterns spatiaux de données réelles pour approcher les processus qui sera abordée, par le biais de cas d'études ainsi que par une exploration et amélioration technique de certains de ses outils statistiques. Nous illustrerons cependant également la façon dont l'utilisation de simulations de processus peut permettre de mettre à l'épreuve les outils statistiques utilisés dans l'approche corrélative partant des patterns spatiaux.

Boîte 1 : Terminologie spatiale

Autocorrélation spatiale (*sensu lato*) : Dépendance spatiale entre des observations. Les observations ne sont pas indépendantes les unes des autres aux distances où une autocorrélation spatiale est présente.

Autocorrélation spatiale positive : Deux observations spatialement proches présentent des valeurs plus proches que par hasard. Exemples de processus sous-jacents : dispersion limitée, dépendance spatiale induite

Autocorrélation spatiale négative : Deux observations spatialement proches présentent des valeurs plus dissimilaires que par hasard. Exemples de processus sous-jacents : Compétition, prédation dépendante de la densité.

Pattern spatial : Un pattern spatial décrit une caractéristique observable d'un système et sa configuration.

Grain : Le grain est la taille ou dimension d'une unité d'échantillonnage (par exemple, une parcelle de 25 × 25 m). Lorsque ces unités échantillonnées sont adjacentes, la résolution spatiale minimum des méthodes d'analyses spatiales les plus fines est le grain. C'est une propriété spatiale d'un design d'échantillonnage.

Etendue : L'étendue est la dimension la plus large qui englobe toutes les unités d'échantillonnage (par ex. quadrats, parcelles). Il s'agit également d'une propriété spatiale d'un design d'échantillonnage.

Echelle : L'échelle est utilisée ici au sens de *focus* (Scheiner 2011), c'est-à-dire, la dimension de la somme des grains constituant un pattern spatial. Afin d'éviter une confusion avec les termes de petite et grande échelle utilisés en géographie, un pattern spatial grossier, d'échelle élevée, sera qualifié de pattern de large échelle, alors qu'un pattern d'échelle plus locale sera qualifié de pattern d'échelle fine (Legendre et Legendre 2012).

III. Autocorrélation spatiale : problème ou opportunité ?

Depuis fort longtemps, la non-indépendance de paires d'observations (autocorrélation spatiale) a posé problème en Science, dès lors que la condition de base des tests statistiques standards (par ex. *ordinary least squares*, OLS, *generalised linear models*, GLM) est l'indépendance des résidus dans un modèle (Legendre et Fortin 1989, Diniz-Filho et Bini 2005, Dormann et al. 2007, Peres-Neto et Legendre 2010, Legendre et Legendre 2012). La présence d'autocorrélation spatiale dans le terme d'erreur d'une variable réponse exprimée en fonction d'un certain nombre de variables explicatives dans un modèle viole donc cette condition. La surestimation du nombre de degrés de liberté réels du modèle découlant de cette violation cause ainsi une augmentation de la probabilité de rejeter l'hypothèse nulle alors que celle-ci est vraie (taux d'erreur de type I), de même qu'un biais dans l'estimation du modèle. La prise en compte explicite des aspects spatiaux d'un jeu de données est donc avant tout née de la volonté de pouvoir utiliser les tests statistiques standards de façon non biaisée. Ceci nous place dans un paradigme de *spatial nuisance* (Peres-Neto et Legendre 2010) au sein duquel l'autocorrélation spatiale est avant tout un problème à résoudre (voir la synthèse de Dormann et al. 2007).

D'autre part, tel qu'expliqué dans la section précédente, la dimension spatiale des communautés est fonctionnelle et mérite une attention toute particulière dès lors qu'elle constitue une porte d'accès privilégiée vers les processus écologiques sous-tendant la mise en place des patterns de distribution (Fig. 2) (Legendre 1993, Dray et al. 2012, Legendre et Legendre 2012). Cette vision de la dimension spatiale des communautés comme une signature des processus a mené à un développement foisonnant d'analyses spatialement explicites visant à détecter et décrire avec un maximum de détail les patterns de distribution des espèces afin de s'approcher des processus qui en sont responsables (par ex. Dray et al. 2006, 2012, Jombart et al. 2009, Velázquez et al. 2016). L'écologie spatiale qui en résulte s'inscrit alors dans le paradigme de *spatial legacy* (Peres-Neto and Legendre 2010), qui sera la vision adoptée et développée dans ce travail.

IV. Méthodes d'analyses spatialement explicites

Au cours des dernières décennies, une importance croissante a été accordée aux analyses spatialement explicites en écologie (Legendre 1993, Griffith et Peres-Neto 2006, Fortin et Dale 2014) mais aussi dans d'autres domaines (par ex. géographie, sciences économiques ; Griffith 1996, Stakhovych et Bijmolt 2008, Patuelli et al. 2011), alors qu'en parallèle d'importantes avancées technologiques permettaient d'obtenir, gérer et stocker d'importantes quantités de données spatiales (par ex. *remote sensing*, *geographic information system*, etc). Ceci a mené au développement d'un nombre important de méthodes d'analyses spatialement explicites dans de nombreux domaines (par ex. géographie, biogéographie, génétique des populations, écologie des communautés) (Griffith 2003, Dray et al. 2006, 2012, Jombart et al. 2008, Fortin et Dale 2014, Wiegand et Moloney 2014, Wagner et Dray 2015, Velázquez et al. 2016).

IV.1. Méthodes spatiales basées sur l'utilisation de vecteurs propres

En écologie, les questions de la structuration des communautés d'organismes vivants ont traditionnellement été abordées par le biais des analyses multivariées (Gauch 1982). Au cours des dernières décennies, ces dernières ont progressivement intégré la dimension spatiale des communautés de façon explicite, transformant ainsi progressivement l'écologie des communautés en écologie spatiale. Ainsi, Legendre (1990) a notamment proposé la *trend surface analysis*, méthode dans laquelle la structuration spatiale d'une variable réponse est capturée au moyen d'une fonction polynomiales des coordonnées géographiques X et Y centrées. Cette approche ne permet néanmoins que la considération de structures spatiales grossières, dès lors qu'il faudrait un nombre beaucoup trop important de termes polynomiaux pour modéliser des structures fines. Ceci a limité cette méthode, dans la pratique, aux polynômes du troisième degré et donc à des structures relativement larges et simples (Borcard and Legendre 2002).

D'autre part, Griffith (1996) a effectué un important travail pionnier en analyses spatiales dans le domaine de la géographie en mettant en évidence que des **vecteurs propres spatiaux** pouvaient être utilisés pour représenter des patterns spatiaux sur carte ou pour contrôler l'autocorrélation spatiale dans les résidus de modèles (voir *spatial nuisance*).

Parallèlement, des développement similaires ont eu lieu en écologie des communautés, notamment avec la méthode des ***principal coordinates of neighbour matrices*** (**PCNM** ; Borcard and Legendre 2002) dont le principe est également de générer des vecteurs propres spatiaux, lesquels permettent la détection de patterns multi-échelles de façon bien plus puissantes et précises qu'avec la *trend surface analysis*.

Les étapes principales de la méthode PCNM consistent à 1) générer une matrice de distances euclidiennes entre les sites au sein desquels l'abondance d'une ou plusieurs espèces a été mesurée, 2) tronquer cette matrice en remplaçant toutes les distances supérieures à un certain seuil par quatre fois la valeur seuil, et 3) effectuer une analyse en coordonnées principales (PCoA) sur la matrice tronquée. Le résultat de la PCoA est un ensemble de vecteurs propres spatiaux pouvant alors servir de variables explicatives spatiales (prédicteurs spatiaux) dans des modèles de type OLS ou GLM pour une variable réponse (cas univarié) ou dans des analyses canoniques de type ordinations contraintes (par ex. analyse de redondance, RDA, Rao 1964) pour l'analyse d'une matrice réponse (cas multivarié).

Dans cette méthode, Borcard et Legendre (2002) proposent que la distance seuil de troncature de la matrice de distance soit celle du plus long axe d'un *minimum spanning tree*. Ce dernier est un schéma de connexion liant des points de sorte à ce que tous soient connectés tout en minimisant la somme des distances des droites les reliant entre eux. Cette distance seuil est donc la plus petite distance qui permet à tous les sites de rester connectés (voir Fig. S1 du Chapitre IV) et permet ainsi de définir que les paires de sites dont la distance est remplacée sont considérés comme non connectés. La multiplication du seuil de connectivité par un facteur de 4 pour les paires de sites distants de plus que le seuil provient de l'observation empirique qu'au-delà de cette valeur, les vecteurs propres restent les mêmes (à une constante multiplicative près).
L'utilisation de la PCoA génère un certain nombre de vecteurs propres associés à une valeur propre négative. Ces derniers ne peuvent être utilisés, dès lors que les coordonnées des sites sur ces axes sont des nombres complexes. Seules les valeurs propres positives représentent la dimension euclidienne des relations de voisinage de la matrice tronquée. En conséquence, pour un design d'échantillonnage régulier, la PCNM génère un maximum de $2n/3$ vecteurs propres spatiaux, où n est le nombre sites échantillonnés.

Dray et al. (2006) ont rendu explicite le formalisme mathématique sous-tendant la proposition empirique des PCNM. Ils ont ainsi montré que les PCNM ne sont qu'un cas particulier faisant partie d'une méthode beaucoup plus globale et flexible qui fut nommée « carte de vecteurs propres de Moran » (***Moran's eigenvector maps***, **MEM**). Dans la méthode MEM, les vecteurs propres spatiaux (aussi appelés **variables MEM** ou **prédicteurs spatiaux**) sont générés en diagonalisant une matrice de pondération spatiale (*spatial weighting matrix*, SWM) doublement centrée (lignes et colonnes). Cette méthode génère ainsi non plus un maximum de 2*n*/3 vecteurs propres, mais *n*-1, pouvant tous être utilisés et ce indépendamment de la régularité du design d'échantillonnage. De plus, Dray et al. (2006) mirent en évidence que la valeur propre de ces axes était proportionnelle à l'indice *I* de Moran – une mesure classique du degré d'autocorrélation spatiale d'une variable – et était donc directement proportionnelle au degré d'autocorrélation spatiale. En conséquence, le premier vecteur propre présente la plus grande valeur propre et donc la structure spatiale positivement autocorrélée de plus large échelle, suivi par le deuxième vecteur propre qui présente une structure légèrement plus fine, et ainsi de suite (Fig. 4a-i). En arrivant aux valeurs propres négatives, les vecteurs propres présentent donc un *I* de Moran négatif, ce qui indique qu'ils présentent des structures spatiales négativement autocorrélées (Fig. 4j-l). Les premiers vecteurs propres associés aux valeurs propres négatives présentent ainsi des structures spatiales négativement autocorrélées mais de relativement plus large échelle que les derniers vecteurs propres associés aux valeurs propres négatives. Le lien entre échelle spatiale et signe de la valeur propre est néanmoins un sujet qui nécessite d'être approfondi (voir discussion du Chapitre III).

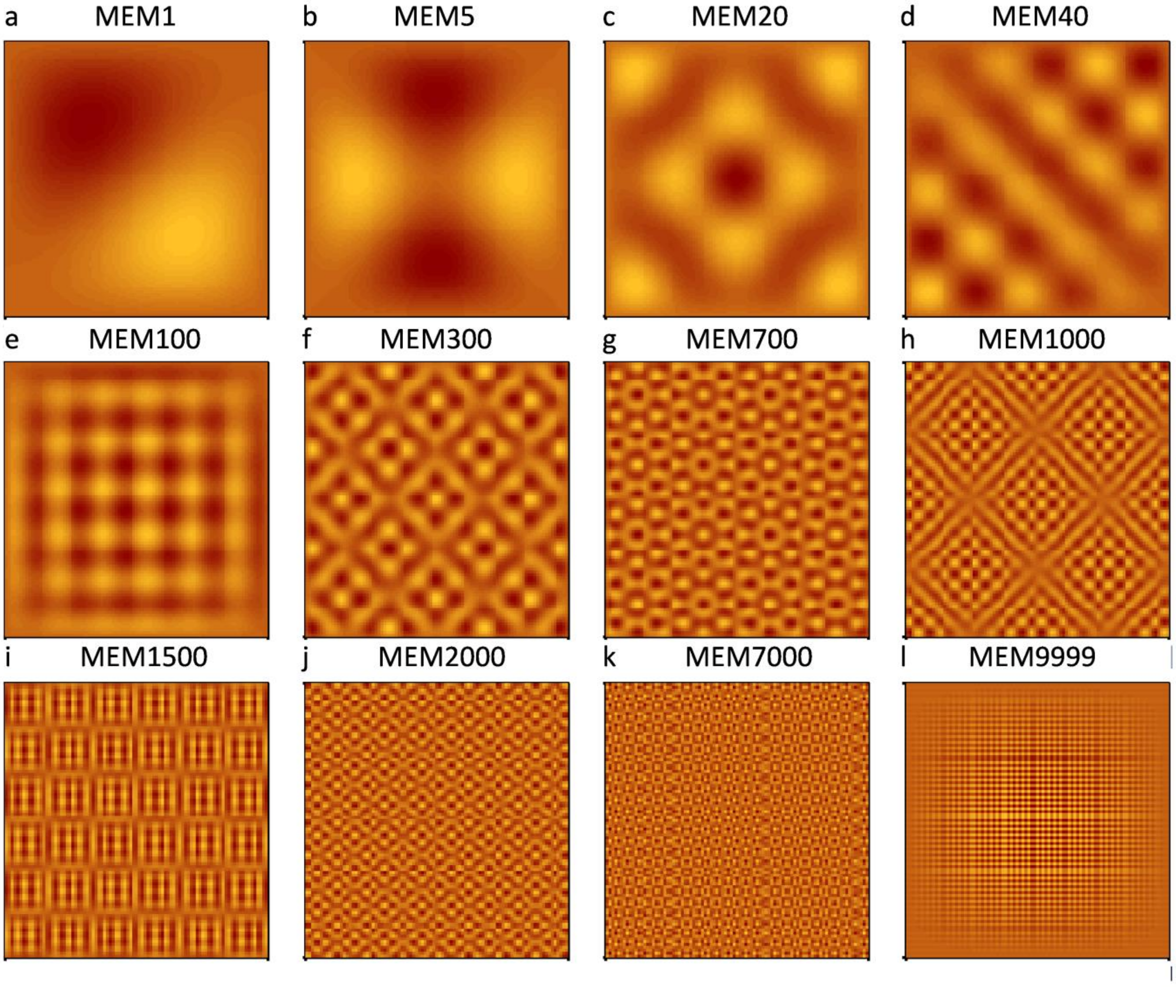


Figure 4 : Illustration de motifs d'autocorrélations spatiale compris dans les vecteurs propres spatiaux (variables MEM) générés sur une grille de 100 × 100 cellules (9999 variables MEM). a-j :

Motifs d'autocorrélation spatiale positive d'échelles larges et fines. k-l : Motifs spatiaux présentant une structure négativement autocorrélée aux échelles les plus fines. Un sous-ensemble de ces cartes de vecteurs propres de Moran peut être utilisé dans divers types d'analyses spatialement explicites afin de contrôler l'autocorrélation spatiale des résidus d'un modèle ou pour une description fine des motifs de structuration spatiale d'une communauté.

Cette généralisation des PCNM par les MEM a ouvert un champ de possibles colossal, dès lors que la SWM peut être définie d'une infinité de façons. La SWM (ou matrice **W**, en langage matriciel) est le résultat du produit Hadamard (terme par terme) d'une matrice de connectivité binaire **B,** définissant les paires de sites connectés (1) et ceux ne l'étant pas (0), et d'une matrice de pondération **A**, donnant un poids à chaque connexion selon une fonction de pondération décroissant généralement avec la distance. La matrice **B** peut définir le statut des connexions (1 ou 0) à partir d'un schéma de connexion (par ex. triangulation de Delaunay, graphe de Gabriel, *minimum spanning tree*, etc, Legendre et Legendre 2012 ; voir Fig. S1 du Chapitre III en annexe), générant alors une *graph-based* SWM, ou à partir d'un critère de distance (i.e. seules les paires de sites séparées par une distance inférieure à un seuil donné sont connectées), menant alors à une *distance-based* SWM. Les PCNM ne sont dès lors qu'un cas particulier de *distance-based* MEM (Dray et al. 2006).

Le fait que les MEM génèrent plus de vecteurs propres (*n*-1) que les PCNM leur confère une puissance supérieure et la possibilité de mettre en évidence des patterns spatiaux plus fins et complexes que les PCNM (Dray et al. 2006). Néanmoins, considérer *n*-1 vecteurs propres dans un modèle visant à expliquer la variabilité d'une variable réponse causerait la saturation du modèle et mènerait à un coefficient de détermination (R^2) de 1.
Une sélection parmi les *n*-1 vecteurs propres générés par une SWM donnée est donc nécessaire avant de pouvoir utiliser les vecteurs propres comme prédicteurs spatiaux dans un modèle. Si de nombreuses méthodes de sélection de vecteurs propres ont été proposées pour régler ce problème (Griffith et Peres-Neto 2006, Tiefelsdorf et Griffith 2007, Blanchet et al. 2008, Bini et al. 2009), celles-ci n'ont pas été confrontées et comparées en termes de performances statistiques (par ex. taux d'erreur de type I, puissance statistique, précision de l'estimation du modèle). Or, il a été montré qu'une sélection inadaptée des vecteurs propres peut gonfler le taux d'erreur de type I et biaiser les coefficients des prédicteurs de modèles dans lesquels les vecteurs propres seraient intégrés (Griffith 2003, Blanchet et al. 2008, Peres-Neto et Legendre 2010, Diniz-Filho et al. 2012). La question de la sélection de vecteurs propres spatiaux (au sein d'une SWM donnée) sera traitée dans le Chapitre III.

En outre, si la gamme des possibilités de définition de la SWM confère aux MEM une flexibilité énorme, elle génère aussi un problème de taille, à savoir, le choix de la SWM parmi un espace potentiellement infini de possibilités. En géographie et en sciences économiques, il a été montré que la sélection d'une SWM est une étape fondamentale de l'analyse pouvant fortement influencer l'estimation de modèles et donc l'interprétation des résultats (Stetzer 1982, Florax et Rey 1995, Griffith et Lagona 1998, Stakhovych et Bijmolt 2008, Kostov 2010, Patuelli et al. 2011). Dray et al. (2006) insistent en effet sur le fait que le choix de la SWM est une étape clé pouvant influencer les patterns spatiaux détectés ainsi que l'estimation des coefficients de modèles. Or, jusqu'à présent, une proportion élevée des études ne précise pas la nature de la SWM utilisée ou continue d'utiliser les PCNM (une méthode limitée et manquant de formalisme mathématique) au lieu d'utiliser les MEM (pouvant offrir des alternatives plus robustes aux designs d'échantillonnage irréguliers, voir Dray et al. 2006). L'étape clé des méthodes basées sur des vecteurs propres spatiaux a donc jusqu'à présent été abordée de façon inadéquate. Cette problématique de la sélection d'une SWM sera traitée dans le Chapitre IV.

IV.2. Vecteurs propres spatiaux et variables environnementales

Les variables MEM sont un outil puissant et flexible qui a été utilisé, associé à des variables environnementales ou pas, dans de nombreuses approches et pour différentes questions fondamentales en écologie (Dray et al. 2012).

Dans un paradigme de ***spatial nuisance***, les variables MEM peuvent être utilisées afin d'éliminer l'autocorrélation spatiale des résidus d'un modèle visant à expliquer une variable ou matrice réponse en fonction d'une série de variables explicatives (par ex. variables environnementales) de façon non biaisée et avec un taux de t'erreur de type I correct (Getis et Griffith 2002, Diniz-Filho et Bini 2005, Griffith et Peres-Neto 2006). Il est intéressant de noter ici qu'un développement similaire aux MEM a eu lieu sur base des distances évolutives d'arbres phylogénétiques d'espèces (voir *phylogenetic eigenvector regression*, Diniz-Filho et al. 1998, 2012, *phylogenetic eigenvector maps*, Guénard et al. 2013, *phylogeny-trait decoupling*, de Bello et al. 2017) afin d'éliminer la composante phylogénétique des traits fonctionnels dans les communautés (autocorrélation phylogénétique) ou au contraire de décrire les patterns de structuration phylogénétique aussi précisément que possible, respectivement dans un parallèle des approches de *nuisance* et *legacy* (McIntire et Fajardo 2009, Peres-Neto et Legendre 2010).

Selon le paradigme de ***spatial legacy***, les variables MEM peuvent être utilisées dans des OLS, GLM ou RDA pour expliquer une variable ou matrice réponse (généralement des abondances d'espèces) (Legendre et Legendre 2012). La lettre **Y** sera utilisée par la suite pour indiquer une variable réponse ou une matrice de variables réponses (ce qu'on vise à expliquer). Les variables MEM permettent de réaliser des tests spatialement explicites de l'effet d'un ensemble de variables explicatives environnementales (**X**) sur **Y**. En outre, ces tests peuvent être effectué indépendamment aux différentes échelles des patterns spatiaux détectés dans **Y** grâce à l'orthogonalité des vecteurs propres. L'échelle d'action des différents facteurs environnementaux sur la communauté peut ainsi être quantifiée de manière claire (Dray et al. 2012, Legendre et Legendre 2012). Les variables MEM peuvent également servir à modéliser les structures spatiales des résidus de **Y** exprimé en fonction d'un ensemble de variables explicatives environnementales (**X**). La présence d'autocorrélation spatiale significative dans ces résidus est le signe qu'un processus – complémentaire à ceux déjà pris en compte dans le modèle – génère un pattern spatial résiduel dans **Y** (Borcard et al. 2004, McIntire et Fajardo 2009, Dray et al. 2012). La spécification a priori de mécanismes hypothétiques pouvant agir aux échelles correspondantes ainsi qu'une caractérisation précise du pattern résiduel peuvent alors aider à générer des hypothèses quant aux processus sous-jacents (voir Chapitres II et VI).

La partition de variation (*variation partitioning* ; VP) est également un cadre d'analyse particulièrement répandu en écologie (Fig. 5) (Borcard et Legendre 1994, Borcard et al. 2004, Peres-

Neto et Legendre 2010). La VP vise à expliquer **Y** à partir de deux ou trois ensembles de variables explicatives au moyen d'une série de régressions linéaires multiples ou de RDA globales et partielles. La VP permet ainsi en théorie de quantifier les portions de variation de **Y** expliquées uniquement (fractions [a] et [c], Fig. 5) et conjointement (fraction [b], Fig. 5) par les différentes composantes explicatives. En écologie, la VP est souvent utilisée sur base de deux composantes explicatives : un ensemble de variables environnementales (**X**) et un ensemble de variables spatiales (variables MEM). Dans ce cadre, le VP permet donc d'expliquer la part de **Y** expliquée par l'environnement seul (fraction [a]), par l'environnement spatialement structuré (covariation de **X** et des variables MEM ; fraction [b]) et par les variables MEM seules (une fois que la variation expliquée par **X** a été retirée ; fraction [c]) (v. Fig. 5).

La significativité des fractions [bc], [a] et [c] sont testées à l'aide d'une procédure de permutation des résidus du modèle de régression ou RDA partielle correspondant (Fig. 5) (Anderson and Legendre 1999). Lorsque toutes les variables environnementales pertinentes ont été mesurées et se trouvent dans **X**, une fraction [c] significative a souvent été interprétée comme étant porteuse de patterns générés par la dispersion limitée des espèces (Peres-Neto et Legendre 2010, Legendre et Legendre 2012). Néanmoins, il est difficile de s'assurer que cette fraction ne soit pas générée par une ou plusieurs variables environnementales non mesurées (qui, si elles étaient mesurées, devrait donc se trouver dans la fraction [b]). L'analyse des propriétés spatiales (par ex. échelle, forme, degré d'autocorrélation spatiale) du pattern spatial de la fraction [c] peut aider à donner plus ou moins de poids à l'une ou l'autre hypothèse, selon les cas (McIntire et Fajardo 2009, Bauman et al. 2016).

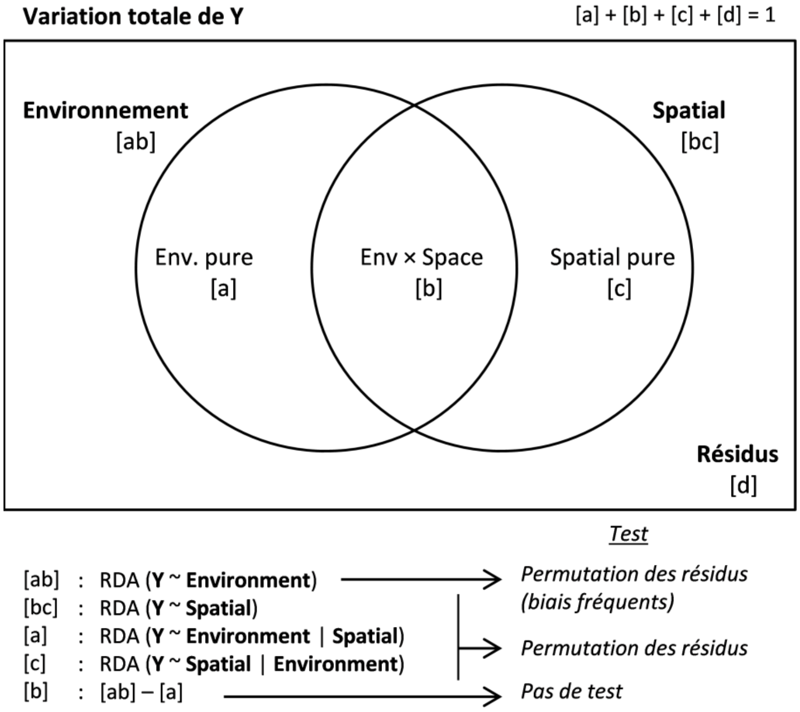


Figure 5 : Illustration schématique du partitionnement de la variation sous la forme d'un diagramme de Venn. La variation au sein de la matrice réponse Y (abondances des espèces de la communauté) est décomposée en composantes environnementale (une matrice de variables environnementales ; [ab]) et spatiale (variables MEM ; [bc]). La variation totale de Y (1) est

répartie en une portion expliquée par le modèle global général ([abc]) et une fraction résiduelle ([d]). La fraction [abc] est décomposée en composantes environnementale et spatiale pure ([a] et [c], respectivement) et une composante partagée par l'environnement et les variables MEM. Cette dernière fraction correspond à un motif de structuration spatiale de Y partagé par l'environnement. Les fractions [bc], [a] et [c] sont testées par permutation des résidus (Anderson and Legendre 1999). Ce test peut être utilisé pour [ab], mais présente généralement un taux d'erreur de type I élevé (voir texte). La fraction [b] est calculée par soustractions d'autres R^2 et n'est pas testable.

La procédure de test par permutation des résidus présente néanmoins un taux d'erreur de type I gonflé pour la fraction [ab] (l'effet global de **X**), dès lors que les résidus du modèle de **Y** en fonction de **X** présente souvent de l'autocorrélation spatiale (une fraction [c] significative) (Peres-Neto et Legendre 2010). La fraction [b] ne peut non plus être testée par le test classique de permutation des résidus dès lors qu'aucun modèle ne soutend cette fraction et que celle-ci est donc calculée par soustraction d'autres fractions et n'a pas de degrés de liberté (Fig. 5). D'un point de vue de l'inférence de processus, cette fraction [b], bien que non testable, a souvent été interprétée comme un effet potentiel de dépendance spatiale induite (ISD) par **X**, dès lors que la structure spatiale de **X** se trouve reflétée dans **Y**. Néanmoins, sans un test approprié de cette fraction, il n'est pas possible de balayer l'éventualité de structures spatiales de **X** et **Y** se superposant par hasard. Ceci a mené un nombre important d'autres études à se baser prioritairement sur la fraction [a] pour aborder les effets de l'environnement sur **Y**, dès lors que cette fraction a l'avantage de pouvoir être testée sans biais spatiaux (Peres-Neto et Legendre 2010). Cette approche pose cependant question, dès lors que la majorité des paramètres abiotiques sont naturellement spatialement autocorrélés (Legendre et Legendre 2012) et que donc une portion significative de la variabilité de **X** pouvant influencer **Y** est attendue dans la fraction [b]. En outre, on peut se demander si la part de **X** non spatialement autocorrélée (fraction [a]) aux échelles capturables par le design d'échantillonnage ne l'est simplement pas à une échelle inférieure à la résolution spatiale de l'étude par exemple, ce qui rend la différence de statut et de traitement des fractions [a] et [b] potentiellement encore plus arbitraire, d'un point de vue écologique. Les effets de dépendances spatiales induites étant à la base de la théorie de la niche, la mise en place d'un test fiable de la fraction [b] ainsi que de la fraction [ab] est donc une priorité. Cette problématique ainsi que deux propositions de solutions originales seront abordées dans le Chapitre V.

IV.3. Analyse de semis de points

Toujours dans le cadre de l'étude des patterns spatiaux comme *proxy* des processus, une approche différente et complémentaire des approches multivariées présentées dans la section précédente est l'analyse spatiale de semis de points (***spatial point pattern analysis***, **SPPA**) (Ripley 1981, Diggle 2003, Illian 2008, Wiegand et Moloney 2014, Velázquez et al. 2016). Le SPPA consiste en un ensemble de techniques statistiques pour l'analyse de semis (ou patterns) de points. Ces derniers sont des localisations d'objets écologiques (dans cette thèse, des arbres) comprises dans une fenêtre d'observation (par ex. Figs. 2 et 3). A la différence des approches multivariées qui résumaient la communauté à des abondances d'espèces dans des sites (par ex. quadras), l'approche du SPPA considère donc la **localisation unique de chaque individu**. A ces derniers peuvent être superposées des « marques » quantitatives (par ex. la taille, l'âge) ou qualitatives (par ex. espèce, survivant *vs* mort) ou encore des covariables environnementales. L'analyse spatiale de ces semis de points permet de tester différentes théories écologiques, dès lors que les patterns de points constituent des « archives écologiques » des processus sous-jacents (par ex. Fig. 2) (Wiegand et al. 2003, McIntire et Fajardo 2009, Velázquez et al. 2016). Pour ce faire, les patterns spatiaux réels sont caractérisés (par ex. densité des points, taille et densité d'agrégats, nombre moyen de points par agrégat) et peuvent être comparés à des modèles nuls de complexité croissante de façon spatialement explicite afin d'inférer les processus pouvant avoir généré le pattern de point réel. En résumé, cette approche consiste à estimer à partir du pattern de point réel un nombre plus ou moins important de paramètres qui caractérisent un ou plusieurs processus écologiques (par ex. limitation de dispersion, dépendance spatiale induite par des covariables environnementales, voir Fig. 2). Une fois ces paramètres estimés, ceux-ci peuvent être utilisés pour générer des patterns de points nuls contraints par la seule influence du ou des processus considérés. Des statistiques descriptives synthétiques (*summary statistics*, SS) – telles que la statistique *K* de Ripley (1981) ou la *pair correlation function* (Wiegand and Moloney 2014) – permettent de caractériser les semis de points au sein d'une gamme de distances. Ces statistiques sont calculées pour le pattern réel ainsi que pour les patterns simulés selon différents processus, ce qui permet de générer une enveloppe de simulations pour chaque SS et de comparer les SS à ces enveloppes. Une SS observée qui sort de ces enveloppes indique alors qu'un processus supplémentaire affecte le pattern de point réel. Cette démarche permet d'estimer de façon spatialement explicite à quelle gamme de distance le semi de point étudié dévie d'un modèle intégrant un ou une combinaison de plusieurs processus (Baddeley et al. 2015, Velázquez et al. 2016). L'approche du SPPA peut être utilisée pour un pattern d'un seul type de points mais peut également être utilisé pour des patterns de points « marqués » (par ex. différentes espèces) afin d'étudier les relations spatiales des points de plusieurs types. Ainsi, l'étude de patterns de points bivariés (par ex. Fig. 3 pour un exemple à deux espèces) de toutes les paires d'espèces d'arbres de plusieurs forêts combinée à l'utilisation de traits fonctionnels a récemment permis de mettre en évidence des patterns fins expliqués par de la différentiation de niche (corrélation spatiale positive entre espèces présentant des traits dissimilaires) et des patterns d'échelle relativement plus large expliqués par un filtre de l'habitat (corrélation spatiale positive entre espèces présentant des traits similaires) (Wiegand et al. 2007, 2012, Velázquez et al. 2015). Cette dernière approche sera utilisée en parallèle d'une approche basée sur les vecteurs propres spatiaux dans le chapitre VI afin de tester la présence de différentiation de niche ainsi que d'un filtre de l'habitat de façon spatialement explicite dans une jeune forêt tempérée de recolonisation.

Objectifs de la thèse

Cette thèse en écologie des communautés s'inscrit dans le cadre de l'inférence de mécanismes écologiques d'assemblage à partir d'une détection précise de patterns spatiaux multi-échelles.

Elle a deux buts principaux :

1) l'évaluation et l'amélioration des analyses spatiales basées sur les vecteurs propres et la partition de variation, dans le cadre des approches corrélatives des mécanismes sous-tendant l'assemblage des communautés,
2) une compréhension – par le biais de ces méthodes spatialement explicites – de l'influence de l'environnement sur les mécanismes de structurations des communautés d'arbres.

Dans le **Chapitre I**, une caractérisation détaillée des propriétés physiques et chimiques du sol sera mise en relation avec un inventaire exhaustif de la communauté des arbres adultes d'une forêt claire tropicale (Miombo) afin de définir si différents habitats peuvent être définis à une échelle locale (< 10 ha). En outre, ce chapitre visera à tester la présence d'espèces d'arbres indicatrices de ces habitats.

Le **Chapitre II** visera à mettre en évidence les déterminants environnementaux de l'assemblage d'une communauté de champignons ectomycorhiziens dans la même forêt de type Miombo. Les variables explicatives utilisées incluront de nombreuses variables de sol ainsi que des *proxy* des stratégies fonctionnelles de la communauté des arbres hôtes (traits fonctionnels). Les hypothèses testées seront que cette communauté fongique est spatialement structurée par l'hétérogénéité spatiale des variables de sol ainsi que par une spécialisation d'association liée aux stratégies fonctionnelles des arbres hôtes. Ce chapitre visera également à définir l'importance relative des processus de la niche et des processus neutres dans la structuration de la communauté fongique symbiotique. Alors que le Chapitre I n'utilise pas encore d'approche spatialement explicite, le Chapitre II illustrera le cadre d'analyse et d'interprétation écologique des analyses spatialement explicites basées sur l'utilisation de vecteurs propres spatiaux et de la partition de variation.

Le **Chapitre III** débute une partie de la thèse consacrée aux développements méthodologiques qui se poursuivront aux Chapitres IV et V. Le Chapitre III aura pour objectif de synthétiser les procédures de sélection de prédicteurs spatiaux dans les méthodes spatiales basées sur les vecteurs propres. Alors qu'il a été suggéré dans la littérature qu'une sélection inadaptée de prédicteurs spatiaux peut mener à des résultats biaisés, aucun consensus n'existe encore à ce stade parmi les nombreuses pratiques de sélection disponibles. Les trois procédures les plus utilisées depuis l'introduction des MEM (2006) et fin 2016 seront comparées par le biais de simulations afin de déterminer leurs performances statistiques (taux d'erreur de type I, puissance et précision de détection de patterns spatiaux) dans une gamme de situations réalistes. Ce chapitre visera *in fine* à produire une suite de recommandations, ou « bonnes pratiques » pour une utilisation non biaisée et optimale des vecteurs propres spatiaux.

Le **Chapitre IV** abordera la question du choix d'une matrice de pondération spatiale à partir de laquelle construire les prédicteurs spatiaux. Alors que le Chapitre III aura abordé la question de la sélection de prédicteurs spatiaux au sein d'une matrice de pondération spatiale donnée, le Chapitre IV s'intéressera à l'impact qu'a en amont la nature de la matrice de pondération spatiale choisie pour construire les prédicteurs spatiaux. Des scénarios de simulations variés reflettant diverses situations réalistes seront utilisés afin de comparer les performances statistiques d'une large gamme de matrices de pondération spatiale. Une nouvelle méthode d'optimisation de sélection de la matrice de pondération spatiale sera proposée. Les performances statistiques de cette nouvelle méthode seront évaluées et comparées à celles des pratiques les plus couramment utilisées depuis 2006.

Dans le **Chapitre V**, l'objectif est de présenter une nouvelle procédure permettant pour la première fois de tester la fraction de la partition de variation correspondant à la variation de composition d'une communauté expliquée conjointement par les variables environnementales et spatiales (fraction [b] de la Fig. 5). Cette fraction – signature potentielle de l'influence d'une structure spatiale de l'environnement sur la communauté (ISD) – ne peut être testée par les approches classiques et a donc jusqu'à présent été interprétée difficilement parce que n'étant pas appuyée statistiquement. Une étude de simulation sera utilisée afin de tester la capacité de cette nouvelle procédure (1) à détecter une ISD lorsque ce processus a bien été utilisé pour générer les données (puissance statistique) et (2) à ne pas détecter d'ISD lorsque l'espèce ou la communauté n'a pas de structure spatiale liée à l'environnement (taux d'erreur de type I).

Le **Chapitre VI** intègrera les avancées méthodologiques des Chapitres III à V dans un cas d'étude visant à déterminer les processus écologiques d'assemblage d'une communauté d'arbres d'une forêt tempérée. Les variables explicatives utilisées seront des variables de sol, topographiques, d'intensité lumineuse ainsi que de traits fonctionnels de croissance et d'acquisition des ressources. Ce chapitre visera également à illustrer l'intérêt de l'utilisation combinée d'approches analytiques spatialement complémentaires dans un but de compréhension plus fine des processus d'assemblage des communautés. Pour ce faire, une communauté d'arbres sera analysée au travers des méthodes des chapitres précédents et au travers d'une analyse de semis de points. Alors que la première méthode aura pour objectif la détection des patterns de composition spécifique à des échelles larges à relativement fines, l'analyse de semis de points permettra d'étudier les processus associés à des patterns spatiaux à l'échelle du voisinage direct des individus.

Références

Anderson, M. J., and P. Legendre. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. Journal of Statistical Computation and Simulation 62:271–303.

Baddeley, A., E. Rubak, and R. Turner. 2015. Spatial point patterns: methodology and applications with R. Chapman & Hall/CRC.

Bauman, D., O. Raspé, P. Meerts, J. Degreef, J. Ilunga Muledi, and T. Drouet. 2016. Multiscale assemblage of an ectomycorrhizal fungal community: the influence of host functional traits and soil properties in a 10-ha miombo forest. FEMS Microbiology Ecology 92:fiw151.

Bell, G., M.J. Lechowicz, and M.J. Waterway. 2006. Neutral interpretations of biological communities. Ecology 87: 1378–1386.

de Bello, F., P. Šmilauer, J. A. F. Diniz-Filho, C. P. Carmona, Z. Lososová, T. Herben et al. 2017.

Decoupling phylogenetic and functional diversity to reveal hidden signals in community assembly. Methods in Ecology and Evolution 8:1200–1211.

Bini, L. M., J. A. F. Diniz-filho, T. F. L. V. B. Rangel, T. S. B. Akre, R. G. Albaladejo, F. S. Albuquerque et al. 2009. Coefficient shifts in geographical ecology: an empirical evaluation of spatial and non-spatial regression. Ecography 32:193–204.

Blanchet, F. G., P. Legendre, and D. Borcard. 2008. Forward selection of explanatory variables. Ecology 89:2623–2632.

Borcard, D., F. Gillet, and P. Legendre. 2011. Numerical Ecology with R. Springer. Springer, New-York.

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

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

Borcard, D., P. Legendre, C. Avois-Jacquet, and H. Tuomisto. 2004. Dissecting the spatial structure of ecological data at multiple scales. Ecology 85:1826–1832.

Bray, R. J., and J. T. Curtis. 1957. An ordination of the upland forest communities of southern Winsconin. Ecological Monographs 27:325–349.

Brown, B. L., E. R. Sokol, J. Skelton, and B. Tornwall. 2016. Making sense of metacommunities: dispelling the mythology of a metacommunity typology. Oecologia 183:643–652.

Cabral, J. S., L. Valente, and F. Hartig. 2017. Mechanistic simulation models in macroecology and biogeography: state-of-art and prospects. Ecography 40:267–280.

De Cáceres, M., P. Legendre, R. Valencia, M. Cao, L. W. Chang, G. Chuyong et al. 2012. The variation of tree beta diversity across a global network of forest plots. Global Ecology and Biogeography 21:1191–1202.

Cale, W. G., G. M. Henebry, and J. A. Yeakley. 1989. Inferring process from pattern in natural communities. BioScience 39:600–605.

Chang, L. W., D. Zelený, C. F. Li, S. T. Chiu, and C. F. Hsieh. 2013. Better environmental data may reverse conclusions about niche- and dispersal-based processes in community assembly. Ecology 94:2145–2151.

Chase, J. M. 2014. Spatial scale resolves the niche versus neutral theory debate. Journal of Vegetation Science 25:319–322.

Chase, J. M., and J. A. Myers. 2011. Disentangling the importance of ecological niches from stochastic processes across scales. Philosophical Transactions of the Royal Society B: Biological Sciences 366:2351–2363.

Chesson, P. 2000. Mechanisms of maintenance of species diversity. Annual Review of Ecology and Systematics 31:343–366.

Clark, J. S., S. Ladeau, and I. Ibanez. 2004. Fecundity of Trees and the Colonization–Competition Hypothesis. Ecological Monographs 74:415–442.

Cornwell, W. K., and D. D. Ackerly. 2009. Community assembly and shifts in plant trait distributions across an environmental gradient in coastal California. Ecological Monographs 79:109–126.

Cottenie, K. 2005. Integrating environmental and spatial processes in ecological community dynamics. Ecology Letters 8:1175–1182.

Diggle, P. J. 2003. Statistical analysis of spatial point patterns. Edward Arnold, London.

Diniz-Filho, J. A. F., and L. M. Bini. 2005. Modelling geographical patterns in species richness using eigenvector-based spatial filters. Global Ecology and Biogeography 14:177–185.

Diniz-Filho, J. A. F., L. M. Bini, T. F. Rangel, I. Morales-Castilla, M. Á. Olalla-Tárraga, M. Á. Rodríguez et al. 2012. On the selection of phylogenetic eigenvectors for ecological analyses. Ecography 35:239–249.

Diniz-Filho, J. A. F., C. E. R. de Sant'Ana, and L. M. Bini. 1998. An eigenvector method for estimating phylogenetic inertia. Evolution 52:1247–1262.

Dormann, C. F., J. M. Mcpherson, M. B. Araújo, R. Bivand, J. Bolliger, G. Carl et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30:609–628.

Dormann, C. F., S. J. Schymanski, J. Cabral, I. Chuine, C. Graham, F. Hartig et al. 2012. Correlation and process in species distribution models: Bridging a dichotomy. Journal of Biogeography 39:2119–2131.

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

Dray, S., R. Pélissier, P. Couteron, M.-J. Fortin, P. Legendre, P. R. Peres-Neto et al. 2012. Community ecology in the age of multivariate multiscale spatial analysis. Ecological Monographs 82:257–275.

Ellison, A. M. 2004. Bayesian inference in ecology. Ecology Letters 7:509–520.

Ezekiel, M. 1929. The application of the theory of error to multiple and curvilinear correlation. Journal of the American Statistical Association 24:99–104.

Florax, R. J., and S. Rey. 1995. The impacts of misspecified spatial interaction in linear regression models. Pages 111–135 New directions in spatial econometrics. Springer, Berlin, Heidelberg.

Fortin, M., and M. R. T. Dale. 2014. Spatial analysis: A guide for ecologists. Cambridge. Cambridge.

Garnier, E., M.-L. Navas, and K. Grigulis. 2016. Plant function diversity: Organism traits, community structure, and ecosystem properties. Oxford University Press.

Garzon-Lopez, C. X., P. A. Jansen, S. A. Bohlman, A. Ordoñez, and H. Olff. 2014. Effects of sampling scale on patterns of habitat association in tropical trees. Journal of Vegetation Science 25:349–362.

Gauch, H. G. 1982. Multivariate analysis in community ecology. Cambridge University Press.

Getis, A., and D. A. Griffith. 2002. Comparative spatial filtering in regression analysis. Geographical Analysis 34:130–140.

Griffith, D. 2003. Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization. Springer, Berlin.

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

Griffith, D. A. 2017. Spatial Weights. Pages 1–7 in D. Richardson, N. Castree, M. F. Goodchild, A. Kobayashi, W. Liu, and R. A. Marston, editors. International Encyclopedia of Geography: People, the Earth, Environment and Technology. Wiley-Blackwell.

Griffith, D. A., and F. Lagona. 1998. On the quality of likelihood-based estimators in spatial autoregressive models when the data dependence structure is misspecified. Journal of Statistical Planning and Inference 69:153–174.

Griffith, D. A., and P. R. Peres-Neto. 2006. Spatial modeling in Ecology: the flexibility of eigenfunction spatial analyses. Ecology 87:2603–2613.

Grime, J. P. 2006. Trait convergence and trait divergence in herbaceous plant communities: Mechanisms and consequences. Journal of Vegetation Science 17:255–260.

Guénard, G., P. Legendre, and P. Peres-Neto. 2013. Phylogenetic eigenvector maps: A framework to model and predict species traits. Methods in Ecology and Evolution 4:1120–1131.

HilleRisLambers, J., P. B. Adler, W. S. Harpole, J. M. Levine, and M. M. Mayfield. 2012. Rethinking community assembly through the lens of coexistence theory. Annual Review of Ecology, Evolution, and Systematics 43:227–248.

Hubbell, S. P. 1997. A unified theory of biogeography and relative species abundance and its application to tropical rain forests and coral reefs. Coral Reefs 16:S9–S21.

Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography. Princeton University Press, Princeton, New Jersey.

Hubbel, S.P. 2006. Neutral theory and the evolution of ecological equivalence. Ecology 87:1387–1398.

Hutchinson, G. E. 1959. Homage to Santa Rosalia or why are there so many kinds of animals? The American Naturalist 93:145–159.

Hyatt, L. A., M. S. Rosenberg, T. G. Howard, G. Bole, W. Fang, J. Anastasia et al. 2003. The Distance Dependence Prediction of the Janzen-Connell Hypothesis: A Meta-Analysis. Oikos 103:590–602.

Illian, J. 2008. Statistical analysis and modelling of spatial point patterns. John Wiley & Sons.

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

Jombart, T., S. Dray, and A. B. Dufour. 2009. Finding essential scales of spatial variation in ecological data: A multivariate approach. Ecography 32:161–168.

Jones, M. M., H. Tuomisto, D. Borcard, P. Legendre, D. B. Clark, and P. C. Olivas. 2008. Explaining variation in tropical plant community composition: Influence of environmental and spatial data quality. Oecologia 155:593–604.

Kostov, P. 2010. Model boosting for spatial weighting matrix selection in spatial lag models. Environment and Planning B: Planning and Design 37:533–549.

Kraft, N. J. B., P. B. Adler, O. Godoy, E. C. James, S. Fuller, and J. M. Levine. 2015. Community assembly, coexistence and the environmental filtering metaphor. Functional Ecology 29:592–599.

Legendre, P. 1990. Quantitative methods and biogeographic analysis. Page Evolutionary biogeography of the marine algae of the North Atlantic. Springer, Berlin, Heidelberg.

Legendre, P. 1993. Spatial autocorrelation: Trouble or new paradigm? Ecology 74:1659–1673.

Legendre, P., and M. J. Fortin. 1989. Spatial pattern and ecological analysis. Vegetatio 80:107–138.

Legendre, P., and L. Legendre. 2012. Numerical Ecology. Elsevier, Amsterdam.

Legendre, P., X. Mi, H. Ren, K. Ma, M. Yu, I.-F. Sun et al. 2009. Partitioning beta diversity in a subtropical broad-leaved forest of China. Ecology 90:663–674.

Leibold, M. A., and M. A. McPeek. 2006. Coexistance of the Niche and Neutral Perspectives in Community Ecology. Ecology 87:1399–1410.

Lowe, W. H., and M. A. McPeek. 2014. Is dispersal neutral? Trends in Ecology and Evolution 29:444–450.

MacArthur, R., and R. Levins. 1967. The Limiting Similarity, Convergence, and Divergence of Coexisting Species. The American Naturalist 101:377.

McIntire, E. J. B., and A. Fajardo. 2009. Beyond description: the active and effective way to infer processes from spatial patterns. Ecology 90:46–56.

McIntire, E. J. B., and D. S. Hik. 2005. Influences of chronic and current season grazing by collared pikas on above-ground biomass and species richness in subarctic alpine meadows. Oecologia 145:288–297.

Mcintire, E. J. B., C. B. Schultz, and E. E. Crone. 2007. Designing a network for butterfly habitat restoration: Where individuals, populations and landscapes interact. Journal of Applied Ecology 44:725–736.

Muledi, J. I., D. Bauman, T. Drouet, J. Vleminckx, A. Jacobs, J. Lejoly et al. 2017. Fine-scale habitats influence tree species assemblage in a miombo forest. Journal of Plant Ecology 10:doi:10.1093/jpe/rtw104.

Munoz, F. 2009. Distance-based eigenvector maps (DBEM) to analyse metapopulation structure with irregular sampling. Ecological Modelling 220:2683–2689.

Munoz, F., M. Grenié, P. Denelle, A. Taudière, F. Laroche, C. Tucker et al. 2018. ecolottery: Simulating and assessing community assembly with environmental filtering and neutral dynamics in R. Methods in Ecology and Evolution 9:693–703.

Ogle, K., R. L. Wolpert, and J. F. Reynolds. 2004. Reconstructing plant root area and water uptake profiles. Ecology 85:1967–1978.

Patuelli, R., D. A. Griffith, M. Tiefelsdorf, and P. Nijkamp. 2011. The use of spatial filtering techniques: the spatial and space-time structure of German unemployment data. International Regional Science Review 34:253–280.

Peay, K. G., S. E. Russo, K. L. McGuire, Z. Lim, J. P. Chan, S. Tan et al. 2015. Lack of host specificity leads to independent assortment of dipterocarps and ectomycorrhizal fungi across a soil fertility gradient. Ecology Letters 18:807–816.

Peres-Neto, P. R., and P. Legendre. 2010. Estimating and controlling for spatial structure in the study of ecological communities. Global Ecology and Biogeography 19:174–184.

Perry, G. L. W., B. P. Miller, and N. J. Enright. 2006. A comparison of methods for the statistical analysis of spatial point patterns in plant ecology. Plant Ecology 187:59–82.

Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research. Sankhya A26:329–358.

Rees, M., P. J. Grubb, and D. Kelly. 1996. Quantifying the impact of competition and spatial heterogeneity on the structure and dynamics of a four-species guild of winter annuals. The American Naturalist 147:1–32.

Ripley, B. D. 1981. Spatial statistics. John Wiley and Sons.

Schultz, C. B., and E. E. Crone. 2001. Edge-mediated dispersal behavior in a prairie butterfly. Ecology 82:1879–1892.

Schurr, F. M., J. Pagel, J. S. Cabral, J. Groeneveld, O. Bykova, R. B. O'Hara et al. 2012. How to understand species' niches and range dynamics: A demographic research agenda for biogeography. Journal of Biogeography 39:2146–2162.

Seidler, T. G., and J. B. Plotkin. 2006. Seed dispersal and spatial pattern in tropical trees. PLoS Biology 4:2132–2137.

Šidák, Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 62:626–633.

Simberloff, D., and D. Tamar. 1991. The guild concept and the structure of ecological communities. Annual Review of Ecological Systematics 22:115–43.

Stakhovych, S., and T. H. A. Bijmolt. 2008. Specification of spatial models: A simulation study on weights matrices. Papers in Regional Science 88:389–408.

Stetzer, F. 1982. Specifying weights in spatial forecasting models: The results of some experiments. Environment and Planning A 14:571–584.

Stoll, P., and D. Prati. 2001. Intraspecific aggregation alters competitive interactions in experimental plant communities. Ecology 82:319–327.

Tiefelsdorf, M., and D. A. Griffith. 2007. Semiparametric filtering of spatial autocorrelation: the eigenvector approach. Environment and Planning A 39:1193–1222.

Turner, M. G., R. H. Gardner, and R. V. O'Neill. 2001. Landscape ecology in theory and practice: pattern and proces. EUA: Springer.

Velázquez, E., I. Martínez, S. Getzin, K. A. Moloney, and T. Wiegand. 2016. An evaluation of the state of spatial point pattern analysis in ecology. Ecography 39:1042–1055.

Velázquez, E., C. E. T. Paine, F. May, and T. Wiegand. 2015. Linking trait similarity to interspecific spatial associations in a moist tropical forest. Journal of Vegetation Science 26:1068–1079.

Vellend, M. 2010. Conceptual synthesis in community ecology. The Quarterly Review of Biology 85:183–206.

Violle, C., M.-L. Navas, D. Vile, E. Kazakou, C. Fortunel, I. Hummel et al. 2007. Let the concept of trait be functional! Oikos 116:882–892.

Vleminckx, J., J.-L. Doucet, J. Morin-Rivat, A. B. Biwolé, D. Bauman, O. J. Hardy et al. 2017. The influence of spatially structured soil properties on tree community assemblages at a landscape scale in the tropical forests of southern Cameroon. Journal of Ecology 105:354–366.

Wagner, H. H., and S. Dray. 2015. Generating spatially constrained null models for irregularly spaced data using Moran spectral randomization methods. Methods in Ecology and Evolution 6:1169–1178.

Weiher, E., and P. A. Keddy. 1995. Assembly rules, null models, and trait dispersion: new questions from old patterns. Oikos 74:159–164.

Whittaker, R. J. 1956. Vegetation of the Great Smoky Mountains. Ecological Monographs 26:1–80.

Wiegand, T., S. Gunatilleke, and N. Gunatilleke. 2007. Species Associations in a Heterogeneous Sri Lankan Dipterocarp Forest. The American Naturalist 170:E77–E95.

Wiegand, T., A. Huth, S. Getzin, X. Wang, Z. Hao, C. V. S. Gunatilleke et al. 2012. Testing the independent species' arrangement assertion made by theories of stochastic geometry of biodiversity. Proceedings of the Royal Society B: Biological Sciences 279:3312–3320.

Wiegand, T., F. Jeltsch, I. Hanski, and V. Grimm. 2003. Using pattern-oriented modeling for revealing hidden information: a key for reconciling ecological theory and application. Oikos 65:209–222.

Wiegand, T., and K. A. Moloney. 2014. Handbook of spatial point pattern analysis in ecology. Chapman & Hall/CRC, Boca Raton.

Yamazaki, M., S. Iwamoto, and K. Seivva. 2009. Distance- and density-dependent seedling mortality caused by several diseases in eight tree species co-occurring in a temperate forest. Plant Ecology 201:181–196.

Chapitre I

Fine-scale habitats influence tree species assemblage in a Miombo forest

Jonathan Ilunga Muledi*, David Bauman*, Thomas Drouet, Jason Vleminckx, Arnaud Jacobs, Jean Lejoly, Pierre Meerts, and Mylor Ngoy Shutcha

** Equally contributing authors*

Published September 2016 in *Journal of Plant Ecology*, vol. 10(6), pp. 958-969

Dans ce premier chapitre, un inventaire exhaustif de la communauté adulte des arbres d'une forêt claire tropicale (Miombo) sera mis en relation avec une caractérisation détaillée des propriétés physiques et chimiques du sol. L'objectif sera ici de chercher à définir si différents habitats sont à la base d'une différentiation significative de composition de la communauté des arbres. Nous tenterons également de définir les espèces indicatrices de ces habitats sur des bases statistiques. Contrairement aux échelles régionales généralement bien plus larges étudiées pour considérer des changements de composition de la communauté des arbres de cette formation végétale, c'est à une échelle locale que nous nous intéresserons ici et donc au rôle de l'hétérogénéité environnementale d'échelle spatiale fine.

Fine-scale habitats influence tree species assemblage in a Miombo forest

Abstract

Aims: Relationships between local habitat heterogeneity and tree communities in miombo woodlands have been very little studied. While some studies have addressed this topic at broad scales and based on few environmental parameters, this study aims at 1) detecting fine-scale habitats (≤ 10 ha) on the basis of a detailed characterisation of soil explicitly considering past anthropogenic disturbances, and an exhaustive census of the tree community, and at 2) searching for indicator tree species corresponding to the resulting habitats.

Methods: The study was carried out in the miombo woodland of Mikembo Forest Reserve, Upper Katanga, The Democratic Republic of the Congo. A complete census of the tree community was conducted in a 10-ha forest dynamics plot comprising 160 adjacent quadrats of 25 × 25 m, with a total of 4604 trees (DBH > 10 cm). Thirty-six physicochemical soil parameters were measured. Studying the frequency distribution of soil charcoal content allowed identifying local signature of past human agriculture in the soil. Two strategies were used to define habitats: 1) a combination of PCA on soil variables and Ward clustering, and 2) Multivariate Regression Trees (MRT) to search for key soil parameters allowing the best prediction of species composition. Tree-habitat associations were tested by means of a robust statistical framework combining the IndVal index and torus randomisations.

Important findings: The forest contained 82 tree species and a significant proportion of wet miombo species (e.g., *Marquesia macroura*). We detected a strong east-west edaphic gradient driven by soil texture; most chemical soil parameters followed this pattern. Five habitats were identified based on soil factors and floristic composition. Nine indicator species of these habitats were found. The key soil factors discriminating habitats were total calcium, available forms of phosphorus, and clay content. Even though past agricultural practices were successfully detected in soils, they did not display any significant influence neither on habitat differentiation nor on the associated tree communities. Based on an unprecedented large number of soil parameters, fine-scale soil heterogeneity and niche partitioning were shown to contribute to the variability of the floristic composition in this forest. Our results indicated that considering the most variable environmental parameters, as in PCA, is a poor manner for defining habitats. In contrast, combining MRT with the IndVal index and torus randomisation has proved to be a much more robust and sensitive approach to highlight tree-habitat associations at this scale. The common dichotomous viewpoint of considering deterministic and neutral effects as acting at broad and fine scales, respectively, is not confirmed when measuring suitable environmental variables, even in a case where the physical environment does not exhibit strong heterogeneity.

Keywords

Forest dynamics plot, Indicator species, Miombo, Multivariate regression trees (MRT), Soil, Torus randomization

I. Introduction

Understanding the mechanisms structuring tree species distribution in tropical forests is a challenging issue in community ecology (Legendre *et al.* 2009; de Oliveira *et al.* 2014; Vleminckx *et al.* 2015). On one hand, niche differentiation allows, to a certain extent, to predict the species composition of a given community on the basis of measurable environmental parameters (Legendre and Legendre 2012). On the other hand, neutral processes (dispersion limitation, ecological drift) also influence continually the community in an unpredictable manner (Hubbell 2005; Nguyen *et al.* 2016). The current consensus among ecologists is that both deterministic and neutral processes act together to shape living communities (Chase 2014; Velázquez *et al.* 2015). So far, deterministic processes acting on tree communities have been associated to broad scales, related to large environmental gradients, while neutral processes are often considered as a matter of fine scales presenting relative homogeneity of edaphic or climatic parameters (Borcard *et al.* 2011). Although an influence of deterministic processes is deemed able to occur at fine scales in some cases (Legendre and Legendre 2012), such fine-scale environmental heterogeneity influence on tree species community has not been thoroughly investigated.

Miombo woodlands are the most common savanna type in the southern hemisphere, covering ca. 2.7 million km^2 (i.e., 10% of the African continent) (Millington *et al.* 1994). The miombo is a semi-deciduous formation, with a tree layer characterised by the abundance of three genera of Fabaceae (subfamily Caesalpinioideae): *Brachystegia*, *Julbernardia*, and *Isoberlinia* (White 1983; Campbell 1996). The miombo plays an important role in the regulation of regional climate (Malmer and Nyberg 2008), carbon sequestration (Williams *et al.* 2007; Zahabu 2008), and the conservation of soil and water resources. These forests, although they occupy a larger area than tropical African rainforest (Campbell 1996), have received comparatively little attention and the deterministic processes influencing their woody species distribution remain mainly unknown.

Forest dynamics plots are essential tools to monitor forest composition and dynamics, and to unravel the mechanisms promoting biodiversity and the coexistence of species (e.g., habitat heterogeneity, historical events), and design management strategies (Legendre *et al.* 2009). The structure and dynamics of miombo woodlands have received much attention in the Zambezian region (Campbell *et al.* 2007). However, this miombo is extremely variable in terms of floristic composition and comprises regions of dry and wet miombo (Chidumayo 1987; Kanschik and Becker 2001; Backéus *et al.* 2006). The structure and functioning of the wet miombo has been relatively little studied in comparison with the dry miombo. In order to fill that gap of knowledge, the first 10-ha forest dynamics plot was installed in 2009 in Upper Katanga (DRC) in order to monitor growth and dynamics of the tree community. Earlier work on the miombo revealed extensive variation in floristic composition at the landscape scale, in part accounted for by variation in soil factors (Duvigneaud 1958; Sys and Schmitz 1959; Schmitz 1971; Munishi *et al.* 2011; Mwakalukwa *et al.* 2014). Duvigneaud (1958) explored soil-vegetation relationships in the miombo with the methods of geobotany using topo-lithological transects and proposed a system of ecological groups of indicator species in relation to soil drainage, depth, and texture. He recognised four main types of miombo forests (i.e., plateau miombo in deep soil with *Brachystegia longifolia*, *Brachystegia spiciformis, Erythrophleum africanum*, miombo on slopes with compact gravelly yellow soil, with *Brachystegia utilis*, miombo on shallow rocky soil with *Brachystegia microphylla* and *Brachystegia bussei*, miombo on poorly drained lateritic crust, with *Isoberlinia tomentosa* and *Brachystegia stipulata*). Based on the methods of Zürich-Montpellier, Schmitz (1971) published a phytosociological survey recognising three alliances (i.e., Berlinio-Marquesion (semi-evergreen miombo), Mesobrachystegion (mesic, usually deep soil), and Xerobrachystegion (shallow, dry stony soil)).

Until now, no study using modern statistical methods has investigated soil-vegetation relationships in the miombo of Katanga. Specifically, we investigate the existence of indicator tree species and species assemblages characterising habitats resulting from fine-scale soil heterogeneity. Since several studies highlighted the long-lasting influence of human activities on vegetation distribution patterns (Stroomgaard 1991; Van Gemerden *et al.* 2003; Vleminckx *et al.* 2014), we also explicitly consider the effect of past human agricultural activities as potential driver of habitat and tree community differentiation. Detecting precise relationships between species and their habitat preferences allows establishing precise locations for species plantations (Dray *et al.* 2012). In addition, the detection of indicator species has been shown to be a requisite tool in the field of nature monitoring, conservation, and management, and is a more robust method of assessing ecologically meaningful habitats than the use of diversity indices (Dufrêne and Legendre 1997).

So far, the reported species assemblage observations have generally been studied at large scales in frameworks implicitly acknowledging miombo forests as relatively large homogeneous units (White 1983; Campbell 1996). Therefore, we addressed the following specific questions: 1) Is soil heterogeneity at fine scale (< 10 ha) sufficient to characterise different habitats? 2) Can we define species assemblages related to these fine-scale habitats? 3) Are some tree species indicators of the defined habitats? 4) Did past agricultural activities influence the soil parameters and the forest species composition?

II. Material and methods

II.1. Study site

The study was undertaken in the Mikembo Forest Reserve (11°28'57" to 11°29'5"S, 27°40'12" to 27°40'28"E, ~1 200 m above sea level), an 800-ha private nature protection area located in Upper Katanga, about 35 km northeast of Lubumbashi, the Democratic Republic of the Congo (Fig. 1). The mean annual temperature is 20.3 °C, and the average annual precipitation is 1200 mm, occurring mainly from November to March or April. The climate is Sudanian, corresponding to Cwa in Köppen's classification (Peel *et al.* 2007). Upper Katanga, also referred to as southern Katanga, represents the northern part of the Zambezian centre of endemism (White 1983; Malaisse 1996). The eastern part of Upper Katanga belongs in the Katango-Zambian sector (Duvigneaud 1958; Werger and Coetzee 1978; Malaisse 1996). The landscape showed a flat topography, but regularly punctuated (~3/ha) by termite mounds (up to 8 m high). The forest is located on geological substrates dominated by dolomitic shales and siltstones from Neoproterozoic Nguba and Roan Groups (Batumike *et al.* 2006). Soils of this region are mostly haplic and xanthic Ferralsols characterised by low pH and nutrient content and a sandy loam to clay loam texture (Baert *et al.* 2009). The vegetation is a mixed tropical dry season woodland with a mean canopy height of 14 m and basal area of ~20 m^2 ha^{-1}. The reserve was established in 2003, and since then, it has been protected from fire practices and fuel-wood cutting.

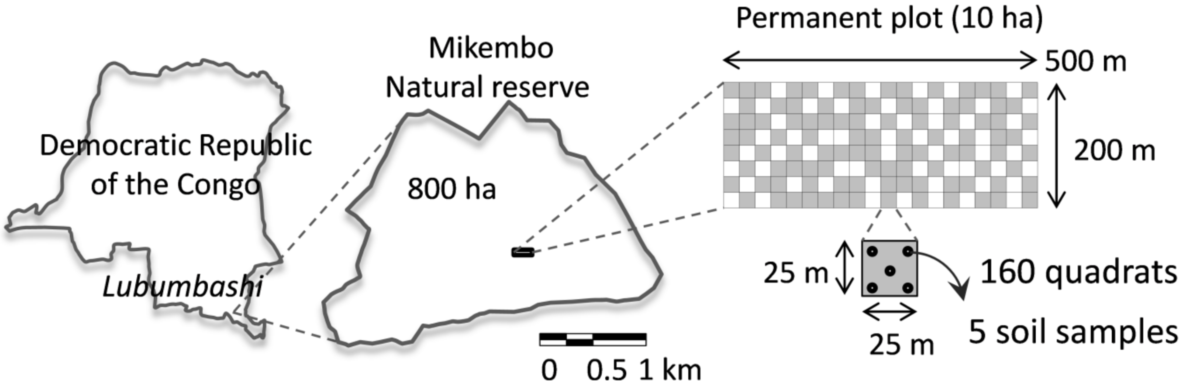


Figure 1: Location of the study site at the Mikembo Reserve (Upper Katanga, DRC), schematic representation of the dynamic permanent plot and of a sampled quadrat. Sampled quadrats are in grey, soil samples location are filled circles.

II.2. Inventory design

In 2009, a forest dynamics plot of 10 ha (200 m × 500 m) divided into 160 quadrats of 25 × 25 m was installed (Fig. 1). The area was systematically inventoried for all living and dead trees ≥ 10 cm diameter at 130 cm (diameter at breast height, DBH) following Picard and Gourlet-Fleury (2008). Trees were tagged, mapped, identified to the species level, and measured for DBH in 2014. Heights of the 25 largest trees were measured using a clinometer (Suunto Co., Finland) to achieve 100 trees measured per ha as recommended by Rondeux (1993). Family (APG III, 2009), genus, and species were determined for all trees in the experimental plot following Meerts (2016). Thirty-tree termite mounds were present in the dynamic plot but were not considered in this study (neither their tree community nor their soil parameters) due to their completely distinct floristic assemblage and ecological conditions. The identification of tree species was completed using Flora Zambeziaca and the Flore d'Afrique Centrale.

II.3. Soil sampling and analyses

Soil sampling was performed in 102 randomly chosen quadrats of the 160 of the dynamic forest plot (Fig. 1), with the constraint of selecting quadrats comprising no termite mound. In each of these quadrats, five soil cores were collected (four at five metres from the corners, one in the centre) at 0-20 cm depth. A total of 36 soil variables were determined on these samples according to conventional protocols (Pansu and Gautheyrou 2006). For each soil sample, the stoniness index (*SI*) was estimated on the field after sieving (2-mm mesh) by a discreet quantitative index taking the values of 0 (no gravels), 1 (half of the sieve area or less covered by gravels) or 2 (all the sieve area covered). Undisturbed soil cores were taken for bulk density (*BD*) measurements (cylinder method). Soil texture (clay, silt and sand) was determined by wet sieving and the pipette method after OM destruction with H_2O_2 and clay dispersion by Na citrate. The pH-H_2O and the electrical conductivity (*EC*) were respectively measured with glass electrodes (Mettler-Toledo) and a conductimeter (VWR EC300) on a 1:5 soil:deionised water suspension. The pH-KCl and exchangeable Al (Al_{exch}) were determined on a soil suspension (1:5 ratio) of 1 M KCl and measuring the derivative of the titration curves for $\mathrm{Al}_{\mathrm{exch}}$ (Radiometer Copenhagen TIM900). The ΔpH was obtained by calculating the difference between pH-KCl and pH-H_2O. The plant-available elements ((*Ca*, *Mg*, *K*, *Al*, *Fe*, *Mn*, *B*, and *Zn*)$_{\mathrm{avail}}$) were extracted with 0.5 M ammonium acetate 0.03 M EDTA at pH 4.65 and measured by inductively coupled plasma optical emission spectroscopy (ICP-OES) with CCD detector (Varian, Vista MPX). Bioavailable phosphorus (P_{Olsen}) was extracted with Na bicarbonate and determined by colorimetry, a second form (P_{EDTA}) was extracted with ammonium acetate EDTA and measured by ICP-OES. Total forms of elements ((*Ca*, *Mg*, *K*, *Al*, *Fe*, *Mn*, *B*, *Mo*, *P*, and *Zn*)$_{tot}$) were taken in solution by complete dissolution of finely ground soil samples by a tri-acid attack (HCl-HNO_3-HF) in Teflon vials on a hot plate. The dry residue was re-dissolved in HNO_3 and total element concentrations were determined by ICP-OES. The effective cation exchange capacity (*CEC*) was calculated as the sum of exchangeable Ca, K, Mg concentrations and titrated Al (Al_{exch}), expressed in $\mathrm{cmol_c\ kg^{-1}}$. The Al saturation rate (*Al-Sat*) of the exchange complex corresponds to the proportion of Al^{3+} on the total *CEC*. The soil carbon-to-nitrogen ratio (*C/N*) was computed after measuring soil nitrogen and carbon contents by flash combustion at 1350 °C in a CN elemental analyser (Dumas method, ISO 10694). Organic matter content (*OM*) was calculated by mass loss of a sample after dry ashing at 550 °C. The extinction coefficient in visible light (*E4/E6*) allows to know the relative importance of humic and fulvic acids in the soil OM and is related to the humification stage. This coefficient was obtained measuring the absorbance at 465 and 665 nm of a soil extract with 0.5 M NaOH during 16 h after centrifugation (10000 rnd/min). The soil charcoal content was measured as a proxy for past agricultural activities. This parameter was determined by loss on ignition of 1 g of soil sample after a hot H_2O_2 pre-treatment destructing OM but preserving charcoals. At the eastern end of the forest dynamics plot, and for a few hundred metres to the east beyond this limit, the forest was subjected to slash and burn cultivation. In order to assess whether the soil charcoal content was a good proxy of past agricultural practices, we measured the charcoal content in 16 additional quadrats beyond the eastern limit of the plot, covering a supplementary area of 6 ha with attested human disturbance. The frequency distribution of the charcoal content values clearly indicated a bimodal distribution pattern, pointing to a trace of human influence beyond the value of 2% of charcoals (Appendix S1). Beyond this threshold, a double amount of charcoal does not necessary indicate a double human activity. Therefore we generated a binary variable (*Man*) taking the value of 1 beyond 2% of charcoal content in the soil (significant past human activity), and taking the value 0 otherwise.

II.4. Data treatment and statistical analyses

Species abundance as well as basal area were determined in each quadrat. For each species, relative frequency (*RF*; absolute frequency divided by the additive frequency (the sum of all species' frequencies), where the absolute frequency is the number of quadrats where the species is present divided by the total number of quadrats), relative density (*RD*; number of individuals of a species divided by the number of all individuals in all species) and relative basal area (*RBA*; total species basal area divided by the total tree basal area of the plot) were calculated. These three indices were used to calculate the importance value index (*IVI*) based on the following equation: $IVI = RF + RD + RBA$ (Cottam and Curtis 1956). Box-Cox transformations of the soil data were conducted before analyses in order to stabilise the variances and bring the variables closer to a normal distribution. We used ordinary kriging to make interpolation on the soil data and to build a map at 6.25 m resolution for each variable (model parameters not presented here). Then, the mean value of each variable was calculated for each quadrat. Relations among all measured soil variables were evaluated with Pearson correlation coefficients, and with the coefficient of intra-class correlation for the relations between the *Man* qualitative variable and the quantitative soil parameters (Appendix S3).

The habitats were defined using two different methods. First, principal component analysis (PCA) was applied to the matrix of soil variables. The selection of the significant axes of the PCA was performed by comparing the distribution of the rank-ordered axes with a broken stick distribution and using the Kaiser-Guttman criterion (Borcard *et al.* 2011). A cluster analysis based on the site scores of the PCA was used to determine the main habitats within the inventory plot (Ward's method), following Borcard *el al.* (2011) for the optimal cutting tree criteria. This procedure defined habitats on the basis of the most variable soil parameters, independently of the tree community. This method is commonly used for summarizing environmental parameters into one or two axes in order to use them as synthetic explanatory variables (e.g., Swaine 1996; Toledo *et al.* 2012; Moraes *et al.* 2016).

A second strategy was used to discriminate habitats by means of multivariate regression trees (MRT) (De'ath 2002). This was carried out in order to model species-environment relationships and to highlight species assemblages (Zhang *et al.* 2016; Wang *et al.* 2016). This approach forms clusters of quadrats by repeating a splitting procedure based on species composition. Each split is characterised by a threshold value of one environmental variable and is made in a way that minimises the dissimilarity within the clusters (within-group sum of squares). Among the numerous possible trees, the retained solution is the one that maximises the predictive power. Therefore, MRT analysis focuses on prediction, making it a useful and powerful tool for ecosystem management and conservation. The result of the analysis is a number of species assemblages to which habitats defined by the threshold values of the environmental variables selected during the splitting procedure correspond. In addition, MRT was shown to outperform the commonly-used redundancy analysis and canonical correspondence analysis (RDA and CCA, respectively) for explaining and predicting species composition. Moreover, MRT needs no model assumptions (linear in RDA, unimodal in CCA) and is indifferent to monotonic transformations of environmental variables, which makes this method very robust (De'ath 2002).

The IndVal index was then used to compute indicator values of individual species within the habitats defined by the PCA and clustering, and by the MRT approaches (Dufrêne and Legendre 1997; Duff *et al.* 2014). The significance of the indices was tested using a torus-randomisation approach (Harms *et al.* 2001; De Cáceres *et al.* 2010; Chuyong *et al.* 2011; Vleminckx *et al.* 2015) in order to correct for spatial autocorrelation. This procedure allowed removing species-habitat associations while maintaining the original abundance spatial patterns. It was repeated 4999 times to build a null distribution of IndVal index values. The observed IndVal index value was then considered significant when it was higher than 95% of the null values. Classical permutations were also run for method comparison. Torus randomisation and permutation tests were performed for each species displaying a minimum of 10 occurrences.

Statistical analyses were conducted using the R statistical software (v. 3.2.2). Data transformations and ordinary kriging were completed using package 'car' (Fox and Weisberg 2011) and 'gstat' (Pebesma 2004), respectively. The PCA and cluster analyses were computed using package 'vegan' (Oksanen *et al.* 2007). The MRT analysis and IndVal index computation were conducted using the package 'mvpart' (De'ath 2006) and 'labdsv' (Roberts 2007), respectively. All R scripts were adapted from Borcard *et al.* (2011).

III. Results

III.1. Floristic composition

A total of 4604 trees with DBH ≥ 10 cm was inventoried in the 10-ha forest dynamics plot. The list of identified species and families is provided in Table 1 with their abundances and *IVI* values; 82 tree species were identified. Mean species density was 43 ± 7 species/ha (range: 32 to 53). All the individual trees occurring on the termite mounds were excluded (8.5% of the total number of individuals). Among them were 14 species exclusively restricted to these mounds (Table 1). The remaining 68 tree species belonged to 32 families and 52 genera. Fabaceae was the most abundant family, comprising 60.4% of the trees. Within this family, Caesalpinioideae was the most abundant subfamily with 42.7% of the individuals followed by Faboideae (14.7%) and Mimosoideae (3.1%). Fabaceae were followed by Apocynaceae (9.3%), Dipterocarpaceae (7.7%), and Combretaceae (3.5%). Eleven species were represented by more than 100 individuals. *Julbernardia paniculata* was the most abundant species with 1240 individuals (26.9%). Other abundant species are *Diplorhynchus condylocarpon* (428), *Brachystegia wangermeeana* (305), *M. macroura* (298), *Julbernardia globiflora* (249), *Pterocarpus angolensis* (247), *Pterocarpus tinctorius* (178), *B. spiciformis* (128), *Pseudolachnostylis maprouneifolia* (119), *Uapaca nitida* (113), and *Albizia antunesiana* (111). The ranking of species according to *IVI* (Table 1) was slightly different. In particular, *B. wangermeeana* ranked third based on frequency and sixth when considering *IVI*s.

Table 1: Table of the identified species and families of the forest plot. The basal area was calculated on the 10 ha of miombo woodland. Species are classified by decreasing IVI. Species marked with an asterisk (*) were exclusively encountered on termite mounds and were not included in the species-habitat association analyses. Ab: Total abundance; RF: relative frequency; RD: relative density; RBA: relative basal area; IVI: importance value index.

Species	Family	Ab	No. quadrats	*RF* (%)	*RD* (%)	*RBA* (%)	*IVI*
Julbernardia paniculata	Fabaceae - Cesalpinioideae	1240	124	6.89	26.93	11.89	45.71
Marquesia macroura	Dipterocarpaceae	298	91	5.06	6.47	27.10	38.63
Diplorhynchus condylocarpon	Apocinaceae	428	150	8.33	9.30	9.14	26.77
Brachystegia spiciformis	Fabaceae - Cesalpinioideae	128	87	4.83	2.78	7.56	15.18
Brachystegia wangermeeana	Fabaceae - Cesalpinioideae	305	65	3.61	6.62	4.35	14.58
Pterocarpus angolensis	Fabaceae - Faboideae	247	97	5.39	5.36	3.74	14.49
Julbernardia globiflora	Fabaceae - Cesalpinioideae	249	78	4.33	5.41	4.01	13.75
Pterocarpus tinctorius	Fabaceae - Faboideae	178	74	4.11	3.87	4.01	11.98
Pseudolachnostylis maprouneifolia	Phyllanthaceae	119	75	4.17	2.58	1.88	8.64
Albizia antunesiana	Fabaceae - Mimosoideae	111	72	4.00	2.41	1.53	7.94
Pericopsis angolensis	Fabaceae - Faboideae	76	53	2.94	1.65	3.25	7.84
Combretum collinum	Combretaceae	85	51	2.83	1.85	1.98	6.66
Uapaca nitida	Phyllanthaceae	113	49	2.72	2.45	1.43	6.60
Combretum molle	Combretaceae	72	44	2.44	1.56	1.59	5.60
Strychnos innocua	Loganiaceae	59	44	2.44	1.28	1.16	4.89
Dalbergia boehmii	Fabaceae - Faboideae	71	45	2.50	1.54	0.47	4.51
Philenoptera katangensis	Fabaceae - Faboideae	61	33	1.83	1.32	1.10	4.26
Monotes katangensis	Dipterocarpaceae	55	35	1.94	1.19	0.99	4.13
Hexalobus monopetalus	Annonaceae	50	42	2.33	1.09	0.49	3.91
Ziziphus mucronata	Rhamnaceae	54	27	1.50	1.17	1.10	3.78
Haplocoelum foliolosum *	Sapindaceae	55	21	1.17	1.19	1.40	3.76
Parinari curatellifolia	Chrysobalanaceae	39	30	1.67	0.85	0.93	3.44
Bobgunnia madagascariensis	Fabaceae - Faboideae	31	27	1.50	0.67	0.80	2.97
Anisophyllea boehmii	Anisophylleaceae	28	27	1.50	0.61	0.47	2.58
Zanthoxylum chalybeum *	Rutaceae	32	24	1.33	0.70	0.42	2.45
Lannea discolor	Anacardiaceae	27	21	1.17	0.59	0.42	2.17
Ficus thonningii	Moraceae	19	14	0.78	0.41	0.47	1.66
Phyllocosmus lemaireanus	Ixonanthaceae	21	19	1.06	0.46	0.14	1.65
Boscia angustifolia *	Capparaceae	18	13	0.72	0.39	0.50	1.61
Albizia adianthifolia	Fabaceae - Mimosoideae	19	17	0.94	0.41	0.25	1.60
Allophylus africanus *	Sapindaceae	21	14	0.78	0.46	0.35	1.58
Uapaca kirkiana	Phyllanthaceae	22	15	0.83	0.48	0.26	1.57
Brachystegia taxifolia	Fabaceae - Cesalpinioideae	16	11	0.61	0.35	0.40	1.36
Ochna schweinfurthiana	Ochnaceae	14	12	0.67	0.30	0.15	1.12
Commiphora glandulosa *	Burseraceae	11	11	0.61	0.24	0.21	1.06
Erythrina abyssinica	Fabaceae - Faboideae	10	9	0.50	0.22	0.33	1.05
Albizia versicolor	Fabaceae - Mimosoideae	10	8	0.44	0.22	0.37	1.03
Erythrophleum africanum	Fabaceae - Cesalpinioideae	10	8	0.44	0.22	0.35	1.01

Hymenodictyon parviflorum	Rubiaceae	14	10	0.56	0.30	0.14	1.00
Uapaca pilosa	Phyllanthaceae	12	11	0.61	0.26	0.08	0.95
Diospyros mespiliformis *	Ebenaceae	12	7	0.39	0.26	0.24	0.89
Ekebergia benguelensis	Meliaceae	10	9	0.50	0.22	0.07	0.79
Ficus ingens	Moraceae	7	6	0.33	0.15	0.23	0.71
Senna abbreviata	Fabaceae - Cesalpinioideae	8	8	0.44	0.17	0.08	0.70
Thespesia garckeana *	Malvaceae	8	5	0.28	0.17	0.24	0.69
Hymenocardia acida	Phyllanthaceae	12	6	0.33	0.26	0.09	0.68
Markhamia obtusifolia	Bignoniaceae	10	6	0.33	0.22	0.09	0.65
Uvariastrum hexaloboides	Annonaceae	7	5	0.28	0.15	0.11	0.54
Vitex fischeri	Lamiaceae	4	4	0.22	0.09	0.22	0.53
Garcinia huilensis	Clusiaceae	6	6	0.33	0.13	0.05	0.51
Diospyros abyssinica *	Ebenaceae	6	3	0.17	0.13	0.19	0.49
Ochna puberula	Ochnaceae	7	5	0.28	0.15	0.05	0.48
Diospyros lycioides *	Ebenaceae	5	5	0.28	0.11	0.08	0.47
Senna singueana	Fabaceae - Cesalpinioideae	6	5	0.28	0.13	0.05	0.45
Commiphora edulis *	Burseraceae	5	4	0.22	0.11	0.12	0.45
Syzygium guineense	Myrtaceae	5	5	0.28	0.11	0.06	0.45
Pappea capensis *	Sapindaceae	4	2	0.11	0.09	0.23	0.43
Strychnos spinosa	Loganiaceae	5	5	0.28	0.11	0.03	0.42
Strychnos cocculoides	Loganiaceae	5	5	0.28	0.11	0.02	0.41
Annona senegalensis	Annonaceae	4	4	0.22	0.09	0.06	0.37
Schrebera trichoclada	Oleaceae	4	4	0.22	0.09	0.06	0.37
Gymnosporia senegalensis	Celastraceae	4	4	0.22	0.09	0.05	0.36
Euclea racemosa	Ebenaceae	4	3	0.17	0.09	0.07	0.32
Combretum celastroides *	Combretaceae	3	2	0.11	0.07	0.03	0.20
Afzelia quanzensis	Fabaceae - Cesalpinioideae	2	2	0.11	0.04	0.02	0.18
Acacia sieberiana	Fabaceae - Mimosoideae	1	1	0.06	0.02	0.10	0.18
Craterosiphon quarrei	Thymelaeaceae	2	2	0.11	0.04	0.02	0.17
Psydrax mutimushii	Rubiaceae	2	2	0.11	0.04	0.02	0.17
Homalium abdessamadii	Salicaceae	2	2	0.11	0.04	0.02	0.17
Vitex mombassae	Lamiaceae	2	2	0.11	0.04	0.01	0.17
Bridelia duvigneaudii	Phyllanthaceae	2	2	0.11	0.04	0.01	0.17
Dichrostachys cinerea *	Fabaceae - Mimosoideae	2	1	0.06	0.04	0.02	0.12
Sterculia quinqueloba *	Malvaceae	1	1	0.06	0.02	0.02	0.10
Gardenia ternifolia	Rubiaceae	1	1	0.06	0.02	0.01	0.09
Dalbergia nitidula	Fabaceae - Faboideae	1	1	0.06	0.02	0.01	0.08
Vitex payos	Lamiaceae	1	1	0.06	0.02	0.01	0.08
Canthium crassum	Rubiaceae	1	1	0.06	0.02	0.01	0.08
Faurea rochetiana	Proteacea	1	1	0.06	0.02	0.00	0.08
Vangueriopsis africana	Rubiaceae	1	1	0.06	0.02	0.00	0.08
Rourea orientalis	Connaraceae	1	1	0.06	0.02	0.00	0.08
Salacia rhodesiaca	Celastraceae	1	1	0.06	0.02	0.00	0.08
Vangueria infausta	Rubiaceae	1	1	0.06	0.02	0.00	0.08
Total		**4604**		**100.0**	**100.0**	**100.0**	**300.0**

III.2. Habitats and species assemblages

The first four principal axes of the PCA accounted for 80% of the total inertia of the soil dataset (Appendix S4). In decreasing order, variables contributing the most to the axes were: Al_{avail}, K_{avail}, Mn_{tot}, charcoal content and Fe_{avail} (Axis 1), Al_{titr}, B_{tot}, pH_{KCl} and Ca_{avail} (Axis 2), P_{tot}, B_{avail}, and K_{tot} (Axis 3). The site scores of the quadrats on the four principal axes were then used in a Ward clustering. All three cutting tree criteria pointed to an optimal division of the dendrogram into five groups (or habitats). Figure 2a describes the habitat membership of quadrats on a map of the forest plot and shows that the resulting habitats are spatially structured in well-defined patches. Table 2 presents the results of the IndVal analysis. *Brachystegia wangermeeana* was mostly present in the fifth habitat (*h*5) and was the only significant indicator species (IndVal = 0.63, $P = 0.031$). No other species could be linked to any habitat based on the IndVal index. Except for *B. wangermeeana*, all significant values obtained with the classical full randomisation were not significant using the more appropriate torus-randomisation procedure (Table 2).

Figure 3 presents the dendrogram resulting from the MRT analysis. While the MRT analysis explained 25.7% (calculated as the reciprocal of the relative error) of the tree community variation, the best predictive model was obtained considering five groups of quadrats (Fig. 3), with a predictive power of 10.3% (the reciprocal of the cross-validated error). The five resulting groups are *H1* to *H5* (Fig. 3). Figure 2b shows the habitat membership of quadrats on a map of the forest plot. The resulting habitats are spatially structured as well-defined patches but show a different distribution in the forest compared to the habitats defined in Figure 2a. The MRT analysis first divided the tree community in two on the basis of Ca_{tot} at a threshold value of 12.4 µg g^{-1}. Below this threshold, the community was split again into two communities based on P_{Olsen} (threshold value: 3.66 µg g^{-1}). In addition, Al_{avail} produced a similar split at this node. Beyond the Ca_{tot} threshold value, the community was split twice, first by P_{EDTA} (threshold value: 2.85 µg g^{-1}), and then by clay content (threshold value: 24.8%) for the sub-community corresponding to the lowest P_{EDTA} concentrations. Moreover, Fe_{avail} and P_{EDTA} gave identical results at this node. Beyond these four discriminant variables, the five habitats also differed significantly for many other soil parameters (Appendix S2). A significant signature of past human activities was detected for only five quadrats within the dynamic plot (corresponding partly to *H4* and *H5*). Ca_{tot} and P_{EDTA} were negatively correlated to past agricultural activities (Appendix S3), but considering the additional sampled area of 6 ha together with the quadrats of the dynamic plot made these correlations non-significant.

Table 2: Table of species IndVal value based on two different habitat definitions. First, five habitats were obtained by a Ward clustering of the first four soil PCA axes. Secondly, computing a multivariate regression tree analysis led to five subcommunities corresponding to soil habitats. IndVal values were computed for each species in the five habitats, for both types of habitats. The IndVal indices were tested by torus randomizations and total randomizations. Codes: ns = non significant, • = P-value < 0.1, * = P-value < 0.05, ** = P-value < 0.01, *** = P-value < 0.001. The first and second symbols give the significance levels based on the torus randomization and total randomization tests, respectively. No symbol means that both tests were non significant. 'max_hab' stands for 'habitat in which the species displayed the highest IndVal value'. Results for species displaying less than 10 individuals in the forest are not presented here.

	PCA & Ward		*MRT*	
Species	**max_hab**	**IndVal**	**max_hab**	**IndVal**
Julbernardia paniculata	1	$0.337^{ns,*}$	1	$0.363^{*,***}$
Diplorhynchus condylocarpon	2	$0.246^{ns,*}$	5	0.218
Brachystegia wangermeeana	5	$0.628^{*,*}$	1	$0.574^{*,***}$
Marquesia macroura	3	$0.240^{ns,*}$	3	$0.312^{**,***}$
Julbernardia globiflora	3	$0.274^{ns,*}$	5	$0.325^{**,***}$
Pterocarpus angolensis	1	$0.210^{ns,\bullet}$	5	$0.219^{ns,*}$
Pterocarpus tinctorius	5	$0.195^{ns,\bullet}$	4	$0.197^{ns,*}$
Brachystegia spiciformis	2	$0.220^{ns,*}$	3	$0.234^{ns,*}$
Pseudolachnostylis maprouneifolia	5	0.123	4	0.113
Uapaca nitida	3	$0.198^{ns,*}$	5	$0.309^{*,***}$
Albizia antunesiana	1	0.138	4	0.122
Combretum collinum	5	$0.189^{ns,*}$	1	$0.172^{ns,*}$
Pericopsis angolensis	2	0.112	5	0.097
Dalbergia boehmii	1	0.105	5	$0.139^{ns,\bullet}$
Strychnos innocua	5	$0.183^{ns,*}$	4	$0.335^{**,***}$
Monotes katangensis	2	0.129	4	$0.170^{\bullet,*}$
Philenoptera katangensis	1	0.069	4	$0.133^{\bullet,*}$
Hexalobus monopetalus	3	0.122	3	0.100
Parinari curatellifolia	3	0.107	5	0.089
Bobgunnia madagascariensis	5	0.070	1	0.097
Anisophyllea boehmii	2	0.106	4	0.070
Uapaca kirkiana	2	$0.093^{ns,*}$	5	0.069
Phyllocosmus lemaireanus	3	0.080	5	0.081
Combretum molle	2	$0.099^{ns,\bullet}$	3	0.069
Albizia adianthifolia	5	$0.126^{ns,*}$	1	0.069
Brachystegia taxifolia	3	0.079	5	$0.130^{*,**}$
Lannea discolor	3	$0.135^{\bullet,*}$	5	$0.146^{*,**}$
Ochna schweinfurthiana	3	0.035	5	0.028
Hymenocardia acida	5	$0.122^{ns,*}$	4	$0.214^{***,***}$
Uapaca pilosa	2	$0.086^{ns,\bullet}$	5	0.063

East-west variation in clay content and stoniness was the most evident gradient in the plot examined. This textural change was associated with chemical modifications along the gradient. *H1* and *H2* have a high SI (>0.6) and low clay content (~19%). This explains the relatively low CEC (~1.6 cmol$_c$/kg) resulting in low available element contents (*Mg*, *K*, *Mn*, *Al*); even tough *Al* saturation was the highest within this group of habitats. High available *Al*, which can cause long-term *P* immobilisation, is likely to be responsible for the low concentrations of the two forms of available phosphorus (P_{EDTA} and P_{Olsen}). Total forms of elements were generally lower than in the other habitats except for *K*. Further, *H4* and *H5* are characterised by high clay content (~35%) and *CEC* values (~1.9 cmol$_c$/kg), and were inversely correlated to the SI (< 0.2). Concentrations of most available cations were favoured by this large *CEC*. Total concentration of elements presented higher values than for the other habitats. Furthermore, *H*3 presented intermediate characteristics compared to the previous groups described, in accordance with its central geographic location. Only total and available forms of *Ca* are inconsistent within the groups of habitat and the overall gradient described. Total concentrations in *Ca* are lower (< 12 µg g^{-1}) in *H*2 and *H*5 located on the opposite sides of the plot. This corresponds to the first bifurcation of the regression tree (Fig. 3).

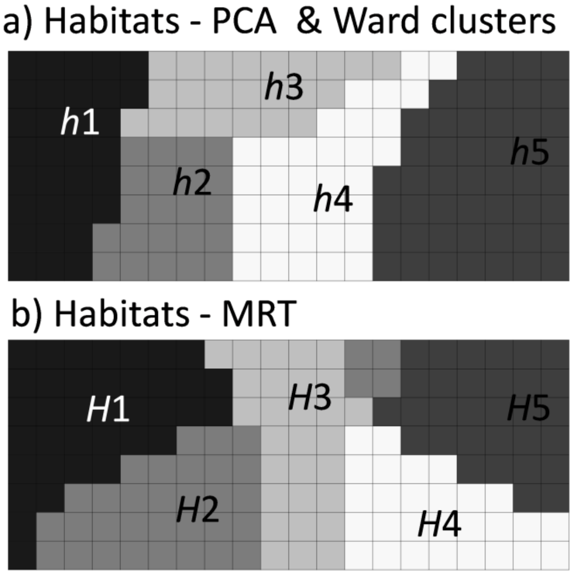


Figure 2: a - Map of the five habitats selected by PCA and Ward clustering (*h*1 to *h*5); b - Map of the five habitats selected on the basis of the Multivariate Regression Trees (*H*1 to *H*5).

The IndVal index was computed and tested for each species on the basis of the habitats in order to detect potential indicator species and search for species assemblages (Table 2). *Julbernardia paniculata* and *B. wangermeeana* were both significantly associated with *H5*, *M. macroura* with *H3*, *Strychnos innocua* and *Hymenocardia acida* with *H4* and *J. globiflora*, *U. nitida*, *Brachystegia taxifolia,* and *Lannea discolor* with *H1*. Within the habitat, species were therefore associated with one another and were significant indicators of their respective soil conditions. Other species were detected as indicator with the classical randomisation test, but this later suffered from inflated type I error, too often considering species to be indicator species, while the torus-randomisation was not significant (Table 2). Besides the relations between significant indicator species and habitats, other interesting features of the tree community are worth mentioning. The five dominant tree species of each habitat are listed in Fig. 3. In addition, *J. paniculata, D. condylocarpon,* and both species of *Pterocarpus* were among the most abundant species of all (or nearly all) the *H*s (see Fig. 3).

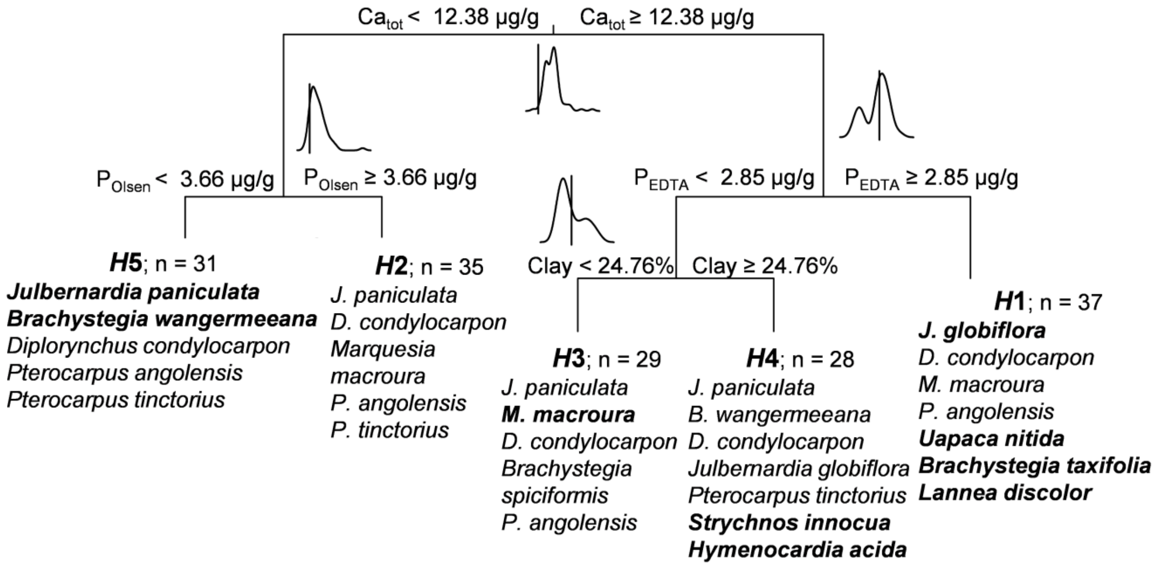


Figure 3: Dendrogram resulting from the Multivariate Regression Trees. Density plots represent the distribution of the corresponding soil parameters through the forest. The vertical bar indicates the splitting threshold value corresponding to the node. *n* indicates the number of quadrats of the cluster. The five dominant species of each habitat are listed in decreasing order of abundance. Species in bold are significant indicator species (4 999 torus randomisations at the α level of 0.05). Statistics of the MRT analysis were: relative error = 0.743; cross-validated error = 0.897; standard error = 0.0419.

IV. Discussion

IV.1. Mikembo forest as a typical wet Miombo

Besides species widely distributed in the Zambezian region (e.g., *D. condylocarpon* and *J. paniculata*), the woody layer of Mikembo forest comprises a number of wet miombo species that are more or less restricted to the northern part of the Zambezian region (*M. macroura, B. wangermeeana, Anisophyllea boehmii, Craterosiphon quarrei,* and *Uvariastrum hexaloboides*). All these species are evergreen or brevideciduous, except *C. quarrei* which is deciduous. Based on Schmitz's (1971) phytosociological system, Mikembo forest is intermediate between Berlinio-Marquesion (rich in evergreen species) and Mesobrachystegion (poor in evergreen species).

Duvigneaud (1958) proposed ecological groups of indicator species for the miombo of Katanga. The most influential soil factors were drainage, soil depth, topography, and soil texture. Soil chemical factors, however, were not measured. Based on his system, Mikembo forest lacks indicators of shallow rocky soil (e.g., *Brachystegia microphylla*) and of impaired drainage on shallow lateritic crust and yellow compact clayey soil (*Brachystegia utilis, B. boehmii, B. stipulata,* and *Isoberlinia* div. sp.). In contrast, the floristic composition of Mikembo forest comprises indicators of deep red soil of plateau miombo (*Brachystegia spiciformis, Julbernardia paniculata*, etc.) and species with a broad ecological amplitude (*Julbernardia globiflora* and *Marquesia macroura*). However, this does not agree very well with our observations since a large part of the plot is established on yellow soils with high gravel load. This discrepancy highlights the need for more detailed species preference characterisations and suggests that any generalisation of indicator status must be done with caution.

IV.2. Modelling species-habitat associations and species assemblages

Very few studies used explanatory multivariate analyses in order to detect soil-plant relationships in the miombo woodland (Mapaure 2001), and this work is the first to use so many potentially relevant soil variables to do so. In other parts of the miombo ecoregion, statistical soil-vegetation analyses have generally considered larger spatial scales and much less ecological parameters (Chidumayo 1987; Kanschik and Becker 2001; Backéus *et al.* 2006; Munishi *et al.* 2011; Mwakalukwa *et al.* 2014).

In our study site, we observed variation in many edaphic factors (e.g., clay content, *Al-Sat*, P_{EDTA}, P_{Olsen}, and Ca_{tot}). At the same time, a striking result is the high species richness of the forest. Eighty-two woody species have been recorded on a 10-ha area. In other floristic studies of miombo, similar species richness is generally found over considerably larger study areas (Chidumayo 1987; Kanschik and Becker 2001; Backéus *et al.* 2006; Munishi *et al.* 2011; Mwakalukwa *et al.* 2014). Our results suggest that this high floristic richness may be accounted for by fine-scale edaphic variation and niche differentiation. Indeed, our model based only on soil variability allowed a significant explanation of 26% of the community distribution. The remaining variation could be partially explained by using spatially explicit models (McIntire and Fajardo 2009) in order to account for other ecological processes (e.g., dispersal limitation, ecological drift). Furthermore, Bauman *et al.* (2016) detected a relation between the dominant ectomycorrhizal fungi of the Mikembo plot and functional traits related to the 'leaf economics spectrum' of host tree species. These results suggest a potential additive role of the soil microbiota in the tree community assembly in miombo woodlands.

To our knowledge, the present study is the first to explicitly test the potential of tree species as indicators of fine-scale habitats in miombo woodlands. To that end, two methods have been used. The first method identified five habitats based on soil factors only, giving more weight to the greatest environmental gradients, independent of their relevance for the tree community. Only one species was identified as an indicator of one of the habitats. In a second approach, vegetation data were used to weight environmental parameters according to their influence on species assemblages. Using a combination of MRT and IndVal analyses and correcting for spatial autocorrelation, five species assemblages were detected. Nine species were significant indicator species (out of a total of 38 species with 10 specimens or more). This is a remarkable result, considering that statistical significance was corrected for spatial autocorrelation using torus randomisation. Four of the five most frequent species, representing > 50% of the total population, were significant indicators of soil conditions (i.e., *J. paniculata, J. globiflora, B. wangermeeana,* and *M. macroura*). In previous studies at much larger scales, the factors that best explained spatial variation of the miombo were soil *pH*, available *Ca*, and texture (Kanschik and Becker 2001; Mapaure 2001; Mwakalukwa *et al.* 2014), topography and soil colour (Backéus et al. 2006), elevation and slope (Munishi *et al.* 2011; Mwakalukwa *et al.* 2014), rainfall (Kanschik and Becker 2001), fire and elephant herbivory (Mapaure 2001).

In this study, soil factors that best discriminated plant assemblages were Ca_{tot}, P_{Olsen}, P_{EDTA}, and clay content. Soil texture and the associated soil chemical composition are determined by the bedrock (dolomitic shales vs siltstones) in the Lubumbashi Plain (Batumike *et al.* 2006) suggesting that geological heterogeneity at the scale of the 10-ha plot may drive the observed variation in edaphic properties. In addition, Ca_{tot} content is the only variable that does not follow the textural gradient. Local concentration of Ca could be driven by carbonate precipitates induced by the termite mound activity as shown by Mujinya *et al.* (2011). Phosphorus is known to be a limiting resource for miombo trees (Högberg 1986; Campbell 1996; Malmer 2007). However, fine-scale variation of floristic composition in response to variation in soil *P* availability does not seem to have been previously reported. This result suggests that niche differentiation for phosphorus availability may be one of the processes structuring miombo at a fine scale.

These four soil variables predicted 10% of the variation in the community composition while the whole set of soil variables explained 26% of the tree composition variability. In previous studies, niche differentiation accounted for 19 to 48% of tropical plant community composition variability (e.g., Mapaure 2001; Jones *et al.* 2008; Legendre *et al.* 2009; Chang *et al.* 2013; Punchi-Manage *et al.* 2013; Vleminckx *et al.* 2015). The explanatory power of species-habitat associations in the Mikembo plot is therefore of the same order of magnitude than in other tropical plant communities. Nonetheless, comparing the relative relevance of deterministic processes in different communities must be done with caution, since the magnitude of the detection of such processes depends upon the quality of the explanatory environmental data collected (Jones *et al.* 2008; Chang *et al.* 2013), the spatial properties of the sampling design (Garzon-Lopez *et al.* 2014), and the analytical methods used (Jones *et al.* 2008).

Mikembo forest has been subjected, at least in part, to slash and burn cultivation. Traces of human disturbance such as cultivation ridges were observed in the eastern portion of the examined plot (portions of *H4* and *H5*, Fig. 2b). Traces of agriculture were localised in the quadrats with higher clay content and lower gravel load, probably because of easier soil tillage conditions. This preference probably explains the negative correlation between P_{EDTA} and past human activities. Indeed, clay content and P_{EDTA} displayed opposite gradients in the plot and were significantly negatively correlated. Since the five quadrats displaying significant human signature were located at the eastern end of the plot, most quadrats non-affected by human corresponded to intermediate or opposite values of the P_{EDTA} gradient. When testing the relationship between P_{EDTA} and *Man* in *H4* and *H5* only, that is, considering an area of relatively homogeneous soil texture, no significant link could be detected, therefore supporting that the correlation of P_{EDTA} and past human activities is indirect and caused by the choice made by past growers for clayey soils with low gravel loads.

The fine-scale (here ~2 ha) variation of habitat distribution implies practical perspectives for management and restoration of miombo forests. 1) The choice of tree species in restoration programmes should be guided by habitat diversity and by preliminary studies assessing potential preferences of species for specific habitat conditions. 2) Conservation strategies should prioritise heterogeneity rather than the extent of the protected areas.

IV.3. Effect of the observation scale

In this study, we sampled an unprecedented high number of physico-chemical soil parameters at a very fine spatial degree of resolution. This may explain why such fine-scale habitat heterogeneity could be detected and related to species assemblages and indicator species. The results support the idea that previous studies may have not detected fine-scale niche differentiation due to 1) a sampling strategy designed to focus on broader scales, 2) sampling of too few environmental variables (Chang *et al.* 2013). The present study did not match the common view of ecology that deterministic processes of tree species assemblage act at broad scales while neutral processes dominate at finer scales. Indeed, Chang *et al.* (2013) showed that this conception may have arisen from the low number and low resolution of environmental parameters measured in many studies. Therefore, we suggest that our analytical approach, both for soil sampling and data analyses, might allow highlighting similar fine-scale tree-habitat associations in other tropical forests, therefore helping to 1) better understand how environmental heterogeneity contributes to species assemblages at fine scales, and 2) establish well calibrated conservation programs.

IV.4. Conclusion

This study revealed fine-scale differentiation of the tree layer of a miombo forest in response to variation of soil factors. The explicit consideration of anthropogenic historical disturbances in the analyses indicated that the soil heterogeneity responsible for habitat and tree community differentiation is mostly natural. Further work is needed to evaluate whether such local habitat heterogeneity can be extended to other miombo forests. Despite the small extent of the plot studied and of the apparent homogeneity of the environment, five contrasting habitats were highlighted and related to distinct indicator tree species. We advise further studies aiming at restoration and conservation purposes in miombo woodlands to adopt the methodology used here and to address fine-scale tree-habitat associations to guide practical decisions.

Acknowledgements

This work was supported by the 'Fondation Rachel Forest', 'Biodiversité au Katanga', and 'Mikembo ASBL' to JIM, and by the Belgian National Fund for Scientific Research and the Wallonie-Bruxelles International Agency to DB. We gratefully acknowledge M. Anastassiou for the access to the forest dynamics plot and accommodations provided at the Mikembo reserve. The authors declare that they are no competing financial interests in relation to the work described.

Supporting information

See *Annexes* at the end of the thesis.

References

Angiosperm Phylogeny Group. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society* 161: 105–121.

Backéus I, Pettersson B, Strömquist L, Ruffo C. 2006. Tree communities and structural dynamics in miombo (*Brachystegia–Julbernardia*) woodland, Tanzania. *Forest Ecology and Management* 230: 171–178.

Baert G, Van Ranst E, Ngongo ML, Kasongo EL, Verdoodt A, Mujinya BB et al. 2009. *Guide des sols en République Démocratique Du Congo, Tome II: Description et données physico-chimiques de profils types*. Lubumbashi, RDC.

Batumike MJ, Kampunzu AB, Cailteux JLH. 2006. Petrology and geochemistry of the Neoproterozoic Nguba and Kundelungu Groups, Katangan Supergroup, southeast Congo: implications for provenance, paleoweathering and geotectonic setting. *Journal of African Earth Science* 44: 97–115.

Bauman D, Raspé O, Meerts P, Degreef J, Ilunga Muledi J, Drouet T. 2016. Multiscale assemblage of an ectomycorrhizal fungal community: the influence of host functional traits and soil properties in a 10-ha miombo forest. *FEMS Microbiology Ecology* 92(10): fiw151.

Borcard D, Gillet F, Legendre P. 2011. *Numerical ecology with R*. Springer, New-York.

Campbell B. 1996. *The Miombo in Transition: Woodlands and Welfare in Africa*. Center for International Forestry Research (CIFOR), Bogor, Indonesia.

Campbell BM, Angelsen A, Cunningham A, Katerere Y, Sitoe A, Wunder S. 2007. Miombo woodlands – opportunities and barriers to sustainable forest management. Center for international forestry research (CIFOR), Bogor, Indonesia.

Chang LW, Zelený D, Li CF, Chiu ST, Hsieh CF. 2013. Better environmental data may reverse conclusions about niche- and dispersal-based processes in community assembly. *Ecology* 94(10): 2145–2151.

Chidumayo EN. 1987. Species Structure in Zambian Miombo Woodland. *Journal of Tropical Ecology* 3: 109–118.

Chuyong GB, Kenfack D, Harms KE, Thomas DW, Condit R, Comita LS. 2011. Habitat specificity and diversity of tree species in an African wet tropical forest. *Plant Ecology* 212: 1363–1374.

Cottam G, Curtis JT. 1956. The use of distance measures in phytosociology sampling. *Ecology* 37: 451–460.

De'ath G. 2002. Multivariate regression trees: a new technique for modeling species-environment relationships. Ecology 83(4): 1105–1117.

De'ath G. 2006. mvpart: Multivariate Partitioning. R Package Version 1.2-6. Available at: http://cran.r-project.org/.

De Cáceres M, Legendre P, Moretti M. 2010. Improving indicator species analysis by combining groups of sites. Oikos 119: 1674–1684.

de Oliveira AA, Vicentini A, Chave J, Castanho CdT, Davies SJ, Martini AMZ et al. 2014. Habitat specialization and phylogenetic structure of tree species in a coastal Brazilian white-sand forest. Journal of Plant Ecology 7(2): 133–144.

Dray S, Pélissier R, Couteron P, Fortin MJ, Legendre P, Peres-Neto PR et al. 2012. Community ecology in the age of multivariate multiscale spatial analysis. Ecological Monographs 82: 257–275.

Dufrêne M, Legendre P. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecological Monographs 67: 345–366.

Duff TJ, Bell TL and York A. 2014. Recognising fuzzy vegetation pattern: the spatial prediction of floristically defined fuzzy communities using species distribution modelling methods. Journal of Vegetation Science 25: 323–337.

Duvigneaud P. 1958. La végétation du Katanga et de ses sols métallifères. Bulletin de la Société Royale de Botanique de Belgique 90: 127–283.

Garzon-Lopez CX, Jansen PA, Bohlman SA, Ordonez A, Olff A. 2014. Effects of sampling scale on patterns of habitat association in tropical trees. Journal of Vegetation Science 25(2): 349-362.

Harms KE, Condit R, Hubbell SP, Foster RB. 2001. Habitat association of trees and shrubs in a 50-ha neotropical forest plot. Journal of Ecology 89: 947–959.

Högberg P. 1986. Soil nutrient availability, root symbioses and tree species composition in tropical Africa: a review. Journal of Tropical Ecology 2(04): 359-372.

Hubbel SP. 2005. Neutral theory in community ecology and the hypothesis of functional equivalence. Functional Ecology 19(1): 166–172.

Jones MM, Tuomisto H, Borcard D, Legendre P, Clark DB, Olivas PC. 2008. Explaining variation in tropical plant community composition: Influence of environmental and spatial data quality. Oecologia 155(3): 593-604.

Kanschik W, Becker B. 2001. Dry miombo – ecology of its major plant species and their potential use as bio-indicators. Plant Ecology 155(2): 139–146.

Legendre P, Mi X, Ren H, Ma K, Yu M, Sun I-F et al. 2009. Partitioning beta diversity in a subtropical broad-leaved forest of China. Ecology 90(3): 663–674.

Legendre P, Legendre L. 2012. Numerical Ecology, 3rd English edn. Elsevier Science BV, Amsterdam.

Loehle C. 2000. Strategy space and the disturbance spectrum: a life-history model for tree species coexistence. The American Naturalist 156(1): 14–33.

MacArthur RH, Levins R. 1967. The limiting similarity, convergence, and divergence of coexisting species. American Naturalist 102: 377–385.

McIntire EJB, Fajardo A. 2009. Beyond description: The active and effective way to infer processes from spatial patterns. *Ecology* 90: 46–56.

Malaisse F. 1996. Endémisme, biodiversité et spéciation dans le centre "domanial" d'endémisme shabo-zambien: remarques préliminaires. In: Guillaumet, J.-L, Belin, M., Puig, H. (eds) *Phytogéographie tropicale: réalités et perspectives*. pp. 193–204. Paris: ORSTOM.

Malmer A, Nyberg G. 2008. Forest and water relations in miombo woodlands: need for understanding of complex stand management. *Working Papers of the Finnish Forest Research Institute* 98: 70–86.

Mapaure I. 2001. Small scale variations in species composition of miombo woodland in Sengwa, Zimbabwe: the influence of edaphic factors, fire and elephant herbivory. *Systematics and Geography of Plants* 71: 935–947.

Meerts P. 2016. An annotated checklist to the trees and shrubs of the Upper Katanga (D.R. Congo). *Phytotaxa* 258 (3): 201–250. http://dx.doi.org/10.11646/phytotaxa.258.3.1.

Millington AC, Chritchley RW, Douglas TD, Ryan P. 1994. Prioritization of indigenous fruit tree species based on former evaluation criteria: some preliminary results from central region, Malawi. In: *Proceedings of the regional conference on indigenous fruit trees of the Miombo ecozone of Southern Africa, Mangochi, Malawi*, ICRAF, Nairobi 97–105.

Moraes DA, Cavalin PO, Moro RS, Oliveira RAC, Carmo MRB, Marques MCM. 2016. Edaphic filters and the functional structure of plant assemblages in grasslands in southern Brazil. *Journal of Vegetation Science* 27: 100–110.

Mujinya BB, Mees F, Boeckx P, Bodé S, Baert G, Erens H et al. 2011. The origin of carbonates in termite mounds of the Lubumbashi area, D.R. Congo. *Geoderma* 165: 95–105.

Munishi PK, Temu RPC, Soka GE. 2011. Plant communities and tree species associations in a Miombo ecosystem in the Lake Rukwa basin, Southern Tanzania: Implications for conservation. *Journal of Ecology and the Natural Environment* 3(2): 63–71.

Mwakalukwa EE, Meilby H, Treue T. 2014. Floristic composition, structure, and species associations of dry Miombo woodland in Tanzania. *Isrn Biodiversity*, vol. 2014, Article ID 153278.

Nguyen HH, Uria-Diez J, Wiegand K (2016) Spatial distribution and association patterns in a tropical evergreen broad-leaved forest of north-central Vietnam. *Journal of Vegetation Science* 27: 318–327.

Oksanen J, Kindt R, Legendre P, O'Hara B, Simpson GL, Solymos P et al. 2008. The vegan package. *Community ecology package*, 631–637.

Pansu M, Gautheyrou J. 2006. *Handbook of Soil Analysis-Mineralogical. Organic and Inorganic Methods.* Springer, Berlin, DE.

Pebesma EJ. 2004. Multivariable geostatistics in S: the gstat package. *Computers and Geosciences* 30(7): 683–691.

Peel MC, Finlayson BL, McMahon TA. 2007. Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences Discussions* 4(2), 439–473.

Picard N, Gourlet-Fleury S. 2008. *Manuel de référence pour l'installation de dispositifs permanents en forêt de production dans le bassin du Congo*. Commission des Forêts d'Afrique Centrale. CIRAD, 00339816, Version-19 Département Environnements et Sociétés. UPR Dynamique des forêts naturelles.

Punchi-Manage R, Getzin S, Wiegand T, Kanagaraj R, Gunatilleke CVS, Gunatilleke IAUN et al. 2013. Effects of topography on structuring local species assemblages in a Sri Lankan mixed dipterocarp forest. *Journal of Ecology* 101: 149–160.

Roberts DW. 2007. Labdsv: Ordination and multivariate analysis for ecology. 1.3–1.

Rondeux J. 1993. *Mesure des arbres et des peuplements forestiers*. Presses universitaires de Gembloux. Belgique.

Schmitz A. 1971. *La végétation de la plaine de Lubumbashi (Haut-Katanga) : Région d'Elisabethville*. Série scientifique n°113. Institut National pour l'Etude Agronomique du Congo (I.N.E.A.C.) [ed.], Bruxelles, BE.

Swaine MD. 1996. Rainfall and Soil Fertility as Factors Limiting Forest Species Distributions in Ghana. *Journal of Ecology* 84(3): 419–428.

Sys C, Schmitz A. 1959. *Notice explicative de la carte des sols et de la végétation du Congo belge et du Ruanda-Urundi : 9. Région d'Elisabethville (Haut-Katanga).* Institut National pour l'Etude Agronomique du Congo belge (I.N.E.A.C.), Bruxelles, Belgique.

Toledo M, Peña-Claros M, Bongers F, Alarcón A, Balcázar J, Chuviña J, Leaño C, Licona JC, Poorter L. 2012. Distribution patterns of tropical woody species in response to climatic and edaphic gradients. *Journal of Ecology* 100(1): 253–263.

Van Gemerden BS, Olff H, Parren MPE, Bongers F. 2003. The pristine rain forest? Remnants of historical human impacts on current tree species composition and diversity. *Journal of Biogeography* 30: 1381–1390.

Van Wyk B, Van Wyk P. 1997. *Field guide to trees of southern Africa.* Struik Nature (Random House Struik); 2nd revised edition. Cape Town, ZA.

Velázquez E, Paine CET, May F, Wiegand T. 2015. Linking trait similarity to interspecific spatial associations in a moist tropical forest. *Journal of Vegetation Science* 26: 1068–1079.

Vleminckx J, Morin-Rivat J, Biwolé AB, Daïnou K, Gillet J-F, Doucet J-L et al. 2014. Soil charcoal to assess the impacts of past Human disturbances on tropical forests. *PloS One* 9(11): e108–121.

Vleminckx J, Drouet T, Amani C, Lisingo J, Lejoly J, Hardy OJ. 2015. Impact of fine-scale edaphic heterogeneity on trees species assembly in a central African rainforest. *Journal of Vegetation Science* 26: 134–144.

Wang Q, Punchi-Manage R, Lu Z, Franklin S, Wang Z, Li Y et al. 2016. Effects of topography on structuring species assemblages in a subtropical forest. *Journal of Plant Ecology*, doi: 10.1093/jpe/rtw047

Werger MJA, Coetzee BJ. 1978. *The Sudano-Zambezian region*. In Werger M.J.A. (Ed.) Biogeography and ecology of southern Africa. pp. 301–462. Junk, The Hague, NL.

Chapitre II

Multiscale assemblage of an ectomycorrhizal fungal community: The influence of host functional traits and soil properties in a 10-ha miombo forest

David Bauman, Olivier Raspé, Pierre Meerts, Jérôme Degreef, Jonathan Ilunga Muledi, and Thomas Drouet

Published July 2016 in *FEMS Microbiology Ecology*, vol. 92, fiw15

Dans le Chapitre I, il a été montré que la composition de la communauté des arbres du Miombo présente un degré de structuration élevé vis-à-vis des propriétés physiques et chimiques du sol. Or, la communauté des arbres du Miombo est largement dominée par des espèces qui dépendent directement d'une relation particulière avec des champignons : la symbiose ectomycorhizienne. Ce sont les déterminants environnementaux de l'assemblage de la communauté de ces symbiotes fongiques des arbres qui seront au centre du Chapitre II. Alors que les aspects spatiaux de la communauté des arbres n'ont été considérés qu'indirectement et grossièrement dans le Chapitre I – au travers des *patchs* d'habitats émergeant des analyses –, le Chapitre II illustrera le cadre d'analyse et d'interprétation écologique des analyses spatialement explicites basées sur l'utilisation de vecteurs propres spatiaux et la partition de variation. Ce cadre de travail permettra d'évaluer l'importance relative des processus de la niche et des processus neutres dans la structuration de la communauté fongique ectomycorhizienne.

Multiscale assemblage of an ectomycorrhizal fungal community: The influence of host functional traits and soil properties in a 10-ha miombo forest

Abstract

Ectomycorrhizal fungi (EMF) are highly diversified and dominant in a number of forest ecosystems. Nevertheless, their scales of spatial distribution and the underlying ecological processes remain poorly understood. Although most EMF are considered to be generalists regarding host identity, a preference toward functional strategies of host trees has never been tested. Here, the EMF community was characterised by DNA sequencing in a 10-ha tropical dry season forest—referred to as miombo—, an understudied ecosystem from a mycorrhizal perspective. We used 36 soil parameters and 21 host functional traits (FTs) as candidate explanatory variables in spatial constrained ordinations for explaining the EMF community assemblage. Results highlighted that the community variability was explained by host FTs related to the 'leaf economics spectrum' (adjusted R^2 = 11%; SLA, leaf area, foliar Mg content), and by soil parameters (adjusted R^2 = 17%), notably total forms of micronutrients or correlated available elements (Al, N, K, P). Both FTs and soil generated patterns in the community at scales ranging from 75 to 375 m. Our results indicate that soil is more important than previously thought for EMF in miombo woodlands, and show that FTs of host species can be better predictors of symbiont distribution than taxonomical identity.

Keywords

beta diversity, community assemblage, ectomycorrhizal fungi, functional traits, Moran's eigenvector maps (MEM), variation partitioning

I. Introduction

Mycorrhization is the most widespread terrestrial symbiosis on earth (Smith and Read 2008). In south central Africa, although arbuscular mycorrhizae are generally the prevalent mycorrhizal type, tropical dry season forests—referred to as miombo—are characterised by the dominance of ectomycorrhizae (Högberg and Piearce 1986; Allen 1991), which have major functional roles in the dynamics of these forests (Bâ *et al.*, 2011; Felten, Martin and Legué 2012).

Recently, DNA sequencing has revealed an unexpectedly high diversity of ectomycorrhizal fungi (EMF) (van der Heijden *et al.*, 2015). Yet, the relationships between ectomycorrhizal (EM) fungal communities and their environments are still poorly understood in tropical regions (Bâ *et al.*, 2012). This is notably the case of the processes driving the β-diversity of these communities. The type of β-diversity considered here corresponds to variation in community composition among all possible pairs of locations, following Anderson *et al.* (2011) system. Species distribution and coexistence within a community are driven by abiotic environment parameters, contagious biotic processes (e.g., competition, predation, and limited dispersal), historical events (e.g., disturbances), and ecological drift, i.e. random variation of species abundances (Legendre and Legendre 2012; Chase 2014). Assessing the relative importance of these ecological processes underlying the community β-diversity is complicated by their specific modes of action and by the difficulty in quantifying biotic interactions or historical events (Borcard and Legendre 1994). Fortunately, analyses of community spatial patterns can be used as synthetic surrogates of the ecological processes responsible for species assemblage and coexistence (McIntire and Fajardo 2009). However, estimating the importance of deterministic processes in such analyses is strongly influenced by methodological issues including the scale of the study, sampling scheme, and spatial autocorrelation in the community (*sensu* Fortin and Dale 2005; Garzon-Lopez *et al.*, 2013). These confounding factors, which are rarely considered, may result in greatly biased interpretations (Chase 2014). New methods of spatial analysis, such as Moran's eigenvector maps (MEM; Borcard and Legendre 2002; Dray, Legendre and Peres-Neto 2006) and variation partitioning (Borcard, Legendre and Drapeau 1992; Peres-Neto *et al.*, 2006) have recently been developed to address those issues (Legendre and Legendre 2012). Those methods allow disentangling the wide range of spatial scales at which the β-diversity of the community is expressed (Borcard and Legendre 2002; Dray *et al.*, 2006), and consider these scale-dependent components of the β -diversity independently in explanatory models.

The ectomycorrhizal fungal communities are recognised as spatially structured (Tedersoo *et al.*, 2010a, 2011, Matsuoka *et al.*, 2016). Most studies have highlighted spatial structuring of EM fungal communities at scales of a few metres (Lilleskov *et al.*, 2004; Pickles *et al.*, 2010, 2012; Branco, Bruns and Singleton 2013; Talbot *et al.*, 2014; Waring *et al.*, 2015), or <1 – 2 ha (Taylor 2002), therefore focusing on the community variation at a very proximal scale (but see Tedersoo *et al.*, 2012). However, a broad-scale spatial organisation of these communities still needs to be studied with a precise multiscale approach (e.g., patterns >100 m) (Peay, Garbelloto and Bruns 2010), and to be analysed with respect to relevant soil explanatory parameters (Tedersoo *et al.*, 2012). In this study, the broad scale is relative to the scale of the local community (we do therefore not consider regional and global scales which encompass other ecological processes acting on the metacommunity). Since each ecological process acts on living communities at a given scale (Legendre and Legendre 2012; Chase 2014), considering different observation windows allows highlighting different relevant processes. Moreover, the factors structuring EM fungal communities are poorly understood (Koide *et al.*, 2007; Dickie *et al.*, 2015). Although soil parameters have proved to influence EMF distribution in some forest ecosystems (Toljander *et al.*, 2006; Branco, Bruns and Singleton 2013; Peay *et al.*, 2010, 2013, 2015), this could not be shown in miombo woodlands (Tedersoo *et al.*, 2011). Regarding symbiotic association specificity, although some host genera and families present specific associations with given EMF taxa (Molina, Massicotte and Trappe 1992; den Bakker *et al.*, 2004), generalist EMF comprise up to 90% of EMF species, and multihost EMF have appeared to be dominant in many ecosystems (e.g., Horton and Bruns 2001; Verbeken and Buyck 2002; Peay *et al.*, 2015). Yet, the prevailing idea about

EMF preferences may be biased since many host plants remain unstudied and different habitats are still understudied (Smith *et al.*, 2009). In miombo forests and savannas, EMF were already shown to display low specificity (Tedersoo *et al.*, 2011; Tsamba, Kativu and Sithole-Niang 2015). Nevertheless, the preference of EMF for a particular life strategy of their hosts has never been assessed. Life strategies are linked to evolutionary trade-offs in the resource allocation of organisms. Such trade-offs are reflected at the individual level in many morphological, physiological, and phenological characteristics, commonly called functional traits (FTs) (Ackerly and Cornwell 2007). For example, tree species allocating many resources in stress tolerance could be associated with particular EMF, which are able to reinforce this strategy. In this case, FTs of host trees would influence the composition of the EM fungal community.

In this paper, we apply newly developed spatial analysis techniques in combination with a high number of environmental variables in order to investigate the processes driving the spatial structure of the EM fungal community in a miombo woodland. Three specific questions are addressed: Does the distribution of EMF in the miombo respond 1) to soil physicochemical conditions or 2) to the functional strategies of host trees, reflected by aboveground FTs, and 3) is the EM fungal community influenced by ecological processes acting at scales broader than the scale of direct proximity between EMF and their hosts?

II. Material and methods

II.1. Study site

The study was conducted in a permanent plot of 10 ha (500 × 200 m) at the Mikembo Natural Reserve, an 800-ha miombo woodland located 30 km NE of Lubumbashi, Upper Katanga, DRC (11°28'57" to 11°29'5"S, 27°40'12" to 27°40'28"E; 1200 m above sea level). In this plot established in 2009, all trees of diameter at breast height >10 cm were georeferenced, identified, tagged, and measured (diameter at breast height) once a year until 2014. The climate is tropical wet and dry (mean annual temperature: 20.5 °C, mean annual rainfall: 1 239 mm) with six months of dry season from May to October. The dominant vegetation is typical miombo woodland, with EM tree species (e.g., *Brachystegia spp.*, *Julbernardia spp.*, and *Marquesia macroura*) dominating the woody community.

II.2. Ectomycorrhized roots

II.2.1. Sampling design

The community sampling was conducted in 34 quadrats of 25 × 25 m distributed fairly regularly over the whole 10-ha forest (see Figure S1 in Supporting Information). Since this study focused on potential broad-scale patterns within the community and the corresponding processes, the range of scales considered went from 70 m to 500 m. In each quadrat, 100 g of fine shallow roots (0-cm to 20-cm deep) of eight randomly-chosen EM trees were collected after tracing them. Back at the camp, the roots were gently washed with water, and checked carefully for ectomycorrhizae with a ×18 magnifying glass. Two 1 cm-EM root tips were selected for each individual tree (16 ectomycorrhized root tips/quadrat), and were kept in 1.5 mL cetyltrimethylammonium bromide DNA extraction buffer until DNA isolation. Because some quadrats only presented six or seven EM trees and because only one EM root tip could be found on some individuals, the total number of samples was 426.

II.2.2. Sequencing strategy

A number of studies using next generation sequencing (NGS) on ITS amplification products of mixed EM roots or environmental samples revealed only few Cantharellales OTUs and a limited number of Boletales OTUs (e.g., Tedersoo *et al.*, 2011; Phorsi *et al.*, 2012) compared to the frequency of those two orders in the above-ground EM community. Possible explanations for this underrepresentation of Cantharellales and Boletales in NGS data include amplification stochasticity, amplification bias or failure of amplification (e.g., Tedersoo *et al.*, 2010b; Schmidt *et al.*, 2013; Tedersoo *et al.*, 2015). For Cantharellales, it is known that amplification of ITS is problematic because of large variation in length (Schoch *et al.*, 2012). For Boletales, ITS normally amplifies well with universal primers, but amplification bias when DNA of Boletales is mixed with DNA from other groups remains a possibility. Because of these problems, and because only a maximum of 2 root tips had to be analysed for a large number of samples that had to be individualised, we preferred to use a Sanger sequencing approach than a multiplexed NGS approach. We also used a two-step, two-gene approach to maximize the chance of identification of the EMF. This approach consisted in first amplifying atp6 (mitochondrial ATPase subunit 6), and then ITS for the root tips for which atp6 failed to amplify. Atp6 was chosen because it works very well in the Boletales (e.g., Kretzer and Bruns 1999; Raspé *et al.*, 2016) and Cantharellales (Amalfi *et al.*, unpublished), and it has been considered as a good candidate for barcoding of fungi (Min and Hickey 2007; Vialle *et al.*, 2009). Moreover, many atp6 sequences have already been obtained from identified specimens collected in tropical Africa, including the study area (Raspé, Degreef and De Kesel 2012), while a very limited number of ITS sequences are available in public databases for identified specimens of Boletales and Cantharellales from Africa. However, atp6 performs poorly (low amplification success rates) in a number of fungal groups, including important EM genera like *Amanita* and, to a lesser exent, *Russula*, *Lactarius*, and *Lactifluus* (O. Raspé, pers. obs.), which is why a second barcode is needed for identification of samples that fail to amplify with atp6 primers.

II.2.3. Molecular analyses

Before DNA extraction, the samples were stored at -80°C. Then, they were ground using a Retsch MM 301 mixer mill, in two steps: first at -80°C, then at 60°C in CTAB extraction buffer at higher intensity. The DNA was then purified with chloroform:isoamyl alcohol (24:1) and precipitated with isopropanol. PCR amplification was first conducted with the primers atp6-1M40F and atp6-2M (Raspé *et al.*, 2016) for all samples. A second amplification with the primers its-OF-T and LB-W (Tedersoo *et al.*, 2006), was performed for fungal samples that could not be amplified by the first primers. It is worth noting that there was no risk of detecting the same OTU on the basis of the ATP6 and the ITS markers, therefore counting them as two different OTUs. Indeed, the OTUs that could be amplified by the atp6 primers were never amplified by ITS, and the OTUs amplified by ITS could not be amplified by atp6 in the first place, therefore assuring that each OTU was detected by one pair of primers only. When ITS sequence quality was low but still allowed detecting the taxonomic order of the fungal sample (see below), a third amplification was carried out with the its-OF-T primer and a primer specific to the order revealed by the BLAST (see Tedersoo *et al.*, 2011). PCR products were checked on 1% agarose gels under UV light and were then purified using 1 U of Exonuclease I and 0.5 U FastAP Alkaline Phosphatase (Thermo Scientific, St Leon Rot, Germany) at 37 °C for 1 hour, followed by inactivation at 80 °C for 15 min. Sanger sequencing was performed by Macrogen Europe (Amsterdam, The Netherlands) on each of the ectomycorrhized root tips separately with the M13F-pUC and M13F primers for atp6 and with the its4 and its5 primers for ITS (Innis *et al.*, 1990). When a specific primer had been used for amplification, sequencing was done with the its5 primer and the specific one. Consensus sequences were constructed using Geneious (version Pro 5.1.7). Then, BLAST were performed against the internal database of the Botanic Garden Meise, which contains ATP6 sequences of over 200 taxa of Boletales and Cantharellales (Raspé, Degreef and De Kesel 2012; Raspé *et al.*, unpublished; Amalfi *et al.* unpublished) and GenBank for the ATP6 sequences, or GenBank (including UNITE) only for the ITS sequences. For ITS, the lower boundary of identity similarity was fixed at 97% to delimit different operational taxonomic units (OTUs) (Smith, Douhan and Rizzo 2007). For atp6, specimens displaying more than three mutations were considered different OTUs (Raspé, Degreef and De Kesel 2012). This stepwise sequencing approach optimised the identification process by utilising the high precision of the atp6 database of the Botanic Garden Meise for African EMF with the ITS data available on GenBank (where most ITS sequences from vouchered specimens correspond to temperate EMF).

II.3. Soil physical and chemical parameters

Five soil samples (20-cm deep) were collected following a grid in the 34 quadrats where community data had been collected and in 68 additional quadrats (Figure S1). These 68 quadrats were used only for exploratory purposes (i.e., for constructing semivariograms, see below), but were not integrated in the explanatory models. Undisturbed soil cores were additionally taken for bulk density (BD) measurements (cylinder method). All sampling and measurements were carried out according to conventional protocols (Pansu and Gautheyrou 2007), briefly presented in Appendix S1. Measured soil parameters were a stoniness index, soil texture (clay, silt, and sand), electrical conductivity, pH-H_2O, pH-KCl, ΔpH, exchangeable Al, plant-available elements $((\mathrm{Al, B, Ca, Fe, K, Mg, Mn, P,\ and\ Zn})_{\mathrm{avail}})$, Olsen phosphorus, total forms of elements $((\mathrm{Al, B, Ca, Fe, K, Mg, Mn, Mo, P,\ and\ Zn})_{\mathrm{tot}})$, effective cation exchange capacity, Al saturation rate, total nitrogen content and soil carbon-to-nitrogen (C/N) ratio, organic matter content, and the extinction coefficient in visible light.

II.4. Functional traits of host trees

Sampling and measurement of the aboveground FTs of EM trees followed Cornelissen *et al.* (2003). Host trees present on the site were *Brachystegia spiciformis*, *B. taxifolia*, *B. wangermeeana*, *Julbernardia globiflora*, *J. paniculata*, *Pericopsis angolensis* (Fabaceae), *Marquesia macroura*, *Monotes katangensis* (Dipterocarpaceae), *Uapaca kirkiana*, *U. nitida*, and *U. pilosa* (Phyllanthaceae). The following traits were measured on ten individuals of each species: maximum height of the species in the entire plot, mean annual increment, wood density, bark thickness, specific leaf area (SLA), leaf area (LA), leaflet area, leaf dry matter content, leaf thickness, leaf decomposability, leaf lignin concentration, foliar element concentrations ((Al, B, Ca, Fe, K, Mg, Mn, N, and P)$_f$) and the foliar C/N ratio. Mean values of FT, weighted by abundances of host species within quadrats, were used as explanatory variables, therefore allowing testing of the effect of the local value of FTs on the local community. Direct effects of individual hosts (taxonomical identity or FT values) on their symbionts within quadrats are not the focus of this study.

II.5. Statistical analyses

All the analyses were conducted in the R environment, version 3.1.0. (R Development Core Team, 2014). The R scripts were adapted from Borcard, Gillet and Legendre (2011) and Dray *et al.* (2012).

II.5.1. Collinearity and model selection

Ecological explanatory variables were Box-Cox transformed in order to stabilise the variances and bring the variables closer to a normal distribution. Pairwise Pearson correlations were computed among all environmental variables and were tested for significance by 999 permutations, using toroidal translations to avoid inflated Type-I error (Clifford, Richardson and Hemon 1989).
We developed a new approach for selecting environmental variables to be used in our explanatory models, optimising the adjusted bimultivariate redundancy statistic (R^2_{adj}) and limiting collinearity. A high level of collinearity in the matrix of independent variables indicates that the same information is provided to the model in more than one way and generates issues in the interpretation of the models (Meloun *et al.*, 2002). While the forward selection with double stopping criterion (Blanchet, Legendre and Borcard 2008) is efficient for model selection and collinearity limitation, it requires the general model to be significant, which is not necessarily the case when dealing with numerous explanatory variables. The new method presented here was designed for models in which few variables might have a significant effect on the response variable(s), but the general model is not significant because of too numerous non-significant explanatory variables, adding noise to the model. The model selection proceeds as follows, for *n* variables: (i) creation of all possible variable combinations (considering each possible group size), (ii) computation within each combination of the variance inflation factors (VIF), a commonly used measure of collinearity (Dormann *et al.*, 2013), (iii) selection of the groups displaying VIF values below a given threshold (here, values <10), (iv) computation of redundancy analyses (RDA) of the community data table on the pre-selected variable groups and selection of the significant models (999 repetitions of a permutation test), and (v) remaining models are reordered according to their adjusted explanatory power (R^2_{adj}). In order to avoid overly conservative adjustments of the adjusted R^2, only final models displaying a number of explanatory variables smaller than half the number of observations were considered (Borcard, Gillet and Legendre 2011). A forward selection with double stopping criterion (Blanchet, Legendre and Borcard 2008) was used for ensuring the choice of parsimonious environmental models.

II.5.2. Environment effect and spatial analyses

The impact of soil and FTs of host trees on the distribution of fungal OTUs was assessed using RDAs. The latter was also used to test and control for a link between the EMF distribution and their host

community (host identity). Community abundance data were Hellinger-transformed prior to analyses (Legendre and Gallagher 2001). General significance of the models and axis-by-axis significance were tested using 9999 permutations. The spatial structures of the EM fungal community were explored using Moran's eigenvector maps (MEM), a robust multivariate spatial method for detecting multiscale spatial patterns in the community (Dray, Legendre and Peres-Neto 2006; Legendre and Legendre 2012). This method constructs orthogonal spatial eigenvectors on the basis of the Cartesian coordinates of the quadrats and models all possible spatial patterns of single species or community data ranging from broad to very fine scales (Dray *et al.*, 2006). In the present case, the spatial properties of the sampling design allowed studying the spatial patterns of the EMF community ranging from 70.5 to 500 m. A connectivity matrix based on a Gabriel graph was used for constructing the spatial eigenvectors. The method proposed by Dray *et al.* (2012) was followed in order to avoid aliasing effects (that is, obtaining a biased view of the spatial scale of a pattern when the sampling interval is too large to resolve the finest fluctuations of the actual pattern, see Platt and Denman 1975). The analysis was based on two tables (matrices): the Hellinger-transformed fungal community table (Y), and the environmental table (E; soil or FTs); Y itself can be decomposed into two tables, F (for fitted community table, the portion of Y explained by E) and R (for residual community table, Y – F).
Scalograms were computed on the first three axes of 1) a PCA of Table Y, 2) on each significant constrained axis of the RDA of Y on the environmental datasets (Table F), and 3) on the first three axes of a partial RDA in which the environmental variables were placed in a covariate matrix (PRA; PCA of Table R). The latter analysis allowed searching for residual spatial patterns in the variability of the community unexplained by the environmental variables. The scalograms are diagrams displaying all the MEM variables in decreasing order of eigenvalue (and therefore of Moran's index) on the *x*-axis, and a statistic indicating the response of the response variable - an ordination axis (PCA, RDA or PRA) - to these spatial variables (*y*-axis; Legendre and Borcard 2006). The scalograms were presented in a smoothed version (Munoz 2009) in which MEM variables were gathered in six groups of five variables and one group of three. The first half of the generated MEM variables represent positively correlated spatial patterns of decreasing scales, and the second half represent negatively correlated patterns (local structures), so that smoothed MEM 1 to 3, and 4 to 6 correspond to positively and negatively correlated spatial structures, respectively. The statistic of the scalograms was the R2Max (Ollier, Couteron and Chessel 2006), and was tested for the smoothed MEM displaying the highest determination coefficient (R^2) by permuting the response variable (999 repetitions; Dray *et al.*, 2012). This procedure allowed determining whether the observed value of the statistic was likely to be obtained in the absence of spatial structure (null hypothesis). The R2Max is expected to peak when a smoothed MEM accounts for a large portion of the ordination axis variance, therefore indicating spatial structuring at the corresponding scale.

In order to help make an accurate hypothesis regarding the processes responsible for unexplained community variability that could not be spatially modelled (and therefore occurred out of the range of our detectable scales), we estimated the spatial scale of variation of all measured variables by fitting semivariogram models to the empirical semivariograms, therefore obtaining their spatial parameter estimates (nugget, partial sill, and range). Any variable displaying spatial variation out of the range of scales captured by the present sample design may have a structuring influence on the community at the corresponding scale. Semivariograms were constructed on the basis of the maximum amount of available data for both soil and FTs (soil: 102 quadrats; FT: 160 quadrats).

Variation partitioning was also performed on the EM fungal community following Peres-Neto *et al.* (2006) and Borcard, Gillet and Legendre (2011) in order to attain an overview of the relative importance of two components on the EM fungal community composition: (i) environmental effects and (ii) spatial variables, possibly separated into different subscales. Spatial variables were selected after testing, comparing numerous connectivity, and weighting matrices based on the corrected Akaike information criterion (see Appendix S2 in Supporting Information).

III. Results

III.1. Fungal diversity

Identifiable sequences of EMF were obtained for 388 root tips (one sequence/tip). New sequences were deposited in GenBank (see Table S1 in Supporting Information). A total of 120 different OTUs were detected (68% from ITS, 32% from ATP6), belonging to 13 genera in eight orders (Table 1); 81 OTUs did not figure in GenBank. The most represented genera were *Lactarius* s.l. (including *Lactifluus*), *Russula*, *Sebacina*, *Clavulina*, and *Tomentella*. Nevertheless, several identifications could not be obtained on a finer taxonomic level than the order (*e.g.*, Thelephorales). The most represented orders were, by decreasing abundance and diversity, the Russulales, Thelephorales, Agaricales, and Sebacinales.

Seventy-seven OTUs occurred only once or twice in the data, and 91 OTUs were encountered in three or less quadrats. The symbiotic fungal community thus displayed a high diversity with an elevated number of rare OTUs and low evenness. In order to minimise the undersampling stochasticity (related to rare OTUs), the subsequent analyses were performed on OTUs presenting at least 10 occurrences (bold OTUs in Table 1). This allowed discussing the unexplained variability of the EMF community in terms of ecological stochasticity without having to consider undersampling stochasticity. Analyses conducted on the whole community showed results similar to those presented below but with lower signals, and are not presented here for brevity.

Table 1: List of the ectomycorrhizal fungal OTUs of the belowground community classified by decreasing number of occurrences of orders and of OTUs within orders. Unpubl. accessions correspond to OTUs obtained with ATP6 primers in a previous, yet Unpubl., study. Bold OTUs were used for environmental and spatial analyses.

		Accession numbers			
Order	OTUs	ITS	atp6	No. quadrats	Tot. abund.
Russulales	***Lactifluus velutissimus***		Unpubl.	10	15
Russulales	***Russula sp14_18B_53C***	KT461290		5	12
Russulales	***Russula sp01_33B_59B***		KT461416	7	10

Russulales	*Lactarius sp04_08A_02B*		KT461426	5	7
Russulales	*Lactarius sp06_08A_06A**	KT461337		4	7
Russulales	*Lactifluus urens*		Unpubl.	6	7
Russulales	*Lactarius sp02_09B_34A*		KT461411	4	6
Russulales	*Lactarius sp03_37A_33B*		KT461413	1	5
Russulales	*Lactifluus pelliculatus*		Unpubl.	2	5
Russulales	*Russula sp16_15C_06A*	KT461296		2	5
Russulales	*Lactifluus sp01_22A_38B**	KT461363		3	4
Russulales	*Russula sp08_20D_109B*		KT461433	3	4
Russulales	*Lactarius sp01_10D_42A*		KT461410	2	3
Russulales	*Lactarius sp09_12C_02B*	KT461403		2	3
Russulales	*Lactifluus gymnocarpoides*		Unpubl.	3	3
Russulales	*Lactifluus madagascariensis**	KT461376		1	3
Russulales	*Lactifluus pelliculatus(cf.)*		Unpubl.	2	3
Russulales	*Russula sp04_14C_03B*	KT461306	KT461425	3	3
Russulales	*RUSSULALE sp01_28D_61B*	KT461336		2	3
Russulales	*Lactarius kabansus*		Unpubl.	1	2
Russulales	*Lactarius sp05_30C_B1B*	KT461310	KT461438	2	2
Russulales	*Lactarius sp07_06A_08B*	KT461396		1	2
Russulales	*Lactifluus cyanovirescens*		Unpubl.	1	2
Russulales	*Lactifluus ruvubuensis*		Unpubl.	1	2
Russulales	*Russula patouillardii_11B_54B*	KT461397	KT461434	1	2
Russulales	*Russula sp02_01D_88A*		KT461419	2	2
Russulales	*Russula sp03_04D_J5A*	KT461315	KT461420	1	2
Russulales	*Russula sp05_23C_38B*	KT461316	KT461427	1	2
Russulales	*Russula sp06_23C_08A*	KT461305	KT461428	1	2
Russulales	*Russula sp11_19B_46A*		KT461436	1	2
Russulales	*Russula sp15_09B_41B*	KT461326		1	2
Russulales	*Russula sp20_24C_64A*	KT461400		2	2
Russulales	*Russula sp22_38B_42A_Clon04.01*		KT461447	2	2
Russulales	*RUSSULALE sp03_02C_14A_Clon06.01*	KT461325	KT461449	1	2
Russulales	*Lactarius sp08_12C_02A*	KT461318		1	1
Russulales	*Lactarius tenellus*		Unpubl.	1	1
Russulales	*Lactifluus laevigatus*		Unpubl.	1	1
Russulales	*Lactifluus pelliculatus_14C_11A*		Unpubl.	1	1
Russulales	*Lactifluus rubroviolascens_04D_J2B*	KT461394		1	1
Russulales	*Russula albofloccosa*		Unpubl.	1	1
Russulales	*Russula brunneorigida_17B_69B*	KT461399		1	1
Russulales	*Russula cellulata_1*		Unpubl.	1	1
Russulales	*Russula cellulata_4*		Unpubl.	1	1
Russulales	*Russula sp07_12C_01A*	KT461304	KT461431	1	1
Russulales	*Russula sp10_20D_26B*		KT461435	1	1
Russulales	*Russula sp12_19B_62B*	KT461319	KT461437	1	1
Russulales	*Russula sp13_30C_154A*	KT461314	KT461439	1	1
Russulales	*Russula sp17_18B_75B*	KT461309		1	1
Russulales	*Russula sp18_13A_57B*	KT461317		1	1
Russulales	*Russula sp21_39B_52A_Clon01.04*		KT461443	1	1
Russulales	*Russula sp23_11B_49A*	KT461323		1	1
Russulales	*Russula sp24_09B_44B*	KT461404		1	1
Russulales	*Russula sp25_23C_40B*	KT461329		1	1
Russulales	*Russula sp26_19B_47B*	KT461332		1	1
Russulales	*Russula sp27_24C_64B*	KT461335		1	1
Russulales	*Russula sp28_39B_55A*	KT461407		1	1
Russulales	*RUSSULALE sp02_17B_79B*	KT461393		1	1
Thelephorales	***Thelephoraceae sp01_36B_75A***	KT461295	KT461414	9	27
Thelephorales	***Tomentella sp01_08A_02B****	KT461388	KT461417	11	14
Thelephorales	*Tomentella sp03_06A_15B**	KT461389	KT461418	3	6
Thelephorales	*Thelephoraceae sp02_12C_J1A*	KT461312	KT461430	2	3
Thelephorales	*THELEPHORALE sp02_01D_J1A*		KT461421	3	3
Thelephorales	*Tomentella sp04_19B_42B*	KT461320	KT461441	1	2
Thelephorales	*THELEPHORALE sp01_33B_J1B*		KT461415	1	1

Thelephorales	*THELEPHORALE sp03_06A_08B*		KT461423	1	1
Thelephorales	*THELEPHORALE sp04_12C_07A*		KT461440	1	1
Thelephorales	*Tomentella sp02_28D_J1A*	KT461311		1	1
Thelephorales	*Tomentella sp05_01D_J2A**	KT461390		1	1
Thelephorales	*Tomentella sp06_19B_62A*	KT461333		1	1
Agaricales	***Inocybe sp01_05C_J3A***	KT461289		13	24
Agaricales	*AGARICALE sp01_06A_06B*		KT461422	5	5
Agaricales	*Cortinarius sp01_10D_71B*	KT461327		2	3
Agaricales	*Amanita sp01_29C_J4B*	KT461292		1	2
Agaricales	*Amanita loosii_37A_26B*	KT461405		1	1
Agaricales	*Amanita sp02_18B_74B*	KT461300		1	1
Agaricales	*Clavaria sp01_05C_J5A_Clon03.03*		KT461445	1	1
Agaricales	*Cortinarius sp02_08A_02A*	KT461331		1	1
Agaricales	*Inocybe sp02_27D_35A*	KT461298		1	1
Agaricales	*Inocybe sp03_25D_89B*	KT461401		1	1
Sebacinales	***Sebacina sp01_04D_67B***	KT461291		7	13
Sebacinales	*Sebacina sp05_14C_34A*	KT461297		4	6
Sebacinales	*Sebacina sp13_38C_26A*	KT461322		4	5
Sebacinales	*Sebacina sp03_35D_31B**	KT461346		4	4
Sebacinales	*Sebacina sp07_02C_102A*	KT461302		2	4
Sebacinales	*Sebacina sp02_35D_J1A*	KT461293		2	3
Sebacinales	*Sebacina sp10_39B_75A*	KT461406		3	3
Sebacinales	*Sebacina sp06_37A_73A*	KT461301		2	2
Sebacinales	*Sebacina sp12_30C_165A*	KT461321		1	2
Sebacinales	*Sebacina sp04_06A_108A*	KT461392		1	1
Sebacinales	*Sebacina sp08_12C_04A*	KT461303		1	1
Sebacinales	*Sebacina sp09_02C_99B*	KT461307		1	1
Sebacinales	*Sebacina sp11_33B_59A*	KT461398		1	1
Sebacinales	*Sebacina sp14_08A_47B_LQ*	KT461324		1	1
Sebacinales	*Sebacina sp15_37A_22B*	KT461402		1	1
Sebacinales	*Sebacina sp16_19B_49B*	KT461328		1	1
Sebacinales	*Sebacina sp17_23C_53A*	KT461334		1	1
Sebacinales	*SEBACINALE 38C_26C_DB_col_09_11*		KT461454	1	1
Cantharellales	***Clavulina sp03_35D_32B****	KT461354		12	20
Cantharellales	*Clavulina wisoli**	KT461359		6	9
Cantharellales	*Clavulina sp02_02B_79A*	KT461308	KT461432	4	6
Cantharellales	*Clavulina sp04_10D_73A**	KT461352		2	3
Cantharellales	*Clavulina sp01_10D_73B*		KT461412	2	2
Cantharellales	*Clavulina sp05_35D_32A*	KT461313		1	1
Boletales	*BOLETALE sp01_14C_03A*		KT461424	3	3
Boletales	*Boletellus sp._ADK3920*		Unpubl.	1	1
Boletales	*Xerocomus sp12_JR5798*		Unpubl.	1	1
Trechisporales	*TRECHISPORALE sp01_23C_43A*	KT461299		1	1
Trechisporales	*TRECHISPORALE sp02_28D_106B*	KT461330		1	1
Hysterangiales	***Aroramyces sp02_09B_50A***	KT461294	KT461409	5	10
Hysterangiales	*Aroramyces sp01_09B_43A*		KT461408	2	3
Hysterangiales	*Aroramyces sp03_35D_J1A*		KT461429	3	3
Hysterangiales	*HYSTERANGIALE sp02_04D_61B*	KT461395		1	2
Unknown	*05C_J5A_Clon03.01*		KT461453	2	2
Unknown	*02C_14A_Clon06.05*		KT461450	1	1
Unknown	*05C_J3A_Clon02.02*		KT461444	1	1
Unknown	*05C_J3A_DB_col_02_13*		KT461451	1	1
Unknown	*05C_J5A_DB_col_02_09*		KT461446	1	1
Unknown	*08A_23A_Clon05.03*		KT461448	1	1
Unknown	*20D_26A*		KT461442	1	1
Unknown	*38C_10B_Clon10.02*		KT461452	1	1

* OTUs displaying >97% similarity with at least one sequence already present in GenBank

III.2. Ectomycorrhizal β-diversity, soil and host functional traits

General descriptive statistics as well as ranges of spatial correlation of all soil variables are presented in Table S2 (Supporting Information). The model selection procedure retained eight soil variables: soil C/N ratio, Ca_{avail}, and the total forms of Al, B, Mn, Mo, P, and Zn.

The selected variables presented significant correlations with other explanatory variables (see Table S3 in Supporting Information). These variables explained a significant proportion of the variation of Y (R^2_{adj} = 0.17, P = 0.003). The estimated Table F had three significant axes modelling 35.8%, 25.0%, and 18.3% of total variability in F, respectively. Triplots of significant constrained axes are presented in Figure S2 (Supporting Information). Soil variables contributing the most to these constrained axes were, respectively, (i) the concentrations of Mn_{tot} and Zn_{tot}, the soil C/N ratio, and the concentration of Ca_{avail} (r = 79.4%, 64.3%, -57.7%, and -47.0%, respectively), (ii) the concentration of Mo_{tot} and Ca_{avail} (r = -47.4% and 43.0%, respectively), and (iii) the concentration of B_{tot} and P_{tot} (r = -65.0% and -69.0%, respectively). Figure 1 presents the first three axes of the PCA (Table Y, Figures 1 a, b, and c), RDA (Table F, Figures 1 d, e, and f) and PRA (Table R, Figures 1 g, h, and i) as well as the scalograms constructed on seven spatial components and the *R2Max* statistics based on 999 permutations. The first three constrained axes of the RDA were mapped and highlighted three structures of the community corresponding to patterns in the environmental table. The first two structures (Figures 1 d and e) were almost exactly the same as those represented by the first axes of the PCA of Y (Figures 1 a and b), indicating that the main variation patterns of the community were successfully modelled by the soil variables. The maps of the three constrained axes corresponded to significant spatial patterns. The first and third axis patterns ranged from 200 m to 375 m (*R2Max* = 41.7%, P = 0.013, Figure 1 d) and from 200 m to 300 m (*R2Max* = 33.0%, P = 0.036, Figure 1 f), respectively. At a finer scale, the second axis corresponded to a pattern ranging from 75 m to 100 m (*R2Max* = 43.2%, P = 0.007, Figure 1 e). Figures 1 g, h, and i correspond to the first non-constrained axes of the RDA (Table R), after all the variation explained by the soil variables was removed. The three maps of the residual site scores correlated significantly with fine-scale (75 m to 100 m) spatial components (*R2Max* = 32.2%, P = 0.035, *R2Max* = 36.1%, P = 0.027, and *R2Max* = 36.9%, P = 0.028, respectively). Analysing maximum ranges of spatial variation of soil variables (Table S2) showed that the stoniness index was the only variable displaying spatial correlation exclusively below the lower spatial resolution of the study (70 m).

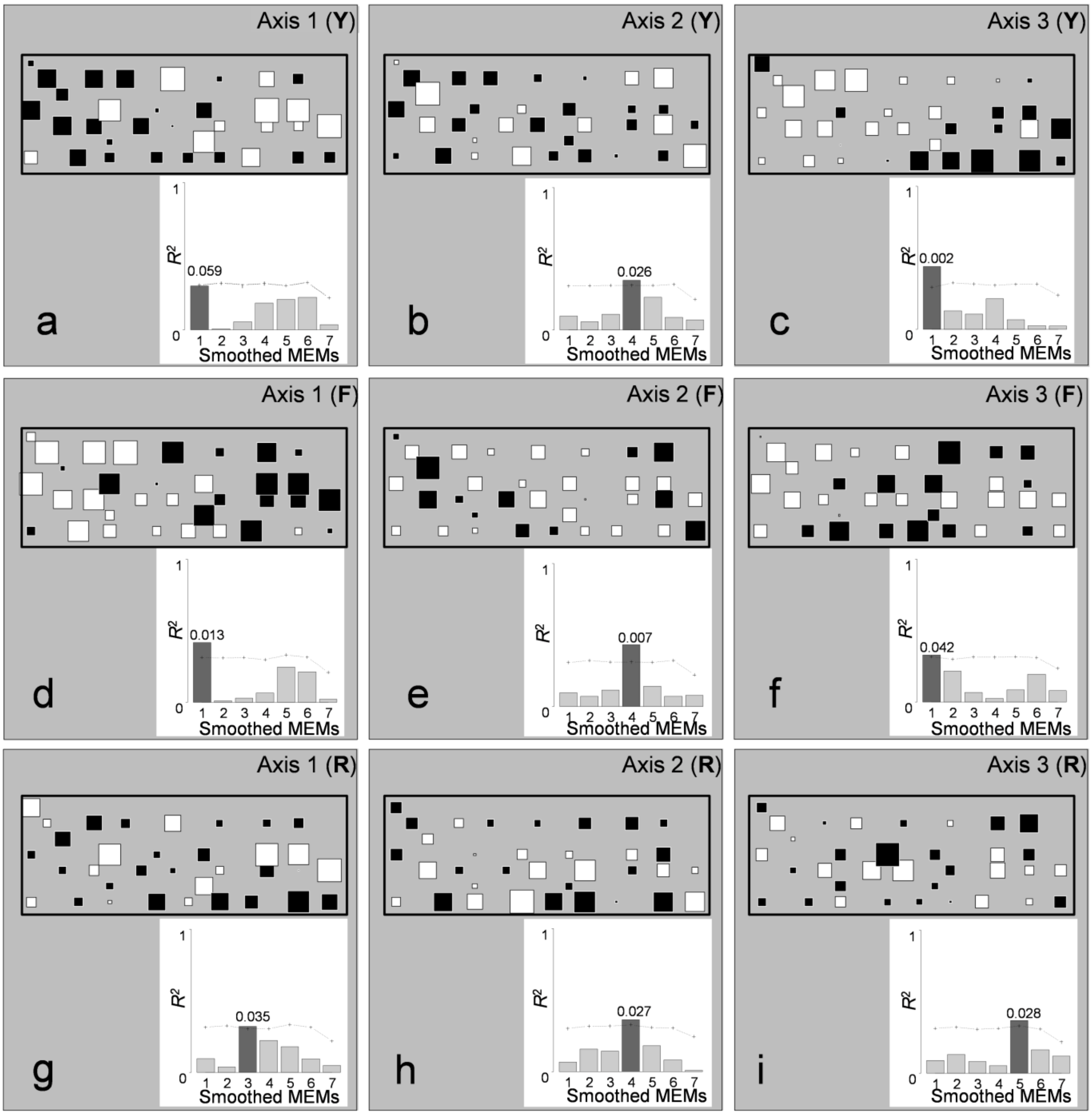


Figure 1: Mapped results of the explanatory analysis of the ectomycorrhizal fungal species distribution based on soil parameters. The nine maps represent the forest plot and are delimitated by the black horizontal rectangles (see Figure S1 in Supporting Information for a more detailed map). The maps illustrate the site scores along the forest plot on the first three axes of the PCA of the raw community data Table Y (a, b and c), the fitted Table F (RDA with the selected soil variables, d, e, and f) and the residual Table R (partial PCA with soil variables as covariables, g, h, and i). The squares inside the maps represent quadrats in which the fungal community was sampled. The size and colour of the quadrats is linked to their position on the corresponding ordination axis: the size of the square is proportional to the absolute value of the site score on the axis, and the black colour corresponds to positive site scores, while the white corresponds to negative site scores. Thus, big quadrats of the same colour present similar community composition and values of environmental variables. For each significant RDA axis, a smoothed scalogram indicates the fraction of total variance (*R^2*) explained by each of the seven spatial components (vertical light and dark grey bars of the bar plots below the maps). The spatial scale corresponding to the highest *R^2* (in dark grey) is tested by 999 permutations of the observed statistic. The *P* values are presented above the corresponding bars (the null hypothesis being that the corresponding map is not spatially structured). The 95% confidence limit is represented by the line of plus signs.

The same analysis was conducted using FTs of the host tree community as explanatory variables. Descriptive statistics as well as ranges of spatial correlation of all FTs are presented in Table S3. The four traits retained by the selection procedure were: the SLA, LA, Mg_f, and leaf thickness. Significant correlations between these variables and other traits are presented in Table S4 (Supporting Information). Figure S3 (Supporting Information) indicates that closely related host species do not have more similar values of the selected FTs. On the contrary, the most similar values of these traits correspond to that of species from different genera and/or families. These selected FTs explained a significant proportion of the variation of the initial community composition Table Y (R^2_{adj} = 0.11, P = 0.003). The fitted Table F had two significant axes that modelled 54.2% and 24.5% of the total variability in F, respectively. The FTs contributing the most to these axes were, respectively, (i) the LA and Mg_f (r = 60.4% and 56.2%) and (ii) the SLA (r = -51.3%). Triplots of significant constrained axes are presented in Figure S2. Figure 2 presents the two first axes of the RDA (Table F, Figures 2 a and b) and of the PRA (Table R, Figures 2 c and d) as well as the scalograms and *R2Max* statistics. Figures 2 a and b corresponded to patterns of 200 m to 375 m and 75 m to 100 m (*R2Max* = 43.0%, P = 0.007, and *R2Max* = 51.3%, P = 0.002), respectively. No significant residual spatial pattern was detected in Table R.

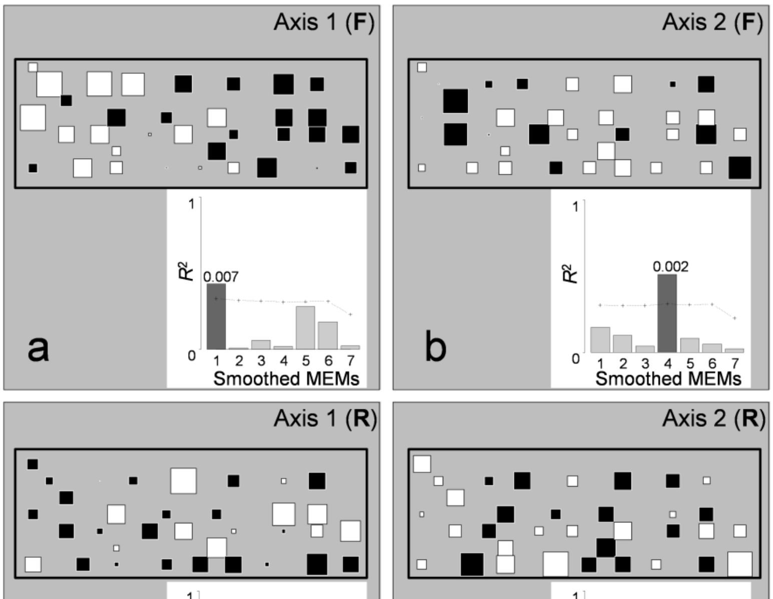


Figure 2: Mapped results of the explanatory analysis of the ectomycorrhizal fungal species distribution based on host FTs. The nine maps represent the forest plot and are delimitated by the black horizontal rectangles (see Figure S1 in Supporting Information for a more detailed map). The maps illustrate the site scores along the forest plot on the first two constrained axes of the fitted Table F (RDA with the selected FT variables, a and b) and the residual table R (partial PCA with host FT variables as covariables, c and d). See the caption of Figure 1 for the meaning of the colour and size of the squares inside the maps. For each axis, a smoothed scalogram indicates the fraction of total variance (R^2) explained by each of the seven spatial components (vertical light and dark grey bars of the barplots below the maps). The spatial scale corresponding to the highest R^2 (in dark grey) is tested by 999 permutations of the observed statistic. The *P* values are presented above the corresponding bars (the null hypothesis being that the corresponding map is not spatially structured). The 95% confidence limit is represented by the line of plus signs.

The analysis of the maximum ranges of spatial variation of the local weighted mean values of FT (Table S3) showed that the leaf decomposability, Al_f, Zn_f, leaf N, and lignin content displayed spatial correlation exclusively below 70 m.

Host species abundances were then used as explanatory variables in the last RDA that showed a link between these variables and the EM fungal distribution (R^2_{adj} = 0.10, P = 0.03). In order to test for the presence of a 'pure functional' signal, a partial RDA was computed using the selected FTs as an explanatory matrix and abundances of host species as a covariance matrix. Results revealed a significant signal of FTs on the EM fungal community (R^2_{adj} = 0.07, P = 0.047). Reversion of the above-mentioned explanatory and covariance matrices led to a non-significant model (P = 0.14).

Results of the variation partitioning of Y are illustrated in Figure 3 a (soil) and Figure 3 b (FT). In the following, *broad* and *fine* scales were used for patterns >200 and <100 m, respectively. Considering Figure 3 a, the analysis detected a slightly significant pure soil signal (R^2_{adj} = 0.07, P = 0.044) and a highly significant fraction of the pure fine-scale spatial component (R^2_{adj} = 0.09, P = 0.005). A fraction corresponding to broad-scale structured environment (R^2_{adj} = 0.08) and another corresponding to the fine-scale structured environment (R^2_{adj} = 0.02) were also detected. The three components together accounted for 24% of the total variability in Y.

Results of Figure 3 b were very similar to those of Figure 3 a except for the pure environmental fraction, which was not significant. All components taken together explained 21% of the total variation within Table Y.

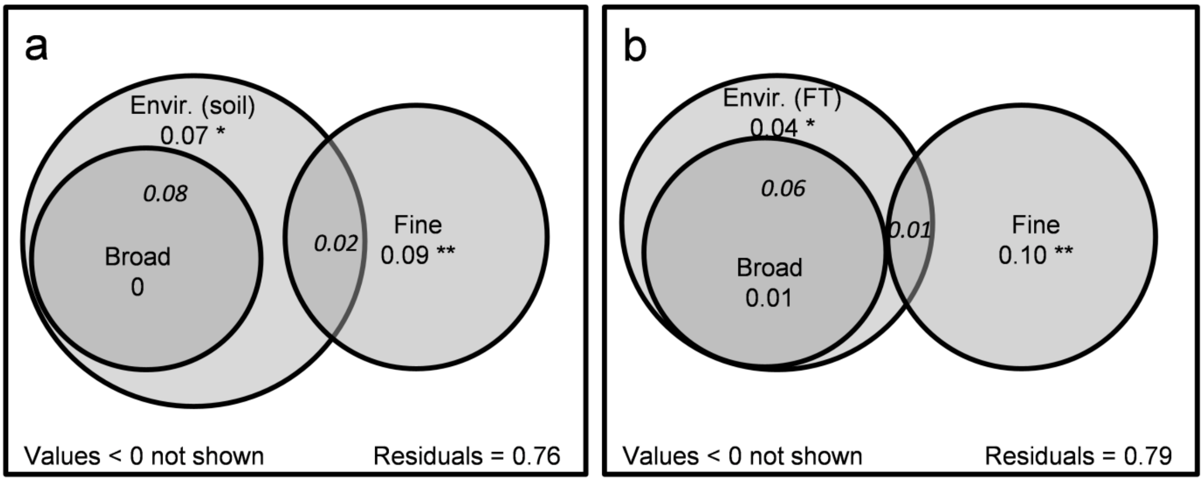


Figure 3: Variation partitioning of the EM fungal community among 1) an environmental component (soil and host FTs in Figures 3 a and 3 b, respectively), 2) the broad-scale spatial variables (> 200 m) and 3) the fine-scale spatial variables (< 100 m). Values in the circles correspond to the R^2_{adj} of the associated partial RDA. The single and double asterisks indicate significance levels (P < 0.05 and P < 0.01, respectively). Italic values correspond to R^2_{adj} computed by subtraction of other R^2, which are therefore non-testable.

IV. Discussion

IV.1. EMF, soil and functional traits of host trees

Using ca. 400 EM fungal sequences was sufficient to show highly significant effects of soil and FTs on the community by means of powerful and robust statistical methods controlling for spatial dependence.

In a previous work (Tedersoo *et al.*, 2011), the influence of soil on miombo EMF could not be detected. This probably has to do with the high number of environmental variables measured and with the scales considered here, and at which soil variables act on the community. Using relevant environmental variables is crucial in order to get a precise idea of the magnitude of deterministic and neutral processes in community assemblages and species coexistence (Chang *et al.*, 2013). In the present study, the assessment of a large panel of environmental variables on the β-diversity of a tropical fungal EM community brought a new insight into the role of soil and host aboveground FTs.

While EMF are often recognised as generalists with respect to host tree species (e.g., Verbeken and Buyck 2002; Toljander *et al.*, 2006; Bâ *et al.*, 2012; Tsamba, Kativu and Sithole-Niang 2015, but see Bruns, Bidartondo and Taylor 2002; Smith *et al.*, 2009), our results showed a significant link between host abundances and the distribution of EMF. Nevertheless, this relation was not detected anymore when controlling for the FT effect. Yet, considering FTs as explanatory variables of the EMF community while controlling for host tree abundances highlighted a significant pure signal of host FTs. This indicates that the effect detected through FTs reflects a real functional effect and not an indirect influence of host identity. This functional influence on EMF distribution does not come either from host phylogenetic relatedness since no clear trait conservatism was detected (Figure S3). The most influential traits were LA, SLA, and foliar Mg concentration—that is, FTs that are tightly associated to the strategy of resource capture and use (the 'leaf economics spectrum'; Reich 2014). In addition, LA and SLA were strongly correlated to several foliar nutrient concentrations (e.g., K, Mn, Zn, C/N ratio), but also to bark thickness (Table S4), which are FTs also directly (foliar nutrient concentrations) or indirectly (bark thickness) related to the leaf economics spectrum. No belowground FTs were measured in the present study, but Kramer-Walter *et al.* (2016) highlighted that aboveground foliar traits were correlated to some belowground FTs, such as the root tissue density. These traits were related to the growth rate and responded in the same way to a soil fertility gradient.

In our study, the influence of host FTs on EMF could be that host trees displaying high values of these traits have higher growth rates related to greater needs for nutrients and would therefore be associated with a portion of the symbiont community compatible with this strategy. From a more fungal point of view, the relevance of these FTs may also be related to the quality/quantity of the sugar produced and transferred to the fungal symbionts. Nevertheless, Peay *et al.* (2015) and Miyamoto *et al.* (2015) showed that host trees and EMF could be influenced by common environmental parameters. Disentangling direct from indirect correlations between patterns of different datasets is therefore a difficult task which can only be achieved by controlled experiments. In the present case, since FTs were to some extent correlated to the soil, part of the FTs influence on the EMF distribution may be indirect. Future studies would need to investigate the relevance of host resource acquisition strategies in a more direct way by testing the effect of root traits directly related to interactions with EMF, such as the transport, production and type of sugars on the EMF community assembly.

The most influential soil factors were the total concentrations of micronutrients (Mn_{tot}, Zn_{tot}, Mo_{tot}, and B_{tot}) and phosphorus (P_{tot}). The effect of total forms of nutrients on the fungal community makes sense, since EMF are able to mobilise nutrients directly from insoluble mineral forms through the excretion of organic acids (Landeweert *et al.*, 2001). Soil C/N and Ca_{avail} also explained a portion of the community distribution. Nevertheless, these selected variables were significantly correlated to other soil variables (Table S4), and the mechanisms explaining the effect of soil on EMF might as well be indirectly linked to the total forms of the chemical elements presented above. Indeed, Mn_{tot} and Zn_{tot} were strongly correlated to the clay content, Al_{avail} K_{avail}, and P_{Olsen}. Differential response to Al toxicity, to K or P uptake, or to hydric stress (linked to the clay content) between host species could also be a mechanism influencing the symbiont community. In addition, P_{tot} and Zn_{tot} (and B_{tot} to a lesser extent) displayed a high correlation with soil nitrogen content, a crucial element which has been shown to influence EM fungal community composition (Lowell and Klein 2001; Treseder and Allen 2002; Avis *et al.*, 2003). Disentangling causal relationships between variables displaying such a high level of collinearity is difficult, especially when different ecological scenarios are likely to be correct (influence of total forms of nutrients, of soil texture, fertility...). Nevertheless, B_{tot} and Mo_{tot} were independent of most other variables. Molybdenum is known to be a key element for the nitrate reductase enzyme of mycorrhizal fungi (Marmeisse *et al.*, 1998), and for nitrogenase enzymes in N fixers (Silvester 1989), therefore constraining the acquisition of new nitrogen in some forests (Barron *et al.*, 2008) so that any EMF associated to Fabaceae (N fixers) could be expected to be positively correlated to Mo. Boron was shown to be actively uptaken by EMF (Lehto *et al.*, 2004), is known to be implied in the ectomycorrhizal symbiosis (Mitchell *et al.*, 1987, 1990; Lehto *et al.*, 2004; Smith *et al.*, 2015) and could have a direct influence on the fungal community.

At a scale of 200 m–375 m, host FTs reflecting the resource acquisition strategy (LA, SLA, and Mg_f) and some soil properties (mainly Mn_{tot}, Zn_{tot}, C/N, Ca_{avail}, B_{tot}, and P_{tot}) generated two nearly identical spatial patterns in the community (Figures 1 d and e and Figures 2 a and b). Coincidental correlation is unlikely to explain how good the match was between these patterns. Therefore, we suggest that the concerned FTs and soil parameters are linked and generated together the corresponding spatial patterns of the fungal community. Yet, one of the patterns of the fungal community (200 m–300 m) mainly related to the concentrations of B_{tot} and P_{tot}, was not generated by host FTs, so that an additional pure effect of soil on fungi is also likely to exist.

IV.2. Hidden ecological processes

Spatial patterns of the residuals of an RDA on the community can be utilised as surrogates for unmeasured ecological processes (McIntire and Fajardo 2009; Dray *et al.*, 2012). In the present study, variation partitioning analyses showed that broad (>200 m) and fine-scale (<100 m) ecological processes were approximately of the same magnitude. While all of the broad-scale variations in the community composition were related to soil or FTs of host trees, the fine-scale variation was mostly unexplained by the environmental variables measured. These unexplained fine-scale spatial patterns also appeared in the scalograms and R2Max statistics associated with residuals of the community data after environmental effects were partitioned out (Figures 1 g, h, and i). These structures could have been generated by dispersal limitation, biotic interactions, or unmeasured site heterogeneity. Dispersal limitation in EMF was shown to occur at scales of less than one to a few metres (Lilleskov *et al.*, 2004; Pickles *et al.*, 2010; Galante, Horton and Swaney 2011; Pickles *et al.*, 2012). On the other hand, Koide *et al.* (2005) showed that competitive interactions among EMF were absent at the scale of the plot, but highlighted competition at the scale of the centimetre, which is the only spatial scale at which biotic interactions among EMF have been shown (e.g., Pickles *et al.*, 2010, 2012; Hortal *et al.*, 2016). Since the residual patterns of the community occurred at scales of 75 m–100 m, it is unlikely they were caused by dispersal limitations or negative biotic interactions. Unmeasured abiotic or biotic heterogeneity (predation and/or symbiotic interactions) are possible drivers responsible for the community residuals patterns, as was already suggested previously (Matsuoka *et al.*, 2016). In particular, EMF depend on their symbiosis with a group of soil bacteria, the mycorrhization and mycorrhiza helper bacteria (MHB) (Frey-Klett, Garbaye and Tarkka 2007). These bacteria play a crucial role at several life stages of the EMF (spore germination, early nutrition of the fungus, mycorrhizal symbiosis establishment, abiotic stress resistance; Garbaye 1994; Frey-Klett, Garbaye and Tarkka 2007; Bonfante and Anca 2009), and were shown to be species-specific (Garbaye and Duponnois 1992), therefore enhancing ectomycorrhization by some fungal species while inhibiting others. As bacteria communities are spatially structured (Burns *et al.* 2015), the MHB species distribution could be a suitable candidate to explain the residual spatial patterns of the fungal EM community.

IV.3. EMF and stochasticity

The unexplained portion of the total variation in the community theoretically comes from its inherent stochasticity (i.e., ecological drift; Legendre and Legendre 2012) and depends on the organism of interest (Tedersoo *et al.*, 2011; Punchi-Manage *et al.*, 2013; Burns *et al.*, 2015). Nevertheless, this portion of variability also depends on several artefacts, such as a range of spatial scales undetectable by a given sampling design, and at which some spatial patterns of the community occur (Legendre *et al.*, 2009), and a limited effort of environmental characterisation (McIntire and Fajardo 2009). Here, a high number of relevant environmental parameters were used for the characterisation of soil and aboveground host FTs, and 97% (soil) and 78% (FT) of the environmental variables displayed their range of spatial variation within the range of scales detectable by our sampling design (Tables S2 and S3). This supports that the sampling design of the EMF was adapted to the ecological variables used to explain the community assemblage. The only variables for which a possible influence at finer scales than 70 m remains to be studied are the SI (soil variables), and LD, Al_f, Zn_f, foliar N, and lignin contents (FT variables). In this study, 24% of the variation in the dominant fungal EM community was successfully modelled. This is an expected amount of explanation considering that fungal communities typically display high β-diversity and a predominance of stochastic processes in the community assemblage (Toljander *et al.*, 2006; Peay, Garbelloto and Bruns 2010; Tedersoo *et al.*, 2011; Lekberg *et al.*, 2012; Waring *et al.*, 2015). The ecological drift, a possible residual undersampling stochasticity and ecological processes acting at scales finer than 70 m (e.g., dispersal limitation, among EMF competition, microsite heterogeneity) are likely to be responsible for the unexplained portion of community variability.

IV.4. Conclusions

The ectomycorrhizal fungal community of miombo dry woodland displayed a great diversity with a high number of rare OTUs. The β-diversity of EMF was correlated with both soil chemical properties and FTs of host trees linked to the leaf economics spectrum. Although environmental variables displayed high levels of collinearity, our selection procedure allowed choosing a limited number of variables, minimising the correlation between them and optimising the explanatory power. This way, we showed that among the most influential soil variables were the total forms of essential nutrients, the impact of which on the EMF community could either be direct or indirect through other parameters such as available Al, K, P or clay soil content, and 24% of the composition variation in the EM fungal community could be explained using a restricted number of environmental variables. A multiscale decomposition of the community β-diversity highlighted that both soil and aboveground host FTs influence the EMF community at scales > 200 m. Three spatial patterns of 75 m to 100 m remained after removing the effect of measured variables. After considering the spatial scales of the patterns and current knowledge about dispersion limitation and competition among EMF, we suggested that these patterns were generated by unmeasured abiotic or biotic heterogeneity. Future studies should aim at 1) better understanding how host FTs are related to soil parameters, 2) confirming that resource acquisition strategy of host trees influence EMF by testing belowground host FTs such as sugar type and production.

Acknowledgements

This work was supported by grants from the Belgian National Fund for Scientific Research (FNRS) and the Wallonie-Bruxelles International Agency (WBI) to DB. We are grateful to M. Anastassiou for his support and the accommodations at the Mikembo sanctuary. We thank A. Jacobs for his valuable help on the field, B. Kasongo for his assistance in the initial search of mycorrhiza, and W. Baert for the supervision of the molecular analyses. M. Devall is thanked for the proofreading of the manuscript.

Supporting information

See *Annexes* at the end of the thesis.

References

Ackerly DD, Cornwell W. 2007. A trait-based approach to community assembly: partitioning of species trait values into within-and among-community components. *Ecology Letters* 10: 135–145.

Allen MF. 1991. *The Ecology of Mycorrhizae.* Cambridge: University Press.

Anderson MJ, Crist TO, Chase JM *et al.* 2011. Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecology Letter* 14:19–28.

Avis PG, McLaughlin DJ, Dentinger BC *et al.* 2003. Long-term increase in nitrogen supply alters above- and below-ground ectomycorrhizal communities and increases the dominance of *Russula spp.* in a temperate oak savanna. *New Phytologist* 160(1):239–253.

Bâ A, Duponnois R, Diabaté M. *et al.* 2011. *Les champignons ectomycorhiziens des arbres forestiers en Afrique de l'Ouest : méthodes d'étude, diversité, écologie, utilisation en foresterie et comestibilité*. Marseille: IRD Editions.

Bâ A, Duponnois R, Moyersoen B *et al.* 2012. Ectomycorrhizal symbiosis of tropical African trees. *Mycorrhiza* 22:1–29.

Barron AR, Wurzburger N, Bellenger JP *et al.* 2008. Molybdenum limitation of asymbiotic nitrogen fixation in tropical forest soils. *Nature Geoscience* 2(1):42–45.

Blanchet FG, Legendre P, Borcard D. 2008. Forward selection of explanatory variables. *Ecology* 89:2623–2632.

Bonfante P, Anca I-A. 2009. Plants, mycorrhizal fungi, and bacteria: a network of interactions. *Annual review of microbiology* 63:363–*383*.

Borcard D, Legendre P, Drapeau P. 1992. Partialling out the spatial component of ecological variation. *Ecology* 73(3):1045–1055.

Borcard D, Legendre P. 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling* 153(1-2):51–68.

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

Borcard D, Gillet F, Legendre P. 2011. *Numerical ecology with R.* New-York: Springer.

Branco S, Bruns TD, Singleton I. 2013. Fungi at a Small Scale: Spatial Zonation of Fungal Assemblages around Single Trees. *PLoS ONE* 8(10):e78295.

Bruns TD, Bidartondo MI, Taylor LD. 2002. Host Specificity in Ectomycorrhizal Communities: What Do the Exceptions Tell Us? Integrative and Comparative Biology 42(2):352–359.

Burns JH, Anacker BL, Strauss SY *et al.* 2015. Soil microbial community variation correlates most strongly with plant species identity, followed by soil chemistry, spatial location and plant genus. *AoB Plants* 7:plv030.

Chang LW, Zeleny D, Li CF *et al.* 2013. Better environmental data may reverse conclusions about niche- and dispersal-based processes in community assembly. Ecology 94(10):2145–51.

Chase JM. 2014. Spatial scale resolves the niche versus neutral theory debate. *Journal of Vegetation Science* 25:319–322.

Clifford P, Richardson S, Hemon D. 1989. Assessing the Significance of the Correlation between Two Spatial Processes. *Biometrics* 45:123–134.

Cornelissen JHC, Lavorel S, Garnier E *et al.* 2003. A handbook of protocols for standardised and easy measurement of plant functional traits worldwide. *Australian Journal of Botany* 51:335–380.

den Bakker HC, Zuccarello GC, Kuyper TW *et al.* 2004. Evolution and host specificity in the ectomycorrhizal genus *Leccinum. New Phytologist* 163:201–215.

Dickie IA, Alexander I, Lennon S *et al.* 2015. Evolving insights to understanding mycorrhizas. *New Phytologist* 205:1369–1374.

Dormann CF, Elith J, Bacher S *et al.* 2013. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36:27–46.

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

Dray S, Pélissier R, Couteron P *et al.* 2012. Community ecology in the age of multivariate multiscale spatial analysis. *Ecological Monographs* 2012;82:257–275.

Felten J, Martin F, Legué V. 2012. Signalling in Ectomycorrhizal Symbiosis. In: Perotto S, Baluška (eds.) *Signaling and Communication in Plant Symbiosis*, Heidelberg: Springer Berlin, 123–142.

Fortin M-J, Dale MRT. 2005. *Spatial analysis: a guide for ecologists*. Cambridge: University Press, 2005.

Frey-Klett P, Garbaye J, Tarkka M. 2007. The mycorrhiza helper bacteria revisited. *New Phytologist* 176:22–36.

Galante TE, Horton TR, Swaney DP. 2011. 95% of basidiospores fall within 1 m of the cap: a field- and modeling-based study. *Mycologia* 103:1175–1183.

Garbaye J. 1994. Helper bacteria: a new dimension to the mycorrhizal symbiosis (Tansley Review, 76). *New Phytologist* 128:197–210.

Garbaye J, Duponnois R. 1992. Specificity and function of mycorrhization helper bacteria (MHB) associated with the *Pseudotsuga menziesii–Laccaria laccata* symbiosis. *Symbiosis* 14:335–344.

Garzon-Lopez CX, Jansen PA, Bohlman SA *et al.* 2013. Effects of sampling scale on patterns of habitat association in tropical trees. *Journal of Vegetation Science* 25:349–362.

Högberg P, Piearce G. 1986. Mycorrhizas in Zambian trees in relation to host taxonomy, vegetation type and successional patterns. *The Journal of Ecology* 74:775–785.

Hortal S, Powell JR, Plett JM *et al.* 2016. Intraspecific competition between ectomycorrhizal Pisolithus microcarpus isolates impacts plant and fungal performance under elevated CO2 and temperature. *FEMS Microbiology Ecology* fiw113.

Horton TR, Bruns TD. 2001. The molecular revolution in ectomycorrhizal ecology: peeking into the black-box. *Molecular ecology* 10(8):1855–1871.

Innis MA, Gelfand DH, Sninsky JJ *et al.* 1990. *PCR protocols: a guide to methods and applications*. San Diego: Academic Press.

Koide RT, Xu B, Sharda J *et al.* 2005. Evidence of species interactions within an ectomycorrhizal fungal community. *New Phytologist* 165(1):305–316.

Koide RT, Shumway DL, Xu B. *et al.* 2007. On temporal partitioning of a community of ectomycorrhizal fungi. *New Phytologist* 174:420–429.

Kramer-Walter KR, Bellingham PJ, Millar TR *et al.* 2016. Root traits are multidimensional: specific root length is independent from root tissue density and the plant economic spectrum. *Journal of Ecology* doi: 10.1111/1365-2745.12562.

Kretzer AM, Bruns TD. 1999. Use of atp6 in fungal phylogenetics: an example from the Boletales. *Molecular Phylogenetics and Evolution* 13:483–492.

Landeweert R, Hoffland E, Finlay RD *et al.* 2001. Linking plants to rocks: Ectomycorrhizal fungi mobilize nutrients from minerals. *Trends in Ecology and Evolution* 16:248–254.

Legendre P, Gallagher E. 2001. Ecologically meaningful transformations for ordination of species data. *Oecologia* 129:271–280.

Legendre P, Borcard D. 2006. Quelles sont les échelles spatiales importantes dans un écosystème. In: Droesbeke JJ, Lejeune M, Saporta G (eds) *Analyse statistique de donnees spatiales*, Paris: TECHNIP, 425–442.

Legendre P, Mi X, Ren H *et al.* 2009. Partitioning beta diversity in a subtropical broad-leaved forest of China. Ecology 90:663–74.

Legendre P, Legendre L. 2012. *Numerical Ecology*. 3rd English edn. Amsterdam: Elsevier.

Lehto T, Lavola A, Kallio E *et al.* 2004. Boron uptake by ectomycorrhizas of silver birch. *Mycorrhiza* 14(3):209-212.

Lekberg Y, Schnoor T, Kjøller R *et al.* 2012. 454-sequencing reveals stochastic local reassembly and high disturbance tolerance within arbuscular mycorrhizal fungal communities. *Journal of Ecology* 100:151–160.

Lilleskov EA, Bruns TD, Horton TR *et al.* 2004. Detection of forest stand-level spatial structure in ectomycorrhizal fungal communities. *FEMS Microbiology Ecology* 49:319–332.

Lowel JL, Klein DA. 2001. Comparative single-strand conformation polymorphism (SSCP) and microscopy-based analysis of nitrogen cultivation interactive effects on the fungal community of a semiarid steppe soil. *FEMS Microbiology Ecology* 365(2-3);85–92.

Marmeisse R, Jargeat P, Wagner F *et al.* 1998. Isolation and characterization of nitrate reductase deficient mutants of the ectomycorrhizal fungus *Hebeloma cylindrosporum*. *New Phytologist* 140(2): 311–318.

Matsuoka S, Mori AS, Kawagushi E *et al.* 2016. Disentangling the relative importance of host tree community, abiotic environment, and spatial factors on ectomycorrhizal fungal assemblages along an elevation gradient. *FEMS Microbiology Ecology* 92(5):fiw044.

McIntire EJB, Fajardo A. 2009. Beyond description: The active and effective way to infer processes from spatial patterns. *Ecology* 90:46–56.

Meloun M, Militky J, Hill M *et al.* 2002. Crucial problems in regression modelling and their solutions. *Analyst* 127:433–450.

Min XJ, Hickey DA. 2007. Assessing the effect of varying sequence length on DNA barcoding of fungi. *Molecular Ecology Notes* 7(3):365–373.

Mitchell RJ, Garrett HE, Cox GS *et al.* 1987. Boron fertilization, ectomycorrhizal colonization, and growth of Pinus echinata seedlings. Canadian Journal of Forest Research 17(10):1153-1156.

Mitchell RJ, Garrett HE, Cox GS *et al.* 1990. Boron and ectomycorrhizal influences on mineral nutrition of container-grown Pinus ehinata mill. Journal of plant nutrition 13(12):1555-1574.

Molina R, Massicotte H, Trappe JM. 1992. Specificity phenomena in mycorrhizal symbiosis: community-ecological consequences and practical implications. In: Allen M (ed). *Mycorrhizal Functioning: an Integrative Plant-Fungal Process*. New York: Chapman and Hall, 357–423.

Munoz F. 2009. Distance-based eigenvector maps (DBEM) to analyse metapopulation structure with irregular sampling. *Ecological Modelling* 220:2683–2689.

Miyamoto Y, Sakai A, Hattori M *et al.* 2015. Strong effect of climate on ectomycorrhizal fungal composition: evidence from range overlap between two mountains. ISME J 9:1870–9.

Ollier S, Couteron P, Chessel D. 2006. Orthonormal transform to decompose the variance of a life-history trait across a phylogenetic Tree. *Biometrics* 2006;62:471–477.

Pansu M, Gautheyrou J. 2007. *Handbook of Soil Analysis-Mineralogical. Organic and Inorganic Methods*. Berlin: Springer.

Peay KG, Garbelloto M, Bruns TD. 2010. Evidence of dispersal limitation in soil microorganisms: Isolation reduces species richness on mycorrhizal tree islands. *Ecology* 91:3631–3640.

Peay KG, Baraloto C, Fine PVAA. 2013. Strong coupling of plant and fungal community structure across western Amazonian rainforests. *ISME J* 27(9):1852–1861.

Peay KG, Russo SE, McGuire KL *et al.* 2015. Lack of host specificity leads to independent assortment of dipterocarps and ectomycorrhizal fungi across a soil fertility gradient. *Ecology Letters* 18(8):807–816.

Peres-Neto PR, Legendre P, Dray S. *et al.* 2006. Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology* 87:2614–2625.

Phorsi C, Põlme S, Taylor AFS *et al.* 2012. Diversity and community composition of ectomycorrhizal fungi in a dry deciduous dipterocarp forest in Thailand. *Biodiversity and Conservation* 21:2287–2298.

Pickles BJ, Genney DR, Potts JM *et al.* 2010. Spatial and temporal ecology of Scots pine ectomycorrhizas. *New Phytologist* 186(3):755–768.

Pickles BJ, Genney DR, Anderson IC. 2012. Spatial analysis of ectomycorrhizal fungi reveals that root tip communities are structured by competitive interactions. *Molecular Ecology* 21(50):5110–5123.

Platt T, Denman KL. 1975. Spectral analysis in ecology. *Annual Review of Ecology and Systematics* 6:189–210.

Punchi-Manage R, Getzin S, Wiegand T *et al.* 2013. Effects of topography on structuring local species assemblages in a Sri Lankan mixed dipterocarp forest. *Journal of Ecology* 101:149–160.

R Development Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna.

Raspé O, Degreef J, De Kesel A. 2012. *DNA barcoding of African Boletes using the mitochondrial gene ATP6*. 3rd European Conference for the Barcode of Life, Brussels, Belgium, 17-20 September 2012.

Raspé O, Vadthanarat S, De Kesel A *et al.* 2016. *Pulveroboletus fragrans*, a new Boletaceae species from Northern Thailand, with a remarkable aromatic odor. *Mycological Progress* 15(4):1–8.

Reich PB. 2014. The world-wide 'fast–slow' plant economics spectrum: a traits manifesto. *Journal of Ecology* 102:275–301.

Roy M, Dubois MP, Proffit M *et al.* 2008. Evidence from population genetics that the ectomycorrhizal basidiomycete *Laccaria amethystina* is an actual multihost symbiont. *Molecular Ecology* 17:2825–2838.

Schmidt PA, Bálint M, Greshake B *et al.* 2013. Illumina metabarcoding of a soil fungal community. *Soil Biology and Biochemistry* 65:128–132.

Schoch CL, Seifert KA, Huhndorf S *et al.* 2012. Fungal Barcoding Consortium (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the USA* 109:6241–6246.

Silvester WB. 1989. Molybdenum limitation of asymbiotic nitrogen fixation in forests of Pacific Northwest America. *Soil Biology and Biochemistry* 21(2):283-289.

Smith AJH, Potvin LR, Lilleskov EA. 2015. Fertility-dependent effects of ectomycorrhizal fungal communities on white spruce seedling nutrition. *Mycorrhiza* 25(8):649–662.

Smith ME, Douhan GW, Rizzo DM. 2007. Intra-specific and intra-sporocarp ITS variation of ectomycorrhizal fungi as assessed by rDNA sequencing of sporocarps and pooled ectomycorrhizal roots from a Quercus woodland. *Mycorrhiza* 18:15–22.

Smith ME, Douhan GW, Fremier AK *et al.* 2009. Are true multihost fungi the exception or the rule? Dominant ectomycorrhizal fungi on *Pinus sabiniana* differ from those on co-occurring *Quercus* species. *New Phytologist* 182:295-299.

Smith SE, Read DJ. 2008. *Mycorrhizal Symbiosis*. Amsterdam: Elsevier Science.

Talbot JM, Bruns TD, Taylor JW *et al.* 2014. Endemism and functional convergence across the North American soil mycobiome. *Proceedings of the National Academy of Sciences of the USA* 111(8):6341-6346.

Taylor AFS. 2002. Fungal diversity in ectomycorrhizal communities: sampling effort and species detection. Plant Soil 244(1-2):19–28.

Tedersoo L, Suvi T, Larsson E *et al.* 2006. Diversity and community structure of ectomycorrhizal fungi in a wooded meadow. *Mycological Research* 110:734–748.

Tedersoo L, Sadam A, Zambrano M *et al.* 2010a. Low diversity and high host preference of ectomycorrhizal fungi in Western Amazonia, a neotropical biodiversity hotspot. *ISME J* 4:465–471.

Tedersoo L, Nilsson RH, Abarenkov K *et al.* 2010b. 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytologist* 188:291–301.

Tedersoo L, Bahram M, Jairus T *et al.* 2011. Spatial structure and the effects of host and soil environments on communities of ectomycorrhizal fungi in wooded savannas and rain forests of Continental Africa and Madagascar. *Molecular Ecology* 20:3071–3080.

Tedersoo L, Bahram M. 2012. Towards global patterns in the diversity and community structure of ectomycorrhizal fungi. *Molecular Ecology* 21:4160–4170.

Tedersoo L, Anslan S, Bahram M *et al.* 2015. Shotgun metagenomes and multiple primer pairbarcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. *MycoKeys* 10:1–43.

Toljander JF, Eberhardt U, Toljander YK *et al.* 2006. Species composition of an ectomycorrhizal fungal community along a local nutrient gradient in a boreal forest. *New Phytologist* 170:873–884.

Treseder KK, Allen MF. 2002. Direct nitrogen and phosphorus limitation of arbuscular mycorrhizal fungi: a model and field test. *New Phytologist* 155(3):507–515.

Tsamba J, Kativu S, Sithole-Niang I. 2015. Diversity and host associations of ectomycorrhizae fungi in fallow lands of the mid-Zambezi valley area, Zimbabwe. *Transactions of the Royal Society of South Africa* 70:71–77.

van der Heijden MGA, Martin FM, Selosse MA *et al.* 2015. Mycorrhizal ecology and evolution: the past, the present, and the future. *New Phytologist* 205:1406–1423.

Verbeken A, Buyck B. 2002. Diversity and ecology of tropical ectomycorrhizal fungi in Africa. In: Watling R, Frankland AM, Ainsworth AM *et al.* (eds.). *Tropical mycology 1*, New-York: CABI, 11–24.

Vialle A, Feau N, Allaire M *et al.* 2009. Evaluation of mitochondrial genes as DNA barcode for Basidiomycota. *Molecular Ecology Resources* 9(S1):99–113.

Waring BG, Adams R, Branco S *et al.* 2015. Scale-dependent variation in nitrogen cycling and soil fungal communities along gradients of forest composition and age in regenerating tropical dry forests. *New Phytologist* 209(2):845–854.

Chapitre III

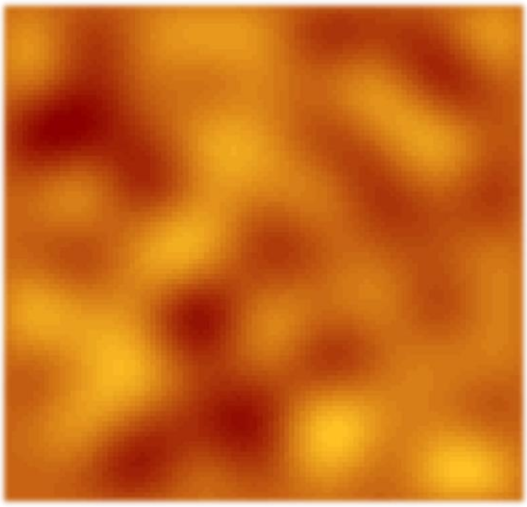
Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors

David Bauman, Thomas Drouet, Stéphane Dray, and Jason Vleminckx

Published December 2017 in *Ecography*, vol. 41(10): 1638–1649.

Dans le chapitre précédent, des prédicteurs spatiaux associés à une partition de la variation ont mis en évidence un jeu d'interactions complexes entre variables de sol, la communauté d'arbres et la communauté fongique ectomycorhizienne associée dans une forêt claire tropicale. Ce cadre de travail a également permis l'estimation de l'importance relative de la sélection et des processus neutres structurés (dispersion) et non structurés (dérive). Enfin, cette analyse a rendu possible une visualisation des patterns spatiaux de la communauté fongique non expliqués par l'environnement.

Dans les Chapitre III, IV et V, ce cadre méthodologique basé sur l'utilisation de vecteurs propres spatiaux (*Moran's eigenvector maps*, MEM) et de la partition de variation sera approfondi sur base de simulations. L'appel à des simulations permettra de générer des distributions d'espèces à partir d'un ou plusieurs processus écologiques afin de reproduire une gamme de scénarios réalistes de distribution d'espèces dont les processus influents seront contrôlés. Les conséquences de différentes étapes clés de ces méthodes spatiales sur l'inférence de processus seront ainsi étudiées.
Dans le Chapitre III, les simulations offriront un cadre d'évaluation des performances statistiques des trois méthodes les plus courantes de sélection de prédicteurs spatiaux, parmi lesquelles la sélection basée sur un critère d'information (AIC), utilisée jusqu'à présent dans de nombreuses études et dans la partition de variation du chapitre précédent.



Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors

Abstract

Eigenvector mapping techniques are widely used by ecologists and evolutionary biologists to describe and control for spatial and/or phylogenetic patterns in their data. The selection of an appropriate subset of eigenvectors is a critical step (misspecification can lead to highly biased results and interpretations), and there is no consensus yet on how to proceed. We conducted a ten-year review of the practices of eigenvector selection and highlighted three main procedures: selecting the subset of descriptors minimising the Akaike information criterion (AIC), using a forward selection with double stopping criterion after testing the global model significance (FWD), and selecting the subset minimising the autocorrelation in the model residuals (MIR). We compared the type I error rates, statistical power, and R^2 estimation accuracy of these methods using simulated data. Finally, a real dataset was analysed using variation partitioning analysis to illustrate to what extent the different selection approaches affected the ecological interpretation of the results. We show that, while the FWD and MIR approaches presented a correct type I error rate and were accurate, the AIC approach displayed extreme type I error rates (100%), and strongly overestimated the R^2. Moreover, the AIC approach resulted in wrong ecological interpretations, as it overestimated the pure spatial fraction (and the joint spatial-environmental fraction to a lesser extent) of the variation partitioning. Both the FWD and MIR methods performed well at broad and medium scales but had a very low power to detect fine-scale patterns. The FWD approach selected more eigenvectors than the MIR approach but also returned more accurate R^2 estimates. Hence, we discourage any future use of the AIC approach, and advocate choosing between the MIR and FWD approaches depending on the objective of the study: controlling for spatial or phylogenetic autocorrelation (MIR) or describing the patterns as accurately as possible (FWD).

Keywords

Akaike information criterion (AIC), ecological processes, eigenvector selection, forward selection, Moran's eigenvector maps (MEM), phylogenetic eigenvector regression (PVR), principal coordinates of neighbour matrices (PCNM), spatial eigenvector mapping (SEVM), spatial patterns, variation partitioning

I. Introduction

Identifying and explaining spatial structures in the distribution of organisms in natural communities are two longstanding challenges in ecology (Cormack and Ord 1979, Legendre and Fortin 1989, Legendre and Legendre 2012). Clustered distributions may result from spatially heterogeneous habitat filters, community intrinsic processes (e.g., dispersal limitation), historical processes or biotic interactions (Legendre and Fortin 1989, Legendre 1993, Barbier et al. 2008, Pinto and MacDougall 2010). Disentangling the relative influence of these processes is a challenging task. Indeed, even if only community-intrinsic processes occur, spatially clustered distributions may overlap by chance with spatially structured environmental variables (‘spatial nuisance’, Peres-Neto and Legendre 2010). On the other hand, spatial patterns of species distribution may give crucial clues about possible underlying ecological processes (McIntire and Fajardo 2009), a view in which space is not considered an issue but a surrogate of the underlying ecological processes ('spatial legacy', Peres-Neto and Legendre 2010, Legendre and Legendre 2012). These paradigms of ‘nuisance’ and ‘legacy’ also apply for temporal and phylogenetic patterns (Diniz-Filho and Bini 2005, Griffith and Peres-Neto 2006, Diniz-Filho et al. 2012, Baho et al. 2015).

Several methods have been developed to overcome the ‘nuisance’ issue or properly detect the ‘legacy’, depending on the objectives of the study. Among others, eigenvector-based methods produce predictors that can be used either to (1) control for autocorrelation in model residuals, or (2) highlight and interpret the patterns of autocorrelation. In the spatial context, Griffith (1996) provided pioneering work in the geographic literature showing how spatial eigenvectors can be used to describe map patterns or control for spatial autocorrelation in residuals. In the same way, evolutionary biologists have been using phylogenetic eigenvector regression (PVR; Diniz-Filho et al. 1998, 2012) to control for phylogenetic autocorrelation in comparative analysis or to study patterns of trait evolution along a phylogeny (see Diniz-Filho et al. 2012 for more details). This paper focuses on spatial eigenvectors, but the general conclusions probably hold for the phylogenetic context.

Eigenvector-based methods have been receiving increasing attention in the past decade in community ecology, especially with the development of Moran’s eigenvector maps (MEM, Dray et al. 2006). This method generalises the ad hoc principal coordinates of neighbour matrices (PCNM, Borcard and Legendre 2002, Griffith and Peres-Neto 2006) that consist of the principal coordinate analysis of a truncated geographic distance. Whereas the initial definition of PCNM suffers from a lack of statistical formalism (e.g., the empirical basis of the definition of the truncated Euclidean distance matrix, or the scaling of the PCNM base functions), MEM provides a well-defined theoretical framework based on the notion of spatial autocorrelation, which encompasses PCNM, and will thus be used in the rest of the paper (see Dray et al. 2006 for further details).

The first step to use MEM is to define the spatial relationships between sampling sites using a spatial weighting matrix (**W**). The latter is built by computing the Hadamard product of a connectivity (**B**) and weighting (**A**) matrix (Dray et al. 2006). As various candidates can be considered for the **B** and **A** matrices, a broad range of **W** matrices can be generated (see Dray et al. 2006 for details). Even if the choice of a spatial weighting matrix is a crucial issue, this work does not deal with this question, and the following developments consider that this step has been performed properly (although there currently is no consensus method for the choice of a **W** matrix, see discussion). The **W** matrix is then doubly centred (i.e., by rows and columns) and diagonalised. This last step generates n-1 eigenvectors (where n is the number of sampling sites) associated with eigenvalues directly proportional to Moran’s

I coefficients. The first eigenvector has the highest Moran's *I* value and thus models broader structures than the second eigenvector, itself displaying a broader spatial pattern than the third eigenvector, and so on. These eigenvectors (hereafter referred to as MEM or MEM variables; see the MEM.ALL matrix in Fig. 1) can then be used 1) as predictors to study the multiscale spatial patterns of the response data (e.g., species abundances; Declerck et al. 2011, Bauman et al. 2016, Vleminckx et al. 2017), or 2) as covariables (spatial filters) to remove spatial autocorrelation from residuals (spatial eigenvector mapping, SEVM; Griffith 2003, Diniz-Filho and Bini 2005, Dormann et al. 2007, Bini et al. 2009, Corkeron et al. 2011). However, all eigenvector-based methods, including MEM, face a common issue. Specifically, a variable selection is required to avoid model overfitting and a loss of statistical power to detect the environmental contribution to the variability of the response data (Griffith 2003, Dray et al. 2006, Blanchet et al. 2008, Peres-Neto and Legendre 2010, Diniz-Filho et al. 2012). Although several procedures have been proposed relatively separately in the scopes of MEM, SEVM, or PVR, there is still no consensus on a most suitable eigenvector selection method. The goal of this study is to test and compare the most common eigenvector selection methods to define a guideline of the best practices, and rule out some biased and underpowered approaches.

I.1. Selecting a subset of eigenvectors

A review of the literature (see below) showed that three main approaches to select spatial variables have been widely used in the framework of eigenvector-based methods.

The first method, further referred to as the 'AIC approach', originates from Dray et al. (2006). This procedure relies on the computation of a corrected version of the Akaike information criterion (AICc; Hurvich and Tsai 1989) extended to the case of multivariate response data (Godinez-Dominguez and Freire 2003). This approach starts by reordering the complete set of MEM (MEM.ALL) by decreasing (canonical) coefficient of determination (R2; Step 3.1 in Fig. 1). The reordered MEM are then added one by one in an explanatory model, and the AICc is computed at each step. The subset of MEM corresponding to the minimum AICc (MEM.AIC in Fig. 1) is then selected.
The AIC approach was originally developed by Dray et al. (2006) to select a spatial weighting matrix **W** among a range of candidates. However, it has been utilised to select a subset of MEM in nearly all the works using the method since 2006 (see 'Results'). This misuse emerged from the subtlety that, although the method selects a subset of MEM (Step 4 in Fig. 1), the only purpose of the latter is to detect the best **W** matrix. Contributing to this ambiguity, Borcard et al. (2011) defined the 'champion model' as the MEM subset best minimising the $\mathrm{AIC_c}$ among the different **W** matrices tested. Hence, they recommended the use of the AIC approach both for selecting a **W** matrix and a 'very best model' within it (and provided an R code to do so). Our study will therefore consider this misuse of the procedure and will assess its suitability for spatial eigenvector selection.

The second approach, proposed by Blanchet et al. (2008) and further referred as the 'FWD approach', is based on a forward selection with double stopping criterion. The method selects a subset of explanatory variables while maintaining a correct type I error rate and avoids model overfitting (MEM.FWD, Step 3.2 in Fig. 1, also presented in Borcard et al. 2011). Blanchet et al. (2008) showed that the type I error rate of the stepwise procedure was correct if and only if a global test of significance was conducted previous to applying the forward selection (i.e., testing the model containing all MEM variables). Moreover, adding to the classical p-value stopping criterion of the forward selection, Blanchet et al. (2008) introduced the adjusted coefficient of determination (R^2_{adj}) of the global model as additional criterion, preventing the selection of models from explaining more than the complete set of explanatory variables, and showed that it allowed avoiding an inflated R^2_{adj}. If both positive and negative spatial autocorrelations are of interest, then the method is performed separately on two subsets of MEM (displaying positive and negative eigenvalues, respectively), and is followed by a *p*-value correction for multiple tests (Blanchet et al. 2008).

The third most widely used variable selection method consists of selecting the smallest subset of predictors that best minimises the spatial autocorrelation in model residuals (Griffith and Peres-Neto 2006, Tiefelsdorf and Griffith 2007). It was shown to be a reliable and efficient way to control spatial (Bini et al. 2009) and phylogenetic autocorrelation (Diniz-Filho et al. 2012), and to maintain a high statistical power. The method consists of removing spatial effects from the residuals of explanatory models (e.g., linear models) that relate a univariate response variable to a set of explanatory predictors (environmental factors, for instance). The method starts by computing the autocorrelation (Moran's *I*) of the residuals and tests its significance by a permutation procedure. If the test is significant, each MEM is added separately to the regression model and the Moran's *I* of the residuals is recomputed for each model. The procedure is repeated until the residuals display no more significant spatial autocorrelation. This minimisation of Moran's *I* in the residuals (MIR approach) both ensures the independence of the residuals and maintains a high statistical power for the explanatory variables of interest as it adds as few MEM variables as possible (Griffith and Peres-Neto 2006). Whereas the FWD and AIC approaches are designed for studies where space is considered a legacy (their criteria maximise the spatial fit of the model), the MIR approach is oriented towards the spatial nuisance viewpoint (MEM used as covariables or spatial filters; e.g., Corkeron et al. 2011, Siesa et al. 2011).

In this paper, we decided to compare the AIC, FWD, and MIR approaches only in a context where space is considered a legacy. This choice is justified because our review demonstrates that these three methods are the most used in the literature and that the MIR approach has also been applied for spatial legacy purposes even if it was not designed for this objective. To test the suitability of this method for a 'spatial legacy approach', we slightly adapted the method so that no environmental variables were needed. To do so, we introduce the MEM variables by the procedure described above in a model that contains only an intercept term (Step 3.3 in Fig. 1).

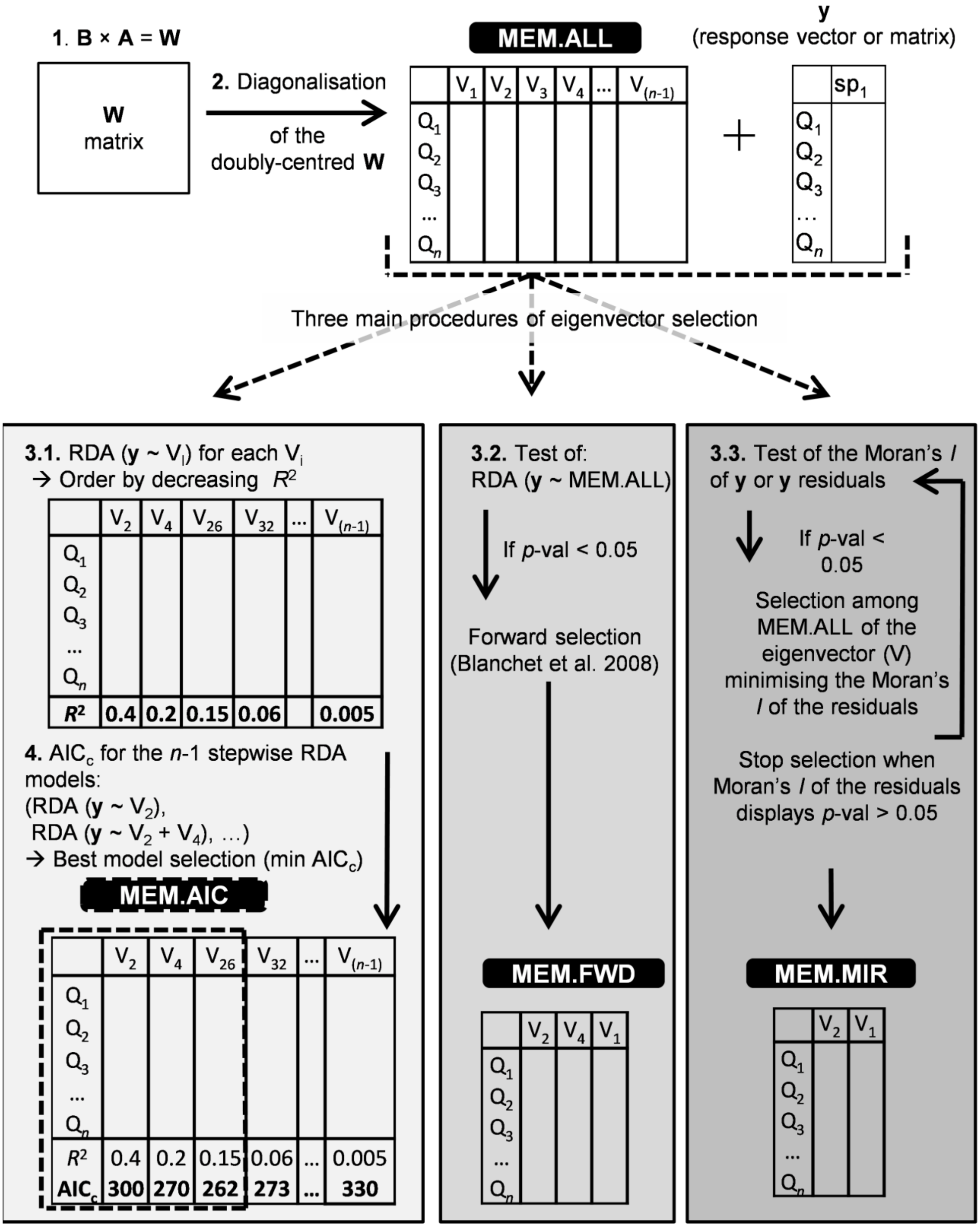

1. B × A = W
W matrix
2. Diagonalisation of the doubly-centred W
MEM.ALL
V1 V2 V3 V4 … V(n-1)
Q1 Q2 Q3 … Qn
+
y (response vector or matrix)
sp1
Three main procedures of eigenvector selection
3.1. RDA (y ~ Vi) for each Vi → Order by decreasing R2
V2 V4 V26 V32 … V(n-1)
R2 0.4 0.2 0.15 0.06 0.005
4. AICc for the n-1 stepwise RDA models: (RDA (y ~ V2), RDA (y ~ V2 + V4), …) → Best model selection (min AICc)
MEM.AIC
R2 0.4 0.2 0.15 0.06 … 0.005
AICc 300 270 262 273 … 330
3.2. Test of: RDA (y ~ MEM.ALL)
If p-val < 0.05
Forward selection (Blanchet et al. 2008)
MEM.FWD
V2 V4 V1
3.3. Test of the Moran's I of y or y residuals
If p-val < 0.05
Selection among MEM.ALL of the eigenvector (V) minimising the Moran's I of the residuals
Stop selection when Moran's I of the residuals displays p-val > 0.05
MEM.MIR
V2 V1

Figure 1: Schematic illustration of the three selection procedures for a univariate response vector y. Q: quadrat; V: spatial, temporal or phylogenetic eigenvector; sp: species; RDA: redundancy analysis. The first step corresponds to the Hadamard product of a connectivity (B) and weighting matrix (A). The spatial weighting matrix W is then doubly centred and diagonalised, leading to the MEM.ALL matrix, a matrix of *n*-1 eigenvectors. At this point, one model selection has to be used to reduce the number of eigenvectors. 1) Step 3.1. In the AIC approach of Dray et al. (2006), the *R*² of each MEM is computed by constrained ordinations, most often a redundancy analysis (RDA; multivariate response), or linear regression models (univariate response), and the MEM variables are ordered by decreasing *R*². At step 4, the ordered MEM are included one by one to the model and an AIC$_c$ value is computed at each step. The subset of MEM corresponding to the min AIC$_c$, although initially designed only to select a W matrix, is then used as the best spatial model of the corresponding W matrix (MEM.AIC). 2) In the forward selection developed by Blanchet et al. (2008; step 3.2.), a global significance test is computed with all MEM (associated to positive and negative eigenvalues separately, using a *p*-value correction for multiple tests). If the global model is significant, a forward selection is performed with the *p*-value and global adjusted *R*² as

stopping criteria, providing a subset of MEM (MEM.FWD) to be used for further analyses. 3) The step 3.3. corresponds to the MIR approach (Griffith and Peres-Neto 2006), and aims at finding the smallest subset of MEM variables minimising the autocorrelation (Moran's *I*) in the residuals. The method first tests the significance of the Moran's of y with a permutation test, and if significant, looks for the predictor best minimising it. The significance of the Moran's *I* is then recomputed, and if significant, the procedure searches for the next eigenvector best minimising the Moran's *I* of the residuals of the model of y as a function of the first added eigenvector. The procedure goes on until the Moran's *I* is not significant anymore.

Other criteria were proposed but are much less used (e.g., using all MEM, only those significantly related to the response data, displaying a Moran's *I* or R^2 greater than a given threshold value, displaying a significant Moran's *I* value (permutation test), or a combination of these criteria).

Our study is structured into three sections. We present a representative ten-year review (2006-2016) of the peer-reviewed literature using eigenvector-based spatial methods. Then, we use simulated data to evaluate the type I error rate, statistical power, and R^2 estimation accuracy of the AIC, FWD, and MIR procedures. Lastly, we explore a real dataset to compare to what extent the different selection methods can affect the results and their ecological interpretation.

II. Material and methods

II.1. Review of the literature

The review aimed to give a representative view of the selection practices of spatial predictors in the past 10 years. A research of the peer-reviewed articles was carried out with the terms 'Moran's eigenvector maps', 'Moran eigenvector maps', and 'spatial eigenvector mapping' in separate Google Scholar searches (on January 2017). We limited the research to these two criteria, as they covered the principal selection methods and consisted of a large number of articles (301). All 'Material and Methods' sections of the articles reviewed were carefully analysed to determine how the spatial predictors were selected.

II.2. Spatial eigenvector selection: type I error rate

We considered a study area of 1000 × 500 cells for both a regular (117 sites positioned on a 13 x 9 grid covering the entire area) and a random sampling design (117 randomly chosen cells, supplementary material, Fig. A1). The type I error rates were computed for the three selection approaches (Fig. 1) by considering a random univariate response variable **y** generated from four different distributions: uniform, normal, exponential, and cubed exponential, following Anderson and Legendre (1999) and Manly (2007).

We built a spatial weighting matrix **W** (Fig. 1) generated with distance-based criteria to define the **B** and **A** matrices. Two sites were considered connected if they were found within a threshold distance corresponding to the shortest distance that keeps all sites connected (i.e., the length of the largest edge of the minimum spanning tree). The links were then weighted by $f(D_{ij}) = 1-(D_{ij}/4t)^2$, where D_{ij} is the Euclidean distance between sites *i* and *j*, and *t* is the threshold value. The MEM predictors were then obtained by the diagonalisation of the doubly centred spatial weighting matrix **W**. This produced 116 eigenvectors (MEM.ALL in Fig. 1). Dray et al. (2006) demonstrated that this particular case of specification of the **W** matrix produced distance-based MEM (db-MEM) that correspond roughly to the original PCNM method. The main difference is that the PCNM approach is originally defined by computing a principal coordinate analysis of a truncated Euclidian distance matrix. Thus, it does not return the full set of *n*-1 eigenvectors and does not respect the strict equivalence between eigenvalues and Moran's index of autocorrelation (more details in Dray et al. 2006).

The three selection methods (FWD, AIC, MIR; Fig. 1) were then applied on the full set of spatial predictors (MEM.ALL). The tests of significance were performed with 999 permutations, and 10000 simulations were conducted for each scenario. Type I error rates were computed as the proportions of significant results (for a significance level of 0.05) among the 10000 simulations.

II.3. Spatial eigenvector selection: power and accuracy

To test the power and a possible bias of the R^2_{adj}, the three selection methods were applied in a second set of simulations. We used the same simulation design as for estimating type I error rates except for building the response variable **y**. Instead of using random numbers, we built spatially positively autocorrelated **y** as the sum of a linear combination of three MEM variables with a random normal noise, following Jombart et al. (2009; details in Appendix A1). We considered three cases where the response variable **y** was structured either at broad (first three MEM), medium (MEM 25 to 27), or fine spatial scale (the last three positively autocorrelated MEM; Appendix A1), for both the regular and random sampling designs. Then, a linear regression explaining **y** by the three corresponding MEM variables was computed, and the resulting R^2_{adj} was considered the reference value of the spatial signal contained in **y**. The three selection approaches were then applied separately for selecting the subset of MEM variables that best explained **y**, and the adjusted R^2 of the resulting models were compared to the reference value, providing a ΔR^2 ($R^2_{\text{AIC, FWD, or MIR}} - R^2_{\text{reference}}$). This procedure was repeated 10000 times. The power was then computed as the proportion of simulations leading to a significant model (significance level of 0.05), while the mean of ΔR^2 was used as a measure of the accuracy of the selection approaches. All R^2 values presented in this study are adjusted R^2 (Ezekiel 1929). Also note that, in this work, 'scale' exclusively refers to the spatial feature 'focus', *sensu* Scheiner (2011), that is, the dimension of the aggregated grains of a spatial pattern, and is never used to refer to the 'extent' or 'grain' of the study to avoid confusion.

II.4. Illustration on a real dataset: tree species of a Miombo woodland

The three methods of selection were also applied to the Mikembo forest data (see Muledi et al. 2017). This dataset corresponds to an exhaustive census and mapping of all individual trees (≥ 10 cm diameter at breast height; 4604 individuals) in a 10-ha tropical dry woodland (500 × 200 m) located in the eastern part of Upper Katanga (Democratic Republic of the Congo). The plot was divided into a grid of 160 quadrats of 25 × 25 m in which 36 soil parameters were measured (texture, soil depth, soil chemistry, etc; see Muledi et al. 2017 for further information on the dataset). The MEM variables were generated and used in a univariate context for the 24 most abundant tree species of the forest (using db-MEM based on the PCNM criterion as in the simulation study), and the selection was performed using the three selection methods (Fig. 1). We compared the proportion of species displaying a significant spatial structure and the corresponding R^2 for the subset of MEM variables selected by each method. Additionally, a variation partitioning analysis (VP; Borcard et al. 1992, Peres-Neto and Legendre 2010) was applied separately for each species displaying a significant spatial structure to assess the pure and shared effects of soil and spatial variables on species distributions. The VP were performed with each MEM subset to illustrate to what extent ecological conclusions can be affected by the different procedures of selection. We focused on fractions corresponding to the effect of a spatially structured environment ($R^2_{ENV\text{-}SPA}$) and the pure spatial fraction ($R^2_{SPACE.PURE}$). A shared effect of soil and spatial variables ($R^2_{ENV\text{-}SPA}$) may indicate an *induced spatial dependence* (i.e., the spatial signature of an influent environmental factor; Legendre and Legendre 2012), while pure spatial effects ($R^2_{SPACE.PURE}$) may relate to dispersal limitations, biotic interactions, or unmeasured environmental parameters. For each species, we computed $\Delta R^2_{\text{AIC-FWD}}$ ($R^2_{\text{MEM.AIC}} - R^2_{\text{MEM.FWD}}$), $\Delta R^2_{\text{AIC-MIR}}$ ($R^2_{\text{MEM.AIC}} - R^2_{\text{MEM.MIR}}$), and $\Delta R^2_{\text{FWD-MIR}}$ ($R^2_{\text{MEM.FWD}} - R^2_{\text{MEM.MIR}}$), for both $R^2_{ENV\text{-}SPA}$ and $R^2_{SPACE.PURE}$. These indices allowed comparing how the selection procedures influenced the estimations of the spatial fractions in VP.

All the analyses were performed in the R environment (v. 3.3.2., R Development Core Team 2014) using the packages *vegan* (Oksanen et al. 2017), *spdep* (Bivand 2006), and *adespatial*. The R code used to run the simulations is provided in Appendix A2. The R function *MEM.moransel* applying the MIR approach on a model that only contains an intercept term is provided in Appendix A3.

III. Results

III.1. Review of the literature

We analysed 301 articles published between 2006 and 2016 using MEM, PCNM, and SEVM (and PVR to a lesser extent; Appendix A4 for the list of references, and Supplementary Table A1 for the methodological details). The AIC approach was used alone in 15% of the studies, and was probably used in an additional 13% of articles lacking accurate information regarding the MEM selection procedure performed (e.g., citing Dray et al. (2006) for the MEM variable selection methodology without explicitly specifying whether the AIC approach was used, or referring to Dray et al. 2006 for further details; 'AIC' in Fig. 2). The AIC approach was also used in combination with the FWD approach in 4% of the studies. The latter generally used the AIC procedure for the selection of a spatial weighting matrix **W** and the FWD approach to select MEM variables within the chosen **W** matrix (another 2% likely did the same but lacked methodological specification; 'AIC+FWD' in Fig. 2a). Therefore, up to 32% of the reviewed studies used the AIC procedure. The FWD approach ('FWD' in Fig. 2) was used alone in 21.5% of the studies. Among all the studies applying this latter method (27.2%; i.e., alone or combined with another approach), few of them explicitly mentioned whether a global test of significance had been conducted. The classic forward selection (i.e., no global test of significance performed) was used in 5.6% of the studies. The minimisation of Moran's I in the residuals was used in 24.5% of the reviewed studies ('MIR' in Fig. 2). Among other approaches, 4.0%

of the studies selected the MEM variables displaying a significant Moran's I ('Signif. Moran's *I*' in Fig. 2a). Also, 5.3%, 4,6 %, and 1.6% of the studies used spatial predictors significantly related to the response variable(s), displaying a Moran's *I* superior to a threshold value (generally 0.1), and associated with an R^2 superior to a threshold value, respectively ('Others' for the three last categories; Fig. 2a). The use of these methods was not exclusive; thus, they represented 9.6% of the reviewed works. All MEM associated with positive eigenvalues were used without variable selection in 10.0% of the studies ('No selection' in Fig. 2).

As illustrated in Fig. 2b, while the number of studies using eigenvector-based methods clearly increased over the years, we observed no clear tendency of an increasing or decreasing use of one method at the expense of the others among the three most-used methods. Finally, only 62% of the studies properly characterised the chosen **W** matrix, and 57% of them used the original PCNM method. Among the latter, 43% claimed they used db-MEM while they used the original PCNM.

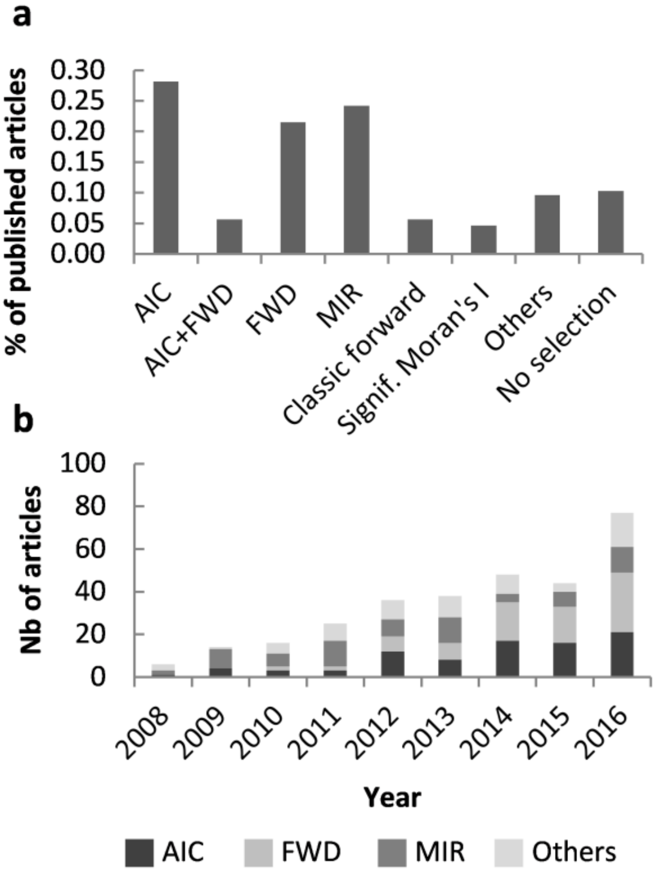


Figure 2: Review of the eigenvector selection methods used from 2006 to 2016 (number of studies: 301). a: Percentage of the articles reviewed using MEM.AIC (AIC), MEM.FWD (FWD), a combination of the two former (FWD+AIC), the MIR approach (MIR), the classic forward selection (without a global test and with the *p*-value as only stopping criterion), all MEM variables displaying a significant Moran's *I* (Signif. Moran's *I*), all the positively autocorrelated MEM variables (No selection), or one of the following criteria: all MEM significantly related to y, or displaying a Moran's *I* or a R^2 superior to a threshold value (Others). b: Number of articles that used MEM.AIC, MEM.FWD, MEM.MIR, or one of the remaining methods from 2008 to 2016 (only few studies were recorded before 2008, and are therefore not represented here).

III.2. Simulation study: type I error rate

The simulations performed using the four distribution types provided very similar results, so that we reported only the results for the uniform distribution (Supplementary Table A2 for the complete results).

Figures 3a illustrates the type I error rate of the MEM variables selected by the three selection approaches. While the FWD and MIR methods presented correct type I error rates (~0.05), the AIC procedure detected a significant spatial signal in the response **y** in 100% of the simulations.

III.3. Simulation study: power and accuracy

Figure 3b shows the power of the selection methods and the mean ΔR^2 (simulated value – real value) obtained after 10000 simulations for broad, medium, and fine spatial scales. The AIC approach always presented a power of 1, regardless of the spatial scale or type of sampling design, and systematically produced overestimated R^2, except at fine spatial scales for the regular sampling design (strong underestimation of the actual R^2 value). The FWD procedure always presented a power of ~1 and allowed retrieving a very accurate estimation of the real values of R^2 regardless of the spatial scale or type of sampling design, except for the fine-scale structure for the regular design. In the latter case, the power was very low (0.04), and R^2 was strongly underestimated. Finally, the MIR approach had a power of 1 at broad and medium scales, regardless of the type of sampling design. The method moderately underestimated the actual R^2 when using the regular design, but provided accurate estimates with the random design. However, the power of this approach to detect fine-scale spatial structures was null for both sampling designs. Regarding the number of MEM, the average number of selected variables was 6 ± 2.5, 6 ± 0, and 3 ± 1, with the FWD, the AIC, and the MIR approach, respectively.

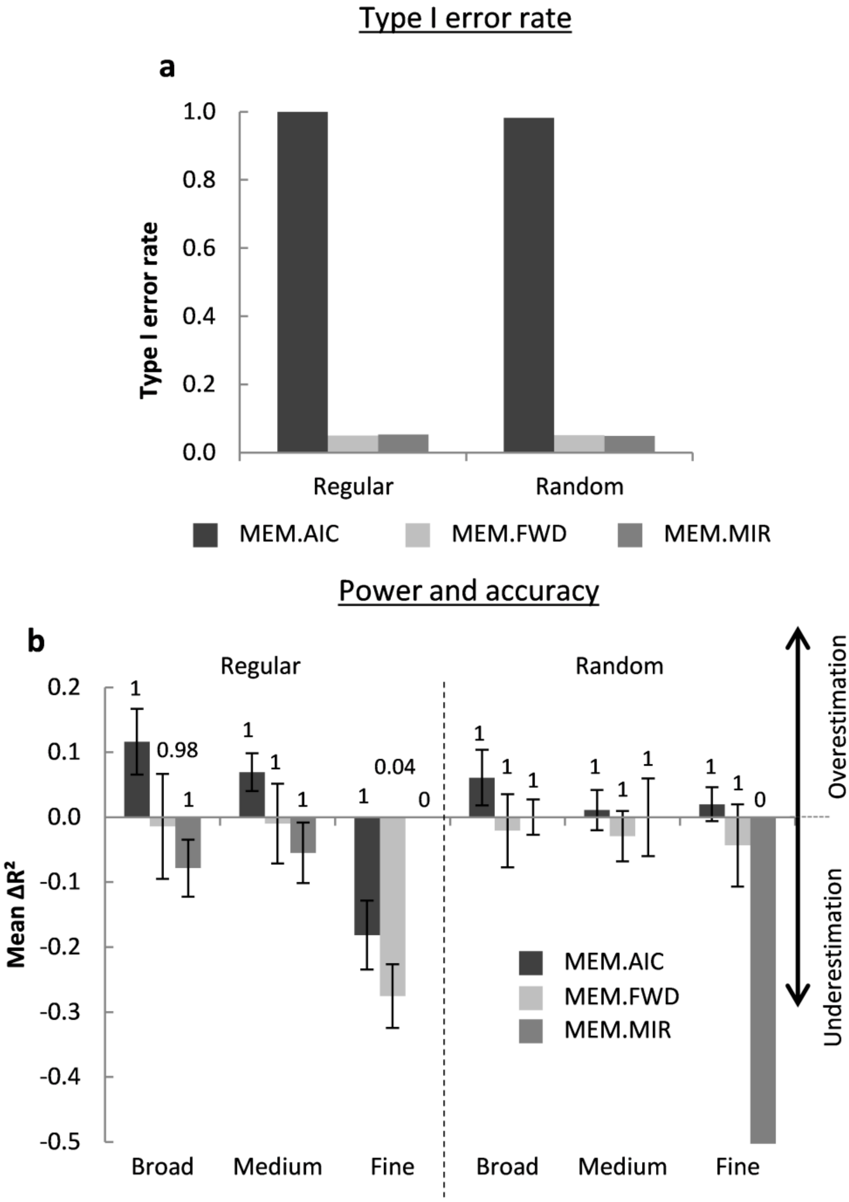


Figure 3: a: Type I error rate of the three eigenvector selection methods. b: Power and R^2 estimation accuracy of the AIC, FWD, and MIR procedures computed on 10000 simulated response variables y structured at broad, medium, or fine scale, both in a regular and in a random sampling design. The y-axis gives the mean of the ΔR^2, that is, $R^2_{\text{simulated}} - R^2_{\text{reference}}$. A positive value indicates an overestimation of the reference R^2, while a negative value corresponds to an underestimation. The absence of a bar for the MIR approach with a response vector structured at fine scale and sampled in the regular sampling design indicates that the ΔR^2 could not be computed, as the ΔR^2 is only computed on the basis of the significant simulations and the statistical power of this scenario was null. Vertical bars represent standard deviations. The number above each standard deviation bar corresponds to the statistical power.

III.4. Selection method effects on ecological interpretations of a real dataset

The three selection methods were used to select a subset of spatial predictors for the 24 most abundant tree species of Mikembo ($20 \leq n \leq 1239$). We observed significant spatial structures for around 50% of the species using the FWD and MIR approaches, while the AIC approach detected highly significant structures for all species. The FWD selection detected a weak spatial pattern for one species that the

MIR approach did not detect, and the MIR approach also detected weak patterns for three species that the FWD did not detect. The main results (Fig. 4) therefore focus on the 11 species that displayed spatial patterns according to both the FWD and MIR methods. The R^2 obtained with MEM.AIC was systematically higher than that obtained with MEM.FWD (up to 23% more) and MEM.MIR (up to 33% more, Fig. 4a) for all species. Regarding VP, as illustrated in Figs. 4b and 4c, most of the R^2 overestimation of the AIC approach ended up in the pure spatial fraction of the VP (ΔR^2 up to 0.19) and, to a minor extent, in the structured environmental fraction (ΔR^2 up to 0.07). The results of $\Delta R^2_{\text{AIC-MIR}}$ were very similar to those of $\Delta R^2_{\text{AIC-FWD}}$ (Fig. 4b) and are therefore not presented here.

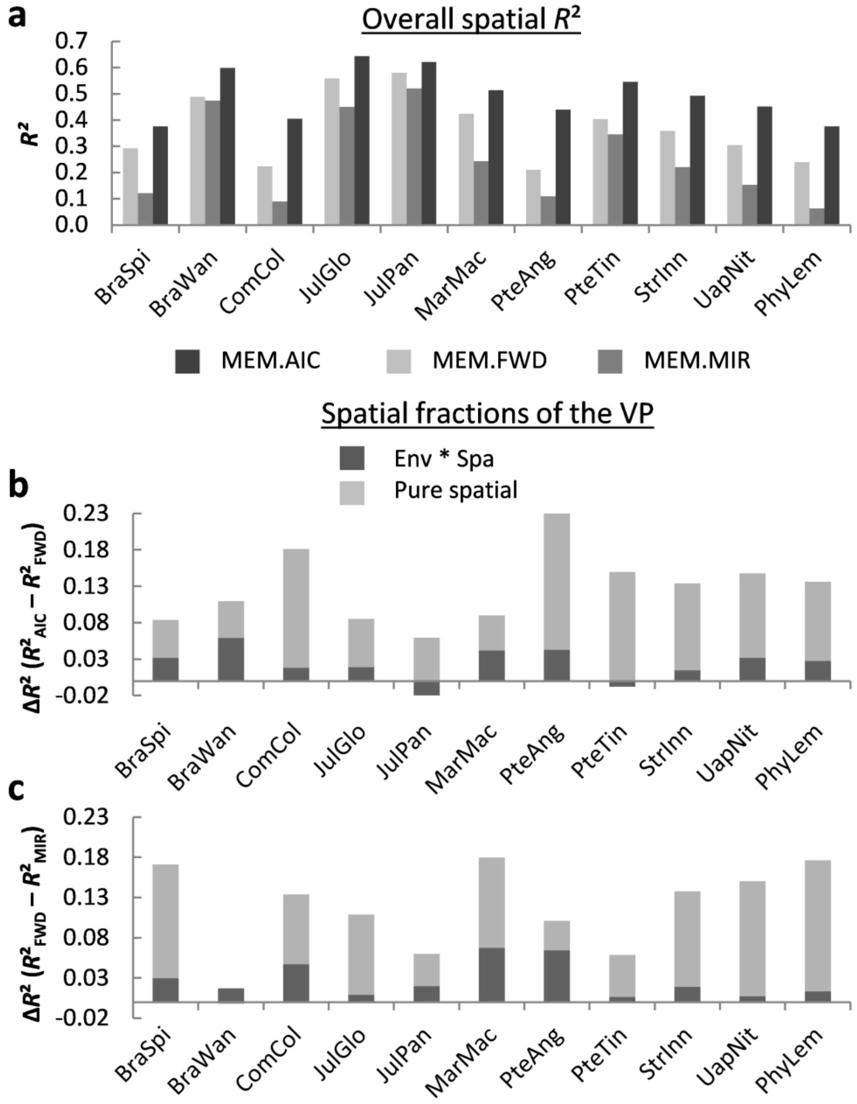


Figure 4: Illustration of the different selection methods on a real dataset. a: Spatial R^2 obtained using MEM.FWD, MEM.AIC, and MEM.MIR for all Mikembo tree species displaying a significant spatial structure according to both the FWD and the MIR procedures (i.e., the methods with a correct type I error rate, see Fig. 3a). b and c: Proportions of the R^2 overestimation ($\Delta R^2_{\text{AIC-FWD}}$ and $\Delta R^2_{\text{FWD-MIR}}$, respectively) ending up in the pure spatial and in the shared environment-space fractions of the VP ('Pure spatial' and 'Env * Spa', respectively). The ΔR^2 were obtained by computing the difference between the R^2 values of MEM.AIC and MEM.FWD (b), and MEM.FWD and MEM.MIR (c). The x-axis corresponds to the tree species abbreviations (details in Muledi et al. 2017) of the species displaying a significant spatial structure according to both the FWD and MIR procedures.

IV. Discussion

IV.1. AIC approach: a highly biased selection method

The simulations revealed that the AIC approach always detected spatial structures when there actually were none and that it always overestimated the actual proportion of spatially structured variability in **y**. The inferential outcomes of these results were illustrated on the real data, showing that spatial structures were wrongly detected for half of the species, while all spatial R^2 were overestimated when using the AIC selection (up to 23% with respect to the most accurate method, the FWD approach, Fig. 3b). This resulted in spurious ecological interpretations regarding the nature of the processes driving the spatial distribution of organisms (mainly an overestimation of community-intrinsic processes or unmeasured environmental parameters, Figs. 4b and 4c). These results show that the AIC approach systematically selects the MEM variables that best fit **y** among a high number of orthogonal explanatory variables, regardless of the global model significance, just as the classical forward selection does (Westfall et al. 1998, Blanchet et al. 2008). This explains why this method displays such a huge type I error rate. Moreover, while the FWD approach controls for R^2 inflation using the R^2 of the global model as second stopping criterion, the AIC approach does not apply any kind of equivalent control, which may partly explain why it always explained more than it should have in the simulations (Fig. 3b). We applied a global test of significance and added the adjusted R^2 stopping criterion to the AIC approach before using it on the simulated structured **y**. The results were nearly identical to those obtained with the FWD approach (Supplementary Table A3), hence indicating that the origin of the biases is not the AIC *per se*, but the absence of a global test of significance and a criterion controlling for model overfitting (such as the global R^2). However, while the classic (univariate) AIC is computed on the basis of a parametric model, Godinez-Dominguez and Freire (2003) proposed a rather simplistic way to compute AIC for the redundancy analysis (RDA; multivariate context). They make a simple but wrong analogy between the univariate residual sum of squares (RSS) and its multivariate analogue, replacing the RSS by the RSSRDA in the classical definition of AIC for linear models. However, as they do not clearly assume any distributional assumptions (e.g., multivariate normal as Pech and Laloë 1997) in the RDA model, their definition of AIC is probably wrong and not applicable. We therefore advocate the use of the R^2_{adj} as a criterion of selection when using the FWD approach, following Blanchet et al. (2008). The AIC approach should be abandoned for future works involving spatial, temporal, or phylogenetic predictors, whether for selecting a subset of predictors (biases highlighted by our results), and probably also for choosing a best-suited **W** matrix (lack of statistical ground for multivariate response data). In addition, the optimisation of the choice of a **W** matrix is likely to be biased (inflated type I error rate) if not associated to a *p*-value correction for multiple testing, as the chance of detecting a significant **W** matrix by chance is expected to increase with the number of **W** matrices compared.

IV.2. Fine-scale positively autocorrelated patterns: limited power and underestimated R²

Our results confirmed a limited power of MEM at fine spatial scales (Layeghifard et al. 2015), and highlighted a systematic underestimation of the actual R^2. A combined approach using MEM for detecting broad and medium-scale patterns and spatial point pattern analysis (i.e., an individual-based approach; see Velázquez et al. 2016 for a review) for fine-scale patterns may be a good strategy to avoid the limited power issue at fine scales for positively autocorrelated structures. The advantage of exploring spatial data through these different methods is that, while MEM is most powerful at broad and medium scales, the spatial point pattern analysis successfully highlights patterns and processes occurring at very local scales, and could therefore compensate the lack of power of MEM. Nevertheless, this would require knowing the spatial location of all individuals, a condition sometimes difficult to meet. In addition, further work is still needed to test and explore the potential of combining individual-based and site-based methods (such as spatial point pattern analysis and MEM) to detect the spatial structures of species and communities and interpret them in terms of ecological processes.

In this study, we focused on positively autocorrelated spatial structures, and the fine scale patterns were therefore positively autocorrelated and generated with the MEM variables associated to small positive eigenvalues. However, although much less investigated by ecologists, fine scale patterns can also be negatively autocorrelated. To explore this point, we also simulated response variables using the last three MEM variables that are the most negatively autocorrelated (smallest eigenvalues) and tested the statistical power and accuracy of the different eigenvector selection approaches as for the other simulation scenarios. These supplementary analyses revealed that the power was close to 1 and that the R^2 estimation accuracy was good for both the FWD and the MIR approaches (Supplementary Table A4). Hence, the statistical power of MEM actually decreases with the absolute value of the eigenvalue associated to spatial predictors, indicating that the relation between statistical power and spatial scale is not straightforward. The relation between power and eigenvalue also explains why the power of the FWD approach was much higher for the fine scale patterns on a random sampling design than on a regular sampling design. Indeed, there are less positive MEM variables when the design is irregular and the last eigenvectors with positive eigenvalues therefore have a higher eigenvalue than the last positively autocorrelated eigenvectors of a regular sampling design. Further works will be necessary to better investigate the relation between the eigenvalues of the spatial eigenvectors, the scales of the corresponding patterns, and the statistical performances of MEM to detect these patterns.

It is also worth mentioning that, although the ecological example chosen to illustrate the simulated results concerned the selection of spatial predictors at a small extent (500 × 200 m), the results would likely be the same for temporal or phylogenetic eigenvectors or if the study were conducted at a macroecological extent. Indeed, as it was shown by the simulations, the differences among the selection approaches arose from the control of the type I error rate, and from the number of selected spatial predictors that either aimed at minimising the spatial autocorrelation (SAC) in the residuals with a minimum number of variables (MIR approach) or at describing the SAC as accurately as possible with a slightly higher number of variables (FWD approach). Hence, the spatial, temporal, or phylogenetic nature of the autocorrelation does not matter; what does matter is whether there is autocorrelation, at what scale (i.e., at what focus, sensu Scheiner 2011), and the characteristics of the sampling design. For instance, an irregular spatial sampling will cause variations in inter-site distances, an irregular time-series will induce differences in sampling time intervals, and a phylogenetic tree with varying branch lengths will highlight that some species are separated from the rest of the species by a much longer evolutionary time. These irregularities will affect the average inter-site distance of the study in the same way, and will therefore influence the ability to detect a pattern (Legendre and Legendre 2012), be it spatial, temporal, or phylogenetic. However, this irregularity should be considered with respect to the geographical extent of the study, so that only the relative position and distance among sites matter and not the extent per se (e.g., the same results could be obtained for a 500 × 200 m or 1000 × 400 km study area if the relative sampling designs have similar characteristics with respect to the extent of the study). These considerations about the extent of the sampling design and the nature of the autocorrelation also hold for the low statistical power that we emphasised at fine scales; the pattern can be spatial, but also temporal or phylogenetic, and the notion of 'fine' is relative to the sampling design or phylogeny used (it was of 25-50 m^2 in our example, but could be of several hundreds of square kilometres in a metacommunity analysis).

The low and null power obtained for fine scale positively autocorrelated patterns using the FWD and MIR approaches, respectively, indicate that ecologists must bear in mind that their pattern detection may be biased towards broad and medium scales. This difference of power across scales is likely to bias the interpretations of many studies. In evolutionary biology for instance, the general tendency of the phylogenetic signals detected in traits, using PVR, is therefore expected to indicate that most traits diverged a long time ago (broad and medium-scale phylogenetic structures), while the traits displaying no signal could have diverged more recently (undetected fine-scale patterns). Consequently, the specific (S) and phylogenetic (P) components of the trait variability (T = S + P, see Diniz-Filho et al. 2012) will often be artificially over-, or underestimated, respectively. Diniz-Filho et al. (2012) reviewed several selection methods in the scope of phylogenetic autocorrelation control and highlighted that the selection method of Griffith and Peres-Neto (2006; MIR approach) was the best performing approach to control for phylogenetic autocorrelation in residuals but that using a small number of eigenvectors may be insufficient to correctly describe the phylogenetic autocorrelation in some cases, mainly for fine-scale complex structures. Our results indeed indicate that, although it successfully removed the autocorrelation from the model, the MIR approach is likely to perform poorly at fine phylogenetic scales. This also questions the suitability of the Moran's I to detect fine-scale spatial structures. The FWD approach would be an alternative providing higher statistical power and a better R^2 estimation accuracy at fine scales. Moreover, it usually selects a slightly higher number of eigenvectors than the MIR approach, therefore ensuring a better characterisation of the phylogenetic signal (Diniz-Filho et al. 2012).

IV.3. Summarising ten years of eigenvector selection practices

Blanchet et al.'s (2008) forward selection appeared to be the selection method yielding both the highest power and R^2 estimation accuracy while displaying correct type I error rates. However, the review revealed that most studies using this method did not specify whether a global test of significance had been conducted, hence possibly obtaining and interpreting spurious results.

Additionally, up to 32% of the studies reviewed used the AIC approach. The number of studies using this method has not decreased over the years (Fig. 2b), hence suggesting more biased studies to come. The review also highlighted a widespread lack of methodological specification regarding the definition of the **W** matrix used, and whether a variable selection was used or not.
Diniz-Filho et al. (2012) showed that selecting all eigenvectors significantly related to the response variable produced over-fitted models, and yet this approach was used in 5.3% of the studies considered here. Similarly, the classical forward selection was used in 5.6% of the reviewed studies in spite of its high type I error rate and model overfitting issues (Blanchet et al. 2008). All together, these 'bad practices' of selection concerned 38.9% of the published works reviewed in this study. Many studies may therefore have been subject to inflated type I error rates and potentially wrongly detected spatial structures, making the reliability of their results uncertain both quantitatively and in terms of ecological interpretations. These methodological biases are likely to have caused false positives and/or overfitted signals (mostly in the pure spatial fraction of VP), and these may impact meta-analyses (e.g., Soininen 2016, for a meta-analysis of the degree of spatial structure across different types of organisms and ecosystems). For future reliable meta-analyses and comparisons to be possible, unbiased methods of eigenvector selection well suited for the biological questions of interest will have to be adopted.

IV.4. Recommendations: eigenvector selection and objective of the study

The results showed that both the FWD and the MIR methods were well adapted to select spatial predictors. The FWD approach returned more accurate R^2 estimates and had more power at fine scales than the MIR approach, but it also selected a slightly higher number of predictors. As both methods displayed a low type I error rate, the choice to use one or the other variable selection method will mostly depend on the purpose of the study and on the univariate or multivariate nature of the model. For a univariate response variable, if the only purpose is to control the spatial autocorrelation in the model residuals, the priority is to introduce as few predictors as possible in the model to avoid losing power; thus, the MIR approach should be preferred. If the response is multivariate, or if it is univariate but the purpose is to model as accurately as possible the multiscale spatial, temporal, or phylogenetic structures, then priority should be given to precision, and we would advocate Blanchet et al.'s (2008) forward selection. A tutorial is provided in Appendix A5 to help users select and code the most suitable selection method in R depending on their purpose and type of data. Additionally, the FWD approach may be a promising solution to the issue of phylogenetic eigenvector selection in PVR (Rohlf 2001, Freckleton et al. 2011, Diniz-Filho et al. 2012) for multivariate response data and when the purpose is to describe trait phylogenetic patterns precisely.

In this study, we disentangle unbiased and accurate from biased and underpowered selection methods by considering the relations between a response variable and the spatial predictors. Although we considered an additional set of environmental predictors in the real dataset, we did not investigate with simulations how the selection methods influenced the type I error rates and estimate accuracy of the VP fractions. When considering VP, previous studies showed that variable selection procedures could lead to inflated type I error rates of the total and pure environmental fractions (Peres-Neto and Legendre 2010, Smith and Lundholm 2010). This can be explained by the fact that ‘space’ is not as accurately described with a subset of spatial predictors as with the complete set of positively autocorrelated predictors, resulting in an incomplete modelling of the spatial signal of the environmental component. However, a selection procedure is essential, as the power to detect the pure environmental fraction can strongly decrease when all MEM variables are used (Peres-Neto and Legendre 2010), and using all MEM variables overcorrects the spatial autocorrelation in the data (Griffith 2003). A solution to this problem would be to 1) apply a forward selection of the spatial predictors (Blanchet et al. 2008), and 2) introduce them in a VP in which the fractions are tested and estimated on the basis of an R^2 corrected by a null model constrained with Moran spectral randomisation (Wagner and Dray 2015) to maintain the spatial structures of the data (see Clappe et al. under review, for details).

The present study focused on the selection of an optimal subset of MEM within a given **W** matrix. However, **W** matrices can be built in many ways and can lead to contrasting results for irregular and clustered sampling designs. The selection of a **W** matrix is therefore a key step of all eigenvector-based methods (Dray et al. 2006), and yet this step appeared in our review as seldom explicitly tackled. Hence, a great proportion of studies are likely to have obtained underpowered results, as the authors used the original PCNM, only used distance-based MEM (while the latter may be an unsuitable choice for irregular or clustered sampling designs, Dray et al. 2006), did not compare different **W** matrices, and, most often did not even specify the **W** matrix that was used. The next step will therefore be to set up an accurate and unbiased optimisation procedure allowing users to test diverse types of **W** matrices and select the most adapted to their dataset.

Acknowledgements

This research was supported by the Belgian National Fund for Scientific Research (F.R.S.-FNRS) to DB. The computational resources have been provided by the Shared ICT Services Centre, Université Libre de Bruxelles. We are grateful to three anonymous reviewers and to the anonymous subject editor for their constructive suggestions and discussions.

Supporting information

See *Annexes* at the end of the thesis.

References

Anderson, M. J. and Legendre, P. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. - J. Stat. Comput. Simul. 62: 271–303.

Baho, D. L. et al. 2015. Assessing temporal scales and patterns in time series: Comparing methods based on redundancy analysis. - Ecol. Complex. 22: 162–168.

Barbier, N. et al. 2008. Spatial decoupling of facilitation and competition at the origin of gapped vegetation patterns. - Ecology 89: 1521–1531.

Bauman, D. et al. 2016. Multiscale assemblage of an ectomycorrhizal fungal community: the influence of host functional traits and soil properties in a 10-ha miombo forest. - FEMS Microbiol. Ecol. 92: fiw151.

Bini, L. M. et al. 2009. Coefficient shifts in geographical ecology: an empirical evaluation of spatial and non-spatial regression. - Ecography 32: 193–204.

Bivand, R. 2006. spdep: spatial dependence: weighting schemes, statistics and models. R package (version 0.6-13).

Blanchet, F. G. et al. 2008. Forward selection of explanatory variables. - Ecology 89: 2623–2632.

Borcard, D. and Legendre, P. 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. - Ecol. Modell. 153: 51–68.

Borcard, D. et al. 1992. Partialling out the spatial component of ecological variation. - Ecology 73: 1045–1055.

Borcard, D. et al. 2011. Numerical Ecology with R. - Springer.

Clappe, S. et al. 2018. Beyond neutrality: disentangling the effects of species sorting and spurious correlations in metacommunity analysis. - Ecology (under review)

Corkeron, P. J. et al. 2011. Spatial models of sparse data to inform cetacean conservation planning: an example from Oman. - Endanger. Species Res. 15: 39–52.

Cormack, R. M. and Ord, J. K. 1979. Spatial and temporal analysis in ecology. - Int. Congr. Ecol. 1978

Declerck, S. A. J. et al. 2011. Scale dependency of processes structuring metacommunities of cladocerans in temporary pools of High-Andes wetlands. - Ecography 34: 296–305.

Diniz-Filho, J. A. F. and Bini, L. M. 2005. Modelling geographical patterns in species richness using eigenvector-based spatial filters. - Glob. Ecol. Biogeogr. 14: 177–185.

Diniz-Filho, J. A. F. et al. 1998. An eigenvector method for estimating phylogenetic inertia. - Evolution 52: 1247–1262.

Diniz-Filho, J. A. F. et al. 2012. On the selection of phylogenetic eigenvectors for ecological analyses. - Ecography 35: 239–249.

Dormann, C. F. et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. - Ecography 30: 609–628.

Dray, S. et al. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). - Ecol. Modell. 196: 483–493.

Ezekiel, M. 1929. The application of the theory of error to multiple and curvilinear correlation. - J. Am. Stat. Assoc. 24: 99–104.

Freckleton, R. P. et al. 2011. Comparative methods as a statistical fix: The dangers of ignoring an evolutionary model. - Am. Nat. 178: E10–E17.

Godinez-Dominguez, E. and Freire, J. 2003. Information-theoretic approach for selection of spatial and temporal models of community organization. - Mar. Ecol. Prog. Ser. 253: 17–24.

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

Griffith, D. 2003. Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization. - Springer.

Griffith, D. A. and Peres-Neto, P. R. 2006. Spatial modeling in Ecology: the flexibility of eigenfunction spatial analyses. - Ecology 87: 2603–2613.

Hurvich, C. M. and Tsai, C.-L. 1989. Regression and time series model selection in small samples. - Biometrika 76: 297–307.

Jombart, T. et al. 2009. Finding essential scales of spatial variation in ecological data: A multivariate approach. - Ecography 32: 161–168.

Layeghifard, M. et al. 2015. Spatial and species compositional networks for inferring connectivity patterns in ecological communities. - Glob. Ecol. Biogeogr. 24: 718–727.

Legendre, P. 1993. Spatial autocorrelation: Trouble or new paradigm? - Ecology 74: 1659–1673.

Legendre, P. and Fortin, M. J. 1989. Spatial pattern and ecological analysis. - Vegetatio 80: 107–138.

Legendre, P. and Legendre, L. 2012. Numerical Ecology. - Elsevier.

Manly, B. F. J. 2007. Randomization, bootstrap and Monte Carlo methods in biology. - Chapman & Hall/ CRC.

McIntire, E. J. B. and Fajardo, A. 2009. Beyond description: the active and effective way to infer processes from spatial patterns. - Ecology 90: 46–56.

Muledi, J. I. et al. 2017. Fine-scale habitats influence tree species assemblage in a miombo forest. - J. Plant Ecol. 10: 958–969.

Oksanen, J. et al. 2017. Package "vegan": Community ecology package (version 2.4-3).

Pech, N. and Laloë, F. 1997. Use of principal component analysis with instrumental variables (PCAIV) to analyse fisheries catch data. - ICES J. Mar. Sci. 54: 32–47.

Peres-Neto, P. R. and Legendre, P. 2010. Estimating and controlling for spatial structure in the study of ecological communities. - Glob. Ecol. Biogeogr. 19: 174–184.

Pinto, S. M. and MacDougall, A. S. 2010. Dispersal Limitation and Environmental Structure Interact to

Restrict the Occupation of Optimal Habitat. - Am. Nat. 175: 675–686.

R Development Core Team. 2014. R: A language and environment for statistical computing. - R Foundation for Statistical Computing.

Rohlf, F. J. 2001. Comparative methods for the analysis of continuous variables: geometric interpretations. - Evolution 55: 2143–2160.

Scheiner, S. M. 2011. Musings on the Acropolis: Terminology for biogeography. - Front. Biogeogr. 3: 62–70.

Siesa, M. E. et al. 2011. Spatial autocorrelation and the analysis of invasion processes from distribution data: a study with the crayfish Procambarus clarkii. - Biol. Invasions 13: 2147–2160.

Smith, T. W. and Lundholm, J. T. 2010. Variation partitioning as a tool to distinguish between niche and neutral processes. - Ecography 33: 648–655.

Soininen, J. 2016. Spatial structure in ecological communities - a quantitative analysis. - Oikos 125: 160–166.

Tiefelsdorf, M. and Griffith, D. A. 2007. Semiparametric filtering of spatial autocorrelation: the eigenvector approach. - Environ. Plan. A 39: 1193–1222.

Velázquez, E. et al. 2016. An evaluation of the state of spatial point pattern analysis in ecology. - Ecography (Cop.). 39: doi: 10.1111/ecog.01579.

Vleminckx, J. et al. 2017. The influence of spatially structured soil properties on tree community assemblages at a landscape scale in the tropical forests of southern Cameroon. - J. Ecol. 105: 354–366.

Wagner, H. H. and Dray, S. 2015. Generating spatially constrained null models for irregularly spaced data using Moran spectral randomization methods. - Methods Ecol. Evol. 6: 1169–1178.

Westfall, P. H. et al. 1998. Forward selection error control in the analysis of supersaturated designs. - Stat. Sin. 8: 101–117.

Chapitre IV

Optimizing the choice of a spatial weighting matrix in eigenvector-based methods

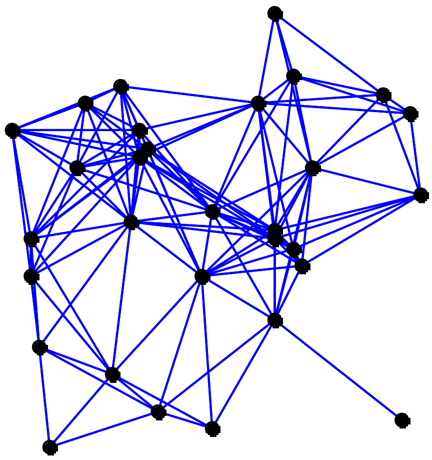
David Bauman, Thomas Drouet, Marie-Josée Fortin, and Stéphane Dray

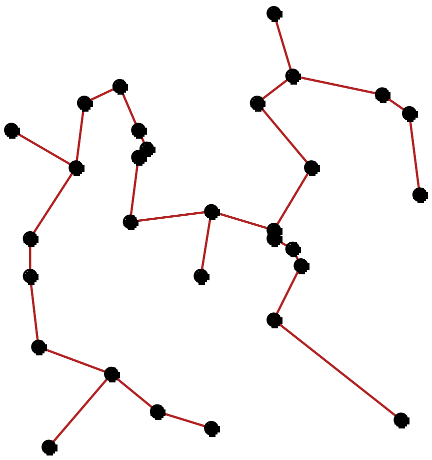
Published July 2018 in *Ecology*, doi: 10.1002/ecy.2469

Le Chapitre III a mis en évidence que près de 40 % des études publiées ont utilisé une sélection de variables spatiales MEM fortement biaisée ou pour le moins sous-optimale et imprécise. Les simulations ont permis de formuler des recommandations sur la méthode de sélection de variables spatiales à privilégier en fonction de la nature univariée ou multivariée des données réponses ainsi que de l'objectif de l'étude (éliminer l'autocorrélation spatiale des résidus d'un modèle ou modéliser et décrire aussi précisément que possible des patterns multi-échelles de structuration).

Le Chapitre IV aborde une question épineuse encore non résolue se situant en amont de la sélection du sous-ensemble de vecteurs propres : le choix de la matrice de pondération spatiale à partir de laquelle les vecteurs propres sont générés. Il a été montré à plusieurs reprises que cette étape est cruciale pour une mise en évidence précise des patterns spatiaux d'une variable ou matrice réponse. Néanmoins, aucun consensus concernant la façon de choisir cette matrice de pondération spatiale n'existe jusqu'à présent. En outre, le chapitre précédent a montré qu'une proportion élevée des études publiées ne précise pas la nature de la matrice de pondération spatiale utilisée ou que les auteurs utilisent une matrice de pondération spatiale inadaptée aux designs d'échantillonnage irréguliers.

Le Chapitre IV utilise à nouveau des simulations, mais cette fois-ci pour comparer les performances statistiques d'une large gamme de matrices de pondération spatiales. Une méthode d'optimisation de la sélection de cette matrice e st proposée et comparée aux pratiques actuellement majoritaires.





Optimizing the choice of a spatial weighting matrix in eigenvector-based methods

Abstract

Eigenvector-mapping methods such as Moran's eigenvector maps (MEM) are derived from a spatial weighting matrix (SWM) that describes the relations among a set of sampled sites. The specification of the SWM is a crucial step, but the SWM is generally chosen arbitrarily, regardless of the sampling design characteristics. Here, we compare the statistical performances of different types of SWMs (distance-based or graph-based) in contrasted realistic simulation scenarios. Then, we present an optimization method and evaluate its performances compared to the arbitrary choice of the most-widely used distance-based SWM. Results showed that the distance-based SWMs generally had lower power and accuracy than other specifications, and strongly underestimated spatial signals. The optimization method, using a correction procedure for multiple tests, had a correct type I error rate, and had higher power and accuracy than an arbitrary choice of the SWM. Nevertheless, the power decreased when too many SWMs were compared, resulting in a trade-off between the gain of accuracy and the loss of power. We advocate that future studies should optimize the choice of the SWM using a small set of appropriate candidates. R functions to implement the optimization are available in the *adespatial* package and are detailed in a tutorial.

Keywords

community ecology, community simulation, connection scheme, Moran's eigenvector maps (MEM), principal coordinates of neighbor matrices (PCNM), spatial autocorrelation, spatial eigenvector mapping (SEVM), multiscale spatial patterns, spatial weighting matrix, optimization, type I error rate inflation

I. Introduction

Spatial (but also temporal or phylogenetic) autocorrelation can be seen either as a curse or as an opportunity for ecologists (Peres-Neto and Legendre 2010, Diniz-Filho et al. 2012, Bauman et al. 2018). Indeed, the lack of independence between observations (spatial autocorrelation, SAC) causes standard statistical procedures to be too liberal (inflated type I error rate; Legendre 1993, Diniz-Filho and Bini 2005). Space, therefore, hinders the correct assessment of a relation between the response variable(s) and a set of predictors (*spatial nuisance*, Peres-Neto and Legendre 2010). Yet, space was also shown to be a surrogate of the effect of ecological processes on living communities (McIntire and Fajardo 2009, Legendre and Legendre 2012; *spatial legacy*, Peres-Neto and Legendre 2010). Hence, several spatially-explicit methods have been developed either to filter SAC from residuals or to depict multiscale spatial patterns and relate them to underlying ecological processes (Griffith 1996, 2004, Plotkin et al. 2000, Borcard and Legendre 2002, Diniz-Filho and Bini 2005, Dray et al. 2006, 2012, Wagner and Dray 2015). The advent of spatial eigenvector-based methods has brought a major advance in this field (Griffith 1996), especially with the development of Moran's eigenvector maps (MEM, Dray et al. 2006) that generalizes the ad hoc principal coordinates of neighbor matrices (PCNM; Borcard and Legendre 2002). MEM allow including multiscale spatial predictors in all kinds of univariate and multivariate models (e.g., generalized linear models, canonical analyses).

MEM variables (also further referred to as spatial predictors or spatial eigenvectors) are generated by the diagonalization of a doubly centered spatial weighting matrix (SWM) **W**. The matrix **W** is obtained as the Hadamard product (element-wise product) of a connectivity matrix (**B**) by a weighting matrix (**A**). The binary matrix **B** defines the pairs of connected and unconnected sites (binary matrix), while the matrix **A** allows weighting the connections, for instance to define that the strength of the connection between two sites decreases with the geographic distance (Dray et al. 2006). The matrix **B** can be distance-based, when the connection status (i.e., 1 or 0) of each pair of sites depends on the distance between them with respect to a connection threshold distance (e.g., Euclidean distances; *db*-MEM), as in the original PCNM method, but it can also be based on geometrical connection schemes, such as the Delaunay triangulation, Gabriel's graph, relative neighborhood graph, or a minimum spanning tree (graph-based MEM, hereafter *gb*-MEM; Dray et al. 2006, Legendre and Legendre 2012). Connections can also be built on the basis of landscape features (physical barriers, resistance to movement; Taylor et al. 1993, Fortin and Payette 2002, Spear et al. 2010).
Several works have investigated how different specifications of SWMs influence the results of spatial analyses (e.g., Stetzer 1982, Florax and Rey 1995, Kostov 2010, Griffith 2017a). The choice of the SWM has been shown to greatly influence the accuracy of parameter estimations and the spatial patterns detected in different types of space-time forecasting models, such as Lagrange Multiplier tests, in econometrics (Stakhovych and Bijmolt 2008), spatial autoregressive models (Griffith and Lagona 1998), and in spatial eigenvector-based methods too, especially for irregular sampling designs (Dray et al. 2006, Patuelli et al. 2011, Griffith 2017). However, a thorough evaluation (in terms of type I error rate, power, and accuracy) is still lacking to understand how spatial eigenvector-based methods are affected by the specification of the SWM with respect to the type of sampling design, the strength of the SAC, and the scale of the pattern. A recent review revealed that most studies used a single – and seemingly arbitrarily chosen – SWM (Bauman et al. 2018). Bauman et al. (2018) also highlighted that only 58% of the studies describe clearly the specification of the SWM, and that over 60% of these studies used either *db*-MEM or the original PCNM approach without justification, even if the latter lacks mathematical formalism, is very sensitive to irregular sampling designs, and present a lower statistical power than its MEM counterpart (Dray et al. 2006). Although Dray et al. (2006) proposed a procedure to select an optimal SWM among a set of candidates, this approach was based on the computation of an Akaike information criterion extended to the case of multivariate response data (Godinez-Dominguez and Freire 2003) which suffers from poor theoretical bases and does not test the candidate matrices against a null model. Hence, this procedure returns an optimal SWM even if there is no spatial structure in the response data, so that its use has been discouraged (Bauman et al. 2018).

Here, we address the issue of the selection of the SWM. We first compare contrasted types of SWMs in terms of type I error rate, statistical power, and R^2 estimation accuracy for: (1) random and clustered sampling designs, (2) weak and strong degrees of SAC, and (3) broad and fine spatial scale patterns. Then, we propose a new procedure that optimizes the selection of the SWM. Finally, we evaluate the performances of this new procedure and compare it to the most common current practices.

II. Material and methods

II.1. Comparing the performance of spatial weighting matrices

We defined a 90 × 90 grid and sampled 120 cells (sites) following a clustered (three clusters of 40 sites) or a random sampling design (right portion of Fig. 1 for an illustration and Appendix S2: Section 1 for methodological details). For each sampling type, we built 21 contrasting SWMs as a combination of connectivity and weighting matrices (see below) and compared their performance (i.e. type I error rate, statistical power, and R^2 estimation accuracy).

Connectivity matrix (**B**): The connectivity matrix (**B**) was generated using four connection schemes (graph-based MEM, hereafter *gb*-MEM: Delaunay triangulation, *del*, Gabriel graph, *gab*, relative neighborhood graph, *rel*, and minimum spanning tree, *mst*), and one distance threshold (*db*-MEM; see Appendix S1: Fig. S1). The latter corresponded to the smallest distance that kept all sites connected (i.e., the PCNM criterion of connectivity; *db* in Appendix S1: Fig. S1). The graph-based connection schemes are inclusive, so that all the links of *mst* are included in *rel*, included in *gab*, itself included in *del*. Hence, the number of connections increases along these graphs (Legendre and Legendre 2012).

Weighting matrix (**A**): Different weighting matrices (**A**) were combined to each **B** matrix. We defined (i) a neutral weighting function (f_{bin}; i.e., no weight added to the connections), (ii) a linear function $f_{lin} = 1 - (d/d_{\max})$, (iii) a concave-down function $f_{down} = 1 - (d/d_{\max})^{\alpha}$, and (iv) a concave-up function $f_{up} = 1/d_{\max}^{\alpha}$, where d is the Euclidean distance between two sites, $d_{\max}$ is the maximum distance between two sites, and $\alpha = 5$ and 0.5 in f_{down} and f_{up}, respectively (see plot of the weighting functions in Appendix S1: Fig. S2). The weighting function $f_{PCNM} = 1\text{-}(d/4t)^2$ was used with the *db*-**B** matrix, where t is the threshold distance beyond which two sites are not connected, and 4 is an empirical value beyond which the eigenvectors remain stable (Borcard and Legendre 2002). This combination of *db* and f_{PCNM} corresponds to the PCNM criteria used in the framework of MEM (*db*-MEM$_{\text{PCNM}}$; see Dray et al. 2006).

For each SWM, we only considered the MEM variables associated to positive eigenvalues (hereafter 'positive MEM variables'), as most studies focus on contagious ecological processes (i.e., displaying positive SAC). Using MEM variables associated to negative or to all eigenvalues yielded very similar results (not shown).

II.1.1. Type I error rate

A random univariate response variable **y** was generated in the sampled cells from four distributions: uniform, normal, exponential, and cubed exponential (Anderson and Legendre 1999, Manly 2007). We then generated MEM variables using the 21 above-described types of SWM. A global test of significance of **y** against each SWM was performed separately by 999 permutations (i.e., regressing **y** against the entire set of positive MEM variables). The above-described simulation procedure was repeated 1000 times by resampling different sets of 120 cells within the 90 × 90 grid, and the type I error rate was the proportion of significant results (significance level of 0.05).

II.1.2. Statistical power and R² estimation accuracy

The SWMs were then evaluated on the basis of their statistical power and R^2 estimation accuracy in a set of scenarios where the response variable **y** was spatially structured. We considered different sampling designs (clustered or random, see previous section), degrees of SAC (low or high), and spatial scales (broad or fine) in these simulations (Fig. 1; Appendix S2: Section 3 for details).

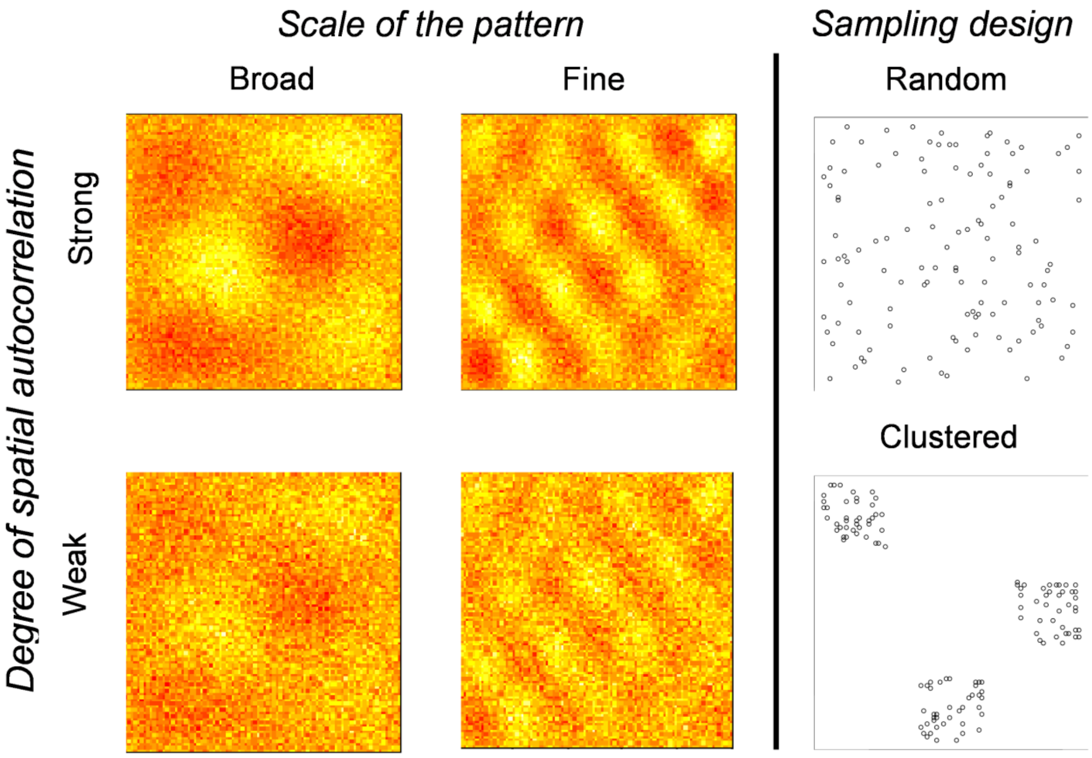


Figure 1: Schematic definition of the simulation design used for evaluating power and accuracy. The response variables were generated with spatial patterns structured either at broad or fine spatial scales, with a strong or a weak degree of SAC on a grid of 8100 cells (90 × 90). Then, 120 cells were sampled either randomly distributed, or following a clustered sampling design. The sampled values were then considered as the response variable (y) to assess the statistical performances of the different types of SWMs and of the optimization method (see Appendix S2: Section 3). The type I error rate evaluation used the same design but considered a random y.

Note that we did not consider regular sampling designs (e.g. equally spaced sites on a transect or grid) as spatial eigenvectors built with different SWMs detect roughly the same structures in this context (see discussion). The response variable (**y**) was regressed on the global set of positive MEM variables generated from the same 21 SWMs, and a forward selection with double stopping criterion (Blanchet et al. 2008) was performed separately on the significant SWMs to select a suitable subset of spatial predictors. Then, **y** was regressed on the forward selected MEM variables of each significant SWM and the spatial contribution (R^2) to the overall variability of **y** was computed for the significant SWMs. The R^2 estimation accuracy of each SWM (hereafter ΔR^2) was defined as the difference between the true R^2 value (R^2_{ref}) and the R^2 estimated by the forward selected MEM variables of a given SWM (R^2_{sim}; i.e., $\Delta R^2 = R^2_{sim} - R^2_{ref}$), so that negative and positive values indicated underestimation and overestimation of the true spatial signal, respectively. The complete simulation procedure was repeated 1000 times by resampling different sets of 120 cells in the 90 × 90 grid, and the power was computed for each SWM as the proportion of simulations returning significant global R^2. The R^2 estimation accuracy was the mean ΔR^2 of the significant simulations. The complete simulation procedure is detailed in Appendix S2: Section 3.

II.2. Optimizing the selection of the spatial weighting matrix

II.2.1. Optimization method

When a number of potential SWMs is considered, it is expected that the probability to accidentally detect a spatial signal for a given response variable will increase with the number of candidates. As a consequence, we proposed a procedure to optimize the selection of a SWM while maintaining a correct type I error rate. After defining a set of potential SWMs, our method consists in (1) performing a global test (based on the R^2 of the model considering all MEM variables as explanatory variables) on each candidate matrix with a p-value correction for multiple tests (corrected by the number of SWMs compared), (2) running a forward selection with double stopping criterion (Blanchet et al. 2008) on the significant SWMs to define the best subset of eigenvectors for each one of them, and 3) selecting the optimal SWM as the one for which the best subset of eigenvectors yields the highest adjusted R^2. In this paper, the p-value is corrected by the Šidak correction (Šidák 1967), where $P_S = 1 - (1 - P)^k$, with P_S = the corrected p-value, P = the uncorrected p-value, and k = the number of tests (i.e., the number of SWMs) but other correction methods can be considered.

The optimization method has been implemented in R functions available in the *adespatial* package (Dray et al. 2018) (details and R tutorial in Appendix S3). These functions provide also alternate optimization procedures (e.g. minimizing residual SAC instead of maximizing adjusted R^2) that can be more suitable depending on the objective of the analysis (see details in Appendix S2: Section 4 and illustration in Appendix S3).

II.2.2. Performance of the optimization method

The type I error rate, power, and accuracy of this optimization method were calculated through 1000 repetitions using the same simulation design. To assess the effect of the p-value correction, and because optimizing the selection of the SWM has so far been done without control of false discovery rate (Bauman et al. 2018), the type I error rate of the optimization method was computed with and without the p-value correction.

Five contrasting candidate SWMs were used in the optimization procedure: *gab* and *mst* (**B** matrices) associated with the f_{lin} and f_{down} functions (**A** matrices), and the *db*-MEM$_{\text{PCNM}}$. The power and accuracy of our optimization procedure were compared to those of the arbitrary choice of *db*-MEM$_{\text{PCNM}}$ (i.e., the most common current practice), and to those of the random selection of a SWM among a set of 57

SWMs (see details in Appendix S2: Section 5). This allowed assessing the benefits of optimizing the choice of the SWM with respect to a randomly-chosen SWM or the common arbitrary choice of *db*-$\mathrm{MEM}_{\mathrm{PCNM}}$.

All analyses were conducted in the R environment (version 3.4.3., R Core Team 2017) using the packages *vegan* (Oksanen et al. 2017), *spdep* (Bivand 2006), and *adespatial* (Dray et al. 2018). The R code of the study is provided in supplementary Data S1.

III. Results

III.1. Comparing the performance of spatial weighting matrices

The four random distributions yielded similar results. Hence, we only present the results of the uniform distribution (the other results are available in Appendix S4).

Figure 2a displays the type I error rate for each combination of **B** and **A** matrices, and shows that all the tested SWMs presented a correct type I error rate (between 0.04 and 0.06), regardless of the sampling design considered.
Figures 2c, and 2e show the R^2 estimation accuracy and statistical power of the SWMs tested with a strong degree of SAC, respectively. Regardless of the degree of SAC, spatial scale, or type of sampling design, the *gb*-MEM (*del*, *gab*, *rel*, and *mst*) systematically yielded a higher accuracy of R^2 estimation than the *db*-MEM (except for the strong degree of autocorrelation at broad scale for the random sampling design). Among the *db*-MEM, the PCNM and binary weighting functions were nearly always associated to the strongest model underestimations. These underestimations were maximal when $\boldsymbol{y}$ displayed a fine-scale pattern. Overall, the *db*-MEM always performed poorly compared with at least one type of the *gb*-MEM models.

High degree of SAC: With a clustered sampling design, the *gb*-MEM slightly underestimated the real R^2 (ΔR^2 down to -0.07 with mst_{bin}), while the *db*-MEM led to more severe underestimations (ΔR^2 down to -0.36 with *db*-$\mathrm{MEM}_{\mathrm{PCNM}}$; Fig. 2c). The results were very similar using the random sampling design, with a slight underestimation for the *gb*-MEM (ΔR^2 down to -0.09), except for the *del*-**B** matrix that led to strong underestimations when considering a fine-scaled pattern, regardless of the **A** matrix (mean ΔR^2 of -0.34). The *db*-MEM led again to strong R^2 underestimations (ΔR^2 down to -0.37).
All the SWMs displayed high statistical power except for the *db*-MEM at fine scale and for the *del*-**B** matrix at fine scale for a random sampling design (i.e., for the above-mentioned cases of strong R^2 underestimation; Fig. 2e).

Low degree of SAC: The results with a low degree of SAC were very similar, except for a general tendency towards a lower statistical power for all SWMs, and a more accurate R^2 estimation for both the *gb*-MEM and *db*-MEM. Yet, the *db*-MEM still underestimated the R^2, regardless of the spatial scale or sampling design considered (see details for the low degree of SAC in Appendix S1: Figs. S3a, c).

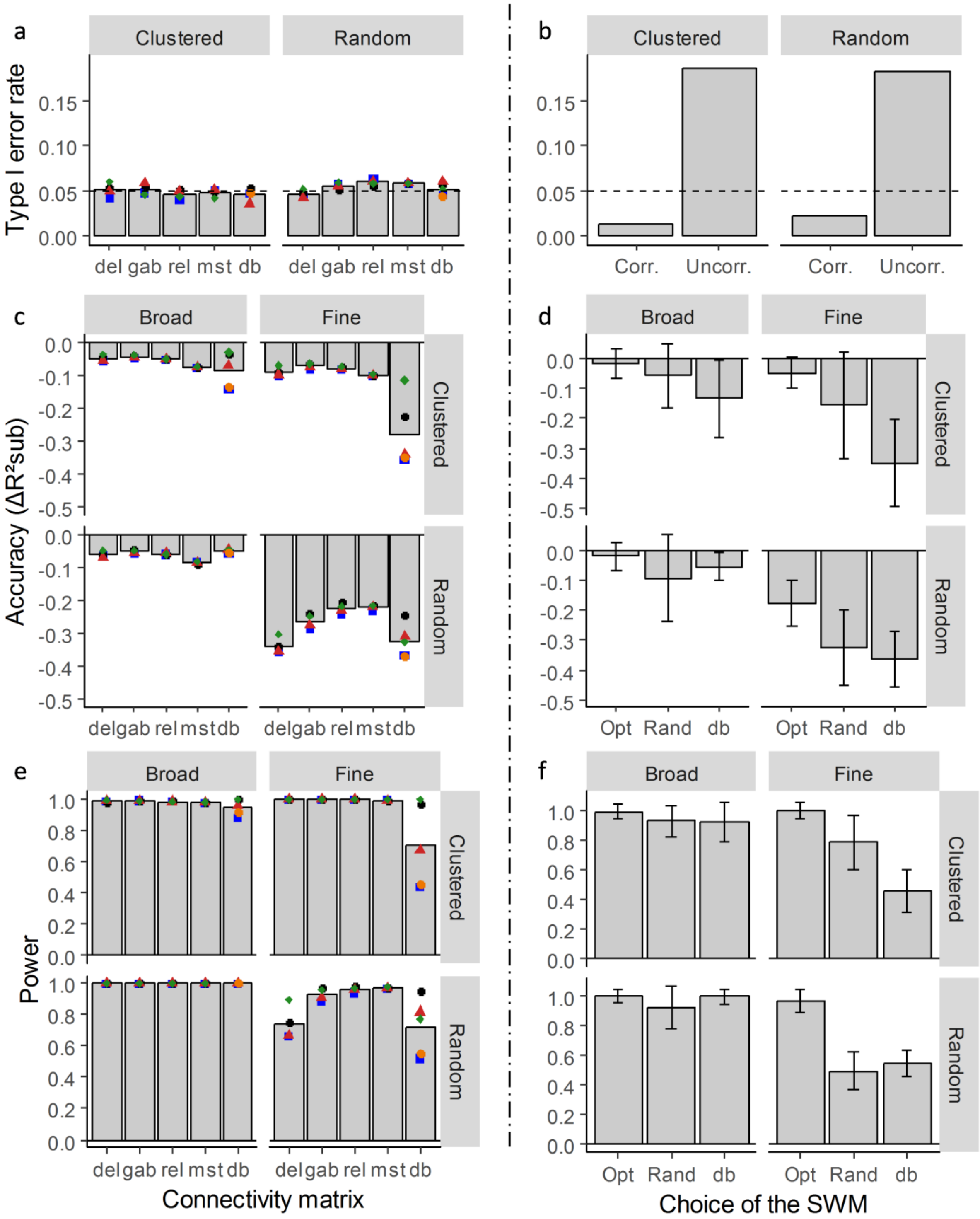


Figure 2: Type I error rate, R^2 estimation accuracy, and statistical power of the different SWMs (a, c, e), the optimization method used with a forward selection criterion, the random choice of a SWM, and the arbitrary choice of *db*-MEM$_{PCNM}$ (b, d, f), at broad and fine spatial scales and for different types of sampling design ('Clustered' and 'Random'). These are the results for the high degree of SAC (see results with a low degree of SAC in Appendix S1: Fig. S3). a, c, e: The grey vertical bars correspond to the mean of the type I error rate (a), mean ΔR^2 (i.e., $R^2_{sim} - R^2_{ref}$) (c), and statistical power (e) of the different A matrices within each B matrix (x-axis). The symbols give the detailed values for the combinations of the matrices B and A. Squares: f_{bin}, black circles: f_{lin}, triangles: f_{down}, diamonds: f_{up}, orange circles: f_{PCNM}. b: Type I error rate of the procedure with

(‘Corr.’) and without (‘Uncorr.’) the Sidak correction of the global *p*-value for multiple tests. d, f: R^2 estimation accuracy (d) and statistical power (f) of the optimization procedure with *p*-value correction (‘Opt’), the random choice of a SWM among 57 candidates (‘Rand’), and the *db*-MEM$_{PCNM}$ (‘db’). a, b: The dashed line is the correct type I error rate (i.e. 0.05). c, d: Negative and positive values of ΔR^2 correspond to underestimations and overestimations of the actual R^2 (i.e. R^2_{ref}), respectively.

III.2. Optimizing the selection of the spatial weighting matrix

Figure 2b shows the type I error rates of the optimizing method with and without *p*-value correction. Without correcting the *p*-value for multiple tests, optimizing the choice of the SWM among the five candidates tested inflated the type I error rate (0.18 for both sampling designs), while the method presented a correct type I error rate when correcting the *p*-value for the number of SWMs tested (0.01 and 0.02 for the clustered and random sampling designs, respectively). As expected, without *p*-value correction, the type I error rate inflation increased with the number of SWMs tested (results not shown).

In all simulation scenarios, the optimization method had a higher or equal power (Fig. 2f) and was more accurate (Fig. 2d) than the random choice of a SWM and the arbitrary choice of *db*-MEM (*db*-MEM$_{PCNM}$). Indeed, while the mean ΔR^2 of the optimization was always close to 0, the mean ΔR^2 of the random choice and the *db*-MEM went down to -0.33 and -0.37, respectively, hence causing severe underestimations of the spatial signal.

db-MEM performed the worst in most cases, mostly when the sampling design was clustered and for detecting fine-scaled patterns. The statistical power of this practice was particularly low for detecting fine-scaled patterns, and so was the power of the random choice of a SWM (although less markedly).

The benefit of the gain of precision of the optimization method was more marked for high degrees of SAC (see results for the low degree of SAC in Appendix S1: Figs. S3b, d), and more specifically for the fine-scaled patterns, both for clustered and random sampling designs (Fig. 2d). Moreover, at fine scales, the power of the optimization method was ~1, while the powers of the random choice and the *db*-MEM both went down to ~0.5 (Fig. 2f). Increasing the number of candidate SWMs in the optimization procedure enhanced the R^2 estimation accuracy but also decreased the statistical power, as the corrected significance threshold became more severe (i.e. smaller; results not shown).

IV. Discussion

Properly defining the SWM to be used in spatially explicit analyses of ecological data is important to avoid biases, accurately capture and study the multiscale distribution patterns of living organisms. To do so, it is crucial to evaluate the practices related to the most crucial step of these methods, that is, the selection of a SWM. Bauman et al. (2018) showed that few studies considered this issue and that around half of the published works did not describe precisely how they defined the SWM in their study. It was also highlighted that an arbitrary choice of db-MEM and most often of the original PCNM method was made in the great majority of the studies that specified their SWM. The PCNM method has, however, long been shown to lack mathematical formalism, to generate less spatial predictors (therefore displaying a lower statistical power), and to be particularly sensitive to irregular sampling designs (Dray et al. 2006). It is therefore fundamental to define good practices to wisely choose the SWM for spatial data analysis based on eigenvector-based methods.

Our results showed that the *gb*-MEM always displayed a higher accuracy in R^2 estimation and nearly always had a higher statistical power than the db-MEM (the power and accuracy of db-MEM were particularly low when weighted by the binary or PCNM functions). This result can easily be understood, as distance-based connectivity criterions connect all sites apart from a distance corresponding either to the minimum distance necessary to keep all sites connected (i.e., the largest edge of a minimum spanning tree) or to any threshold distance greater than that. With irregular and clustered sampling designs, this threshold distance is likely to connect too many sites, therefore avoiding a proper detection of fine to medium-scaled spatial patterns. This will cause distant sites, potentially belonging to different clusters of sites (from a biological or ecological standpoint), to be artificially connected, hence leading to misspecifications of the SWM and poor detection of spatial patterns. The db-MEM definition is therefore unsuitable for a proper detection of spatial patterns in a set of irregularly-distributed sites. Unlike db-MEM, gb-MEM do not present the constraint of building connectivity with respect to a distance threshold potentially defined by a single pair of distant sites. Therefore, this family of MEM yields more realistic connections, regardless of the regularity/clustering of the sampled sites, and higher R^2 estimation accuracy and statistical power in our simulation study.

The connections based on the Delaunay triangulation, however, performed poorly in different scenarios. This was most likely caused by long-distance connections known to be generated at the edges of the sampling design by the Delaunay criterion (Kenkel et al. 1989). These edge effects artificially connect distant sites, hence misspecifying the SWM and causing the observed underestimations. The Delaunay triangulation should therefore be avoided, unless used with an edge effect correction (e.g., Lane et al. 1994). A solution may be to use minimum planar graphs (MPG) (Fall et al. 2007), a generalization of Delaunay triangulation that accounts for the resistance to connect sites, hence providing least-cost paths and avoiding excessively long links as those obtained with *del* in this study.

The optimization procedure that we proposed to select the SWM achieved a higher R^2 estimation accuracy than the random choice of a SWM, hence highlighting the importance of the SWM selection. However, the optimization had a high false discovery rate when not associated to a *p*-value correction for multiple tests, which confirmed our expectation. Previous studies following Dray et al. (2006) and Borcard et al. (2011) to select a SWM therefore not only probably had an inflated type I error rate when selecting spatial eigenvectors (Bauman et al. 2018), but likely also for the choice of the best SWM. The latter inflation was caused by the lack of an adapted control for the potentially very high number of candidate SWMs tested (up to ~100 in Borcard et al. 2011). This type I error rate inflation issue also occurred in additional simulations with varying parameters values (e.g., different values used as α exponent in the concave-down or concave-up functions, or different threshold distances in db-MEM, Borcard et al. 2011). This issue can be solved by correcting the global *p*-value according to the number of SWMs tested before selecting a subset of predictors (see details of these additional simulations in Appendix S2: Section 2).

The *p*-value correction can quickly become very severe, however, as the number of tests not only increases with the B matrices compared, but also with the number of connectivity distance thresholds of db-MEM or the number of parameters within each A matrix, the latter being generally used to weight each of the tested B matrices. In our simulations, the cost of the optimization procedure was nearly inexistent in most cases. However, we only performed the optimization on the basis of five SWMs. Comparing a higher number of candidates rapidly lowered the statistical power of the procedure (results not shown). A trade-off is thus necessary between the benefit and the cost of the optimization, that is, the gain of accuracy against the loss of statistical power.

In this study, the optimization was performed using a criterion associated to the fit of spatial predictors to a response variable (adjusted R^2). This procedure is relevant for any framework focusing on capturing all the spatial patterns of y (e.g., variation partitioning, Peres-Neto and Legendre 2010). However, the questions regarding the selection of the best SWM and an optimal subset of spatial predictors are also relevant when the objective is to remove residual SAC in a model considering additional explanatory variables (e.g. environmental; see spatial eigenvector mapping, Diniz-Filho and Bini 2005). In this case, the most adapted eigenvector-selection method aims to minimize the residual SAC with a small number of spatial predictors (MIR method in Bauman et al. 2018) (Griffith and Peres-Neto 2006, Bauman et al. 2018), and we recommend performing the optimization of the SWM with the same criterion (see details in Appendix S2: Section 4 and Appendix S3). In both previous cases, the selection of the SWM is based on the selection of a subset of spatial eigenvectors (maximizing the fit or minimizing the residual SAC), as subsequent analyses will consider only this subset of predictors. However, some other methods require the complete set of spatial eigenvectors (e.g., Moran spectral randomizations, Wagner and Dray 2015, or smoothed MEM, Munoz 2009). In this case, the optimization of the SWM should be performed on the basis of the whole set of MEM variables without considering any procedure of selection of a subset of eigenvectors (details in Appendix S2: Section 4 and Appendix S3).

It is worth mentioning that optimizing the choice of the SWM when the sampling design is roughly regular is less interesting, as the MEM variables originating from different SWMs detect roughly the same patterns (Dray et al. 2006). In those cases, rook (shared edge) or queen (shared edge or vertex) neighbor definitions or db-MEM can be used, for instance (see Appendix S3). For irregular sampling designs, visualizing the different connection schemes would help identifying B matrices worth being tested and compared. In addition, considering the landscape ecology (e.g., natural barriers) should help improving the definition of the connectivity among sites in matrix B. Regarding the A matrix, plotting connectivity against distance and visualizing the curve of the different functions with several values of parameter (Appendix S1: Fig. S2) should also help choosing appropriate weighting functions and parameter values. Note that our study is not exhaustive and other functions (e.g., negative exponential functions) and connectivity schemes (e.g., k nearest neighbors) may be relevant. Visualizing the B and A matrices bearing in mind the above-mentioned trade-off between power and accuracy should be a fundamental step to reduce the number of candidates and improve the performance of the optimization.

It has been shown that explicitly integrating the resistance to movement/dispersal in spatial weighting (or connectivity) matrices allowed obtaining much more precise results than simple distance-based criteria (e.g. Rayfield et al. 2010, Hanks and Hooten 2013, Saura et al. 2014, Ver Hoef et al. 2018). Incorporating the cumulative effects of landscape fragmentation and land use change into connectivity matrices makes the distance between locations more ecologically realistic and the integration of landscape connectivity in spatial eigenvector methods is an exciting challenge.

Acknowledgment

This research was supported by the Belgian National Fund for Scientific Research (F.R.S.-FNRS) to DB. This work was performed using the Shared ICT Services Centre, Université Libre de Bruxelles and the computing facilities of the CC LBBE/PRABI. We are thankful to two anonymous reviewers and to Dr. Karen Abbott for their constructive suggestions.

Supporting information

See *Annexes* at the end of the thesis.

References

Anderson, M.J., Legendre, P. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. Journal of Statistical Computation and Simulation 62(3): 271–303.

Bauman, D., T. Drouet, S. Dray, and J. Vleminckx. 2018. Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors. Ecography 41(10): 1638–1649.

Bivand, R. 2006. spdep: spatial dependence: weighting schemes, statistics and models. R package (version 0.6-13).

Blanchet, F. G., P. Legendre, and D. Borcard. 2008. Forward selection of explanatory variables. Ecology 89:2623–2632.

Borcard, D., F. Gillet, and P. Legendre. 2011. Numerical Ecology with R. Springer. Springer, New-York.

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

Diniz-Filho, J. A. F., and L. M. Bini. 2005. Modelling geographical patterns in species richness using eigenvector-based spatial filters. Global Ecology and Biogeography 14:177–185.

Diniz-Filho, J. A. F., L. M. Bini, T. F. Rangel, I. Morales-Castilla, M. Á. Olalla-Tárraga, M. Á. Rodríguez et al. 2012. On the selection of phylogenetic eigenvectors for ecological analyses. Ecography 35:239–249.

Dray, S., D. Bauman, G. Blanchet, D. Borcard, S. Clappe, G. Guenard et al. 2018. adespatial: Multivariate multiscale spatial analysis. R package version 0.2-0.

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

Dray, S., R. Pélissier, P. Couteron, M.-J. Fortin, P. Legendre, P. R. Peres-Neto et al. 2012. Community ecology in the age of multivariate multiscale spatial analysis. Ecological Monographs 82:257–275.

Fall, A., M. J. Fortin, M. Manseau, and D. O'Brien. 2007. Spatial graphs: Principles and applications for habitat connectivity. Ecosystems 10:448–461.

Florax, R. J., and S. Rey. 1995. The impacts of misspecified spatial interaction in linear regression models. Pages 111–135 *in* L. Anselin and R. J. G. M. Florax, editors. New directions in spatial econometrics. Springer, Berlin, Heidelberg.

Fortin, M. J., and S. Payette. 2002. How to test the significance of the relation between spatially autocorrelated data at the landscape scale: A case study using fire and forest maps. Ecoscience 9:213–218.

Godinez-Dominguez, E., and J. Freire. 2003. Information-theoretic approach for selection of spatial and temporal models of community organization. Marine Ecology Progress Series 253:17–24.

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

Griffith, D. A. 2004. A spatial filtering specification for the autologistic model. Environment and Planning A 36:1791–1812.

Griffith, D. A. 2017. Some robustness assessments of Moran eigenvector spatial filtering. Spatial Statistics 22:155–179.

Griffith, D. A., and F. Lagona. 1998. On the quality of likelihood-based estimators in spatial autoregressive models when the data dependence structure is misspecified. Journal of Statistical Planning and Inference 69:153–174.

Griffith, D. A., and P. R. Peres-Neto. 2006. Spatial modeling in Ecology: the flexibility of eigenfunction spatial analyses. Ecology 87:2603–2613.

Hanks, E. M., and M. B. Hooten. 2013. Circuit theory and model-based inference for landscape connectivity. Journal of the American Statistical Association 108:22–33.

Ver Hoef, J. M., E. E. Peterson, M. B. Hooten, E. M. Hanks, and M.-J. Fortin. 2018. Spatial autoregressive models for statistical inference from ecological data. Ecological Monographs 88:36–59.

Kenkel, N. C., J. A. Hoskins, and W. D. Hoskins. 1989. Edge effects in the use of area polygons to study competition. Ecology 70:272–274.

Kostov, P. 2010. Model boosting for spatial weighting matrix selection in spatial lag models. Environment and Planning B: Planning and Design 37:533–549.

Lane, S. N., K. S. Richards, and J. H. Chandler. 1994. Developments in monitoring and modelling small-scale river bed topography. Earth Surface Processes and Landforms 19:349–368.

Legendre, P. 1993. Spatial autocorrelation: Trouble or new paradigm? Ecology 74:1659–1673.

Legendre, P., and L. Legendre. 2012. Numerical Ecology. Elsevier, Amsterdam.

Manly, B.F.J. 2007. Randomization, bootstrap and Monte Carlo methods in biology. Chapman & Hall/CRC, London.

McIntire, E. J. B., and A. Fajardo. 2009. Beyond description: the active and effective way to infer processes from spatial patterns. Ecology 90:46–56.

Munoz, F. 2009. Distance-based eigenvector maps (DBEM) to analyse metapopulation structure with irregular sampling. Ecological Modelling 220:2683–2689.

Oksanen, J., F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn et al. 2017. Package "vegan": Community ecology package (version 2.4-3).

Patuelli, R., D. A. Griffith, M. Tiefelsdorf, and P. Nijkamp. 2011. The use of spatial filtering techniques: the spatial and space-time structure of German unemployment data. International Regional Science Review 34:253–280.

Peres-Neto, P. R., and P. Legendre. 2010. Estimating and controlling for spatial structure in the study of ecological communities. Global Ecology and Biogeography 19:174–184.

Plotkin, J. B., M. D. Potts, N. Leslie, N. Manokaran, J. LaFrankie, and P. S. Ashton. 2000. Species-area curves, spatial aggregation, and habitat specialization in tropical forests. Journal of Theoretical Biology 207:81–99.

R Core Team. 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rayfield, B., M. J. Fortin, and A. Fall. 2010. The sensitivity of least-cost habitat graphs to relative cost surface values. Landscape Ecology 25:519–532.

Saura, S., Ö. Bodin, and M. J. Fortin. 2014. EDITOR'S CHOICE: Stepping stones are crucial for species' long-distance dispersal and range expansion through habitat networks. Journal of Applied Ecology 51:171–182.

Šidák, Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 62:626–633.

Spear, S. F., N. Balkenhol, M. J. Fortin, B. H. McRae, and K. Scribner. 2010. Use of resistance surfaces for landscape genetic studies: Considerations for parameterization and analysis. Molecular Ecology 19:3576–3591.

Stakhovych, S., and T. H. A. Bijmolt. 2008. Specification of spatial models: A simulation study on weights matrices. Papers in Regional Science 88:389–408.

Stetzer, F. 1982. Specifying weights in spatial forecasting models: The results of some experiments. Environment and Planning A 14:571–584.

Taylor, P. D., L. Fahrig, K. Henein, and G. Merriam. 1993. Connectivity is a vital element of landscape structure. Oikos 68:571–573.

Wagner, H. H., and S. Dray. 2015. Generating spatially constrained null models for irregularly spaced data using Moran spectral randomization methods. Methods in Ecology and Evolution 6:1169–1178.

Chapitre V

Testing and interpreting the shared space-environment fraction in variation partitioning analyses of ecological data

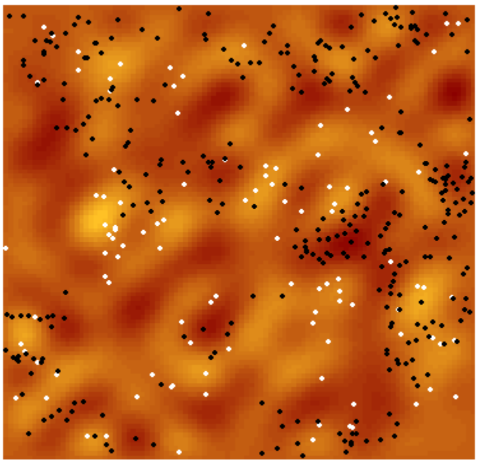
David Bauman*, Jason Vleminckx*, Olivier Hardy, Thomas Drouet

* *Equally contributing authors*

Published August 2018 in *Oikos*, doi: 10.1111/oik.05496.

Les deux chapitres précédents ont abouti à la formulation de recommendations pour les deux étapes déterminantes des méthodes basées sur des vecteurs propres spatiaux, à savoir, la sélection d'une matrice de pondération spatiale et la sélection de vecteurs spatiaux au sein de celle-ci. Ces recommendations ont été synthétisées dans une série de fonctions qui ont été intégrées au package R *adespatial*. Ces fonctions permettent d'effectuer une optimisation de sélection de matrice de pondération spatiale de façon non biaisée selon un critère d'optimisation dépendant de l'objectif visé. Ceci confère à la méthode une flexibilité vis-à-vis de situations variées rencontrées en écologie.

Dans le Chapitre V, une procédure de test de la fraction spatiale et environnementale conjointe de la partition de variation sera créée sur base des avancées méthodologiques des chapitres précédents et de deux procédures de permutations spatiallement contraintes. Une procédure de test non biaisée de cette fraction viendra combler une faille importante dans l'interprétation d'une partition de variation, dès lors que c'est précisément dans cette fraction qu'une influence spatiallement structurée de l'environnement est attendue (*induced spatial dependence*). Il s'agit donc de tester la capacité de la nouvelle procédure de test de cette fraction, c'est-à-dire ne détecter d'effet significatif que lorsqu'un processus environnemental et spatialement structuré influence la distribution des espèces.



Testing and interpreting the shared space-environment fraction in variation partitioning analyses of ecological data

Abstract

Variation partitioning analyses combined with spatial predictors (Moran's eigenvector maps, MEM) are commonly used in ecology to test the fractions of species abundance variation purely explained by environment and space. However, while these pure fractions can be tested using a classical residuals permutation procedure, no specific method has been developed to test the shared space-environment fraction (SSEF). Yet, the SSEF is expected to encompass a major driver of community assembly, that is, an induced spatial dependence effect (ISD; i.e. the reflection of a spatially structured habitat filter on a species distribution). A reliable test of this fraction is therefore crucial to properly test the presence of an ISD on ecological data. To bridge the gap, we propose to test the SSEF through spatially-constrained null models: torus-translations, and Moran spectral randomisations. We investigate the type I error rate and statistical power of our method based on two real environmental datasets and simulations of tree distributions. Ten types of tree distribution displaying contrasted aggregation properties were simulated, and their abundances were sampled in 153 regularly-distributed 20 × 20 m quadrats. The SSEF was tested for 1000 simulated tree distributions either unrelated to the environment, or filtered by environmental variables displaying contrasting spatial structures. The method proposed provided a correct type I error rate (< 0.05). The statistical power was high (> 0.9) when abundances were filtered by an environmental variable structured at broad scale. However, the spatial resolution allowed by the sampling design limited the power of the method when using a fine-scale filtering variable. This highlighted that an ISD can be properly detected providing that the spatial pattern of the filtering process is correctly captured by the sampling design of the study. An R function to apply the SSEF testing method is provided and detailed in a tutorial.

Key-words

Habitat filtering, induced spatial dependence, shared space-environment fraction, Moran's eigenvector maps (MEM), Moran spectral randomisations (MSR), plant community assembly, spatial correlation, spatially structured environmental variable, torus-translation test, variation partitioning

I. Introduction

Understanding the processes controlling the spatial distribution of organisms in natural communities is a long-standing challenge in ecology. Conspecific individuals often display aggregated distributions because of their limited dispersal capacities (Seidler and Plotkin 2006). However, such distributions can also be caused by spatially structured habitat filters (Moran 1953, Legendre 1993, John et al. 2007, Beale et al. 2010, Vleminckx et al. 2017). This reflection of the spatial structure of an environmental variable in the spatial structure of a species or community corresponds to an *induced spatial dependence* effect (Peres-Neto and Legendre 2010, Legendre and Legendre 2012), hereafter ISD, also referred to as *exogeneity* (Fortin and Dale 2005).

Abundance data displaying spatially structured patterns are often modelled through advanced spatial descriptors like *principal coordinates of neighbour matrices* (PCNM, Borcard and Legendre 2002) or their generalised form, *Moran's eigenvector maps* (MEM, Dray et al. 2006), a flexible multivariate method based on the notion of spatial autocorrelation that allows the detection of coarse (i.e. gradients) and complex multiscale spatial patterns. These spatial descriptors can be integrated, together with environmental data, into a variation partitioning analysis (Borcard et al. 1992, Peres-Neto et al. 2006, Peres-Neto and Legendre 2010). The latter consists in quantifying the proportion of variation of a univariate or multivariate response dataset (e.g., abundances of a single species or a community, respectively) explained by the environment alone, by space (i.e., spatial descriptors) alone, and jointly by space and environment (fractions [a], [c], and [b], respectively; Fig. 1) (e.g., Legendre et al. 2009, Lan et al. 2011, Baldeck et al. 2013, Bauman et al. 2016, Vleminckx et al. 2017). The pure and shared contributions of the environmental and spatial components are usually quantified through the coefficients of determinations (R^2) of simple and partial-linear regressions (for univariate response data) or canonical ordinations (e.g., redundancy analysis, RDA; for multivariate response data). These R^2 are adjusted to account for the number of explanatory variables of the model (Peres-Neto et al. 2006).

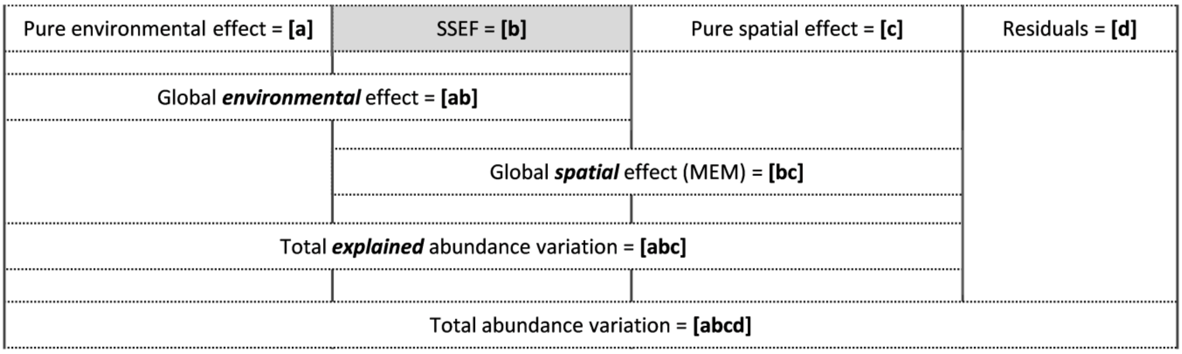

Pure environmental effect = **[a]**
SSEF = **[b]**
Pure spatial effect = **[c]**
Residuals = **[d]**
Global ***environmental*** effect = **[ab]**
Global ***spatial*** effect (MEM) = **[bc]**
Total ***explained*** abundance variation = **[abc]**
Total abundance variation = **[abcd]**

Figure 1: Scheme illustrating the different fractions of the variation partitioning analysis when considering two different explanatory variables (here the environment and the spatial predictors). The SSEF (fraction [b], emphasised in a grey cell) is calculated by subtracting the the pure environmental fraction to the global environmental fraction ([b] = [ab] – [a]), while all other fractions correspond to adjusted R^2 values calculated using linear regressions ([ab] and [bc]) or by subtractions ([a] and [c]).

The global spatial component, the pure environmental, and the pure spatial fractions of the variation partitioning (fractions [bc], [a], and [c], respectively, in Fig. 1) have been shown to be testable with a correct type I error rate and good statistical power when using residuals permutation tests, following Anderson and Legendre (1999) (Peres-Neto and Legendre 2010). Yet, a major issue remains for the global environmental component (fraction [ab]) and the shared space-environment fraction (hereafter, SSEF; fraction [b], in Fig. 1), as neither of them can be tested by classical permutation tests (Peres-Neto and Legendre 2010, Legendre and Legendre 2012). Fraction [ab] cannot be tested by classical residuals permutation test because if the response data is spatially structured, the residuals of the model will display spatial autocorrelation, hence causing an inflation of the false discovery rate (type I error rate) (Dutilleul 1993, Peres-Neto and Legendre 2010). The SSEF – which is the focus of the present study – is therefore generally removed from [ab]. This allows [a] to be tested by the residuals permutation procedure while avoiding type I error rate inflation (Peres-Neto and Legendre 2010; see also Wagner 2003, 2004, who proposes to address this issue by means of multivariate variograms in a technique called "multiscale ordination"). Fraction [a] is thus often considered alone as the reflection of habitat filtering, since it is the only testable portion of the environmental effect. Fraction [b] (i.e., the SSEF), on the other hand, is generally partialled out and ignored or is interpreted with caution (Borcard et al. 2018), as it cannot be tested using classical permutation tests. This is because fraction [b] is computed by the subtraction of other fractions adjusted following Peres-Neto et al. (2006) (e.g. [b] = [ab] – [a]), and hence has zero degree of freedom (Legendre and Legendre 2012). This is a major issue in ecology, because most ecological variables and processes being spatially structured in nature (Legendre and Legendre 2012), the SSEF is likely to represent a major portion of the habitat filtering explaining the spatial distribution of species in natural communities (i.e., an ISD).

Thus the absence of a reliable test for fraction [b] is causing most studies to potentially ignore or wrongly interpret an important niche effect explaining species distribution data. However, although we may expect the SSEF to reflect an ISD, one cannot exclude that an overlap of spatial structures in the community data and the environmental properties may occur by chance. An appropriate method to test the SSEF is therefore crucial to reliably establish the existence of a potential ISD.

In this paper, we present an original testing procedure for the SSEF. This procedure, initially suggested by Vleminckx et al. (2017), is based on two spatially-constrained permutation methods, that is, the torus-translation (TT; Upton and Fingleton 1985, Harms et al. 2001) and the Moran spectral randomisations (MSR; Wagner and Dray 2015) (see *Description of the SSEF testing procedure*). These methods provide randomisation schemes that maintain the spatial structure of the environment, thus allowing generating constrained-null values of the SSEF that can then be compared to the observed SSEF while taking spatial autocorrelation into account.

Our goal, here, is to establish whether an ISD can be reliably detected by testing the SSEF, using either TT or MSR. To do so, we computed the type I error rates and statistical power of our testing procedure on a wide range of simulation scenarios that combined different spatial features of species distribution and environmental variables.

II. Materials and methods

II.1. Description of the SSEF testing procedure

Conditions and precautions prior to perform the variation partitioning: The variation partitioning consists in partitioning the variation of a response variable, or matrix (y) into fractions explained solely and jointly by a set of environmental variables and a set of spatial variables. A preliminary condition to perform such analysis is for the global models of y against the complete set of environmental and spatial variables to both be significant at a predefined significance threshold (usually fixed at 0.05). The global spatial model is tested using classical residuals permutations (Anderson and Legendre 1999). The global environmental model, however, is tested by comparing the observed global adjusted R^2 with 999 null values obtained with a MSR-permutation procedure (Wagner and Dray 2015) in order to take spatial autocorrelation into account, hence avoiding a type I error rate inflation. Testing global models allows avoiding type I error rate inflations when performing model selections (see Blanchet et al. 2008). The variation partitioning (and hence the test of the SSEF) should only be conducted if these preliminary conditions are respected. If both global models are significant, then a variable selection – preferably the forward selection with double stopping criterion (Blanchet et al. 2008; see Appendix S1: Section 1) – must be performed for the spatial variables (Bauman et al. 2018a) and can also be performed for the environmental variables (see details in Appendix S1: Section 2). A variation partitioning is then performed to calculate the SSEF as described in the introduction section.

Testing the SSEF: The procedure to test the SSEF starts by testing whether the selected set of environmental variables displays a significant spatial structure. Although this is not mandatory to perform the variation partitioning, it is a necessary condition to test the SSEF, as there can be no significant SSEF if y and the environment do not both present a spatial structure. This condition allows reducing the risk of a false positive by not performing the test if it does not make sense. The test of this condition is performed by the classical residuals permutation procedure of a global RDA of the environmental variables against the MEM variables generated from an optimised spatial weighting matrix (see Bauman et al. 2018b and Appendix S1: Section 3 for the optimization of a spatial weighting matrix). If the environmental variables are spatially structured, then the SSEF is tested either with torus-translations (TT) or with Moran spectral randomisations (MSR), that is, two spatially-constrained randomisation procedures.

The first randomisation method (TT) applies for regular sampling designs only, where sampling quadrats are disposed according to a grid-like structure. Torus translations have been used in many studies which demonstrated significant environmental signals on floristic assemblages, in diverse regions and ecosystems (e.g., Harms et al. 2001, Noguchi et al. 2007, Itoh et al. 2010, Chuyong et al. 2011, Vleminckx et al. 2014, Vleminckx et al. 2015, Muledi et al. 2017, Vleminckx et al. 2017). Here, the TT consists in spatially translating the values of the environmental variables in each quadrat by a random number of sampling units (one unit corresponding to the space between adjacent quadrats) along the Y and X axes of the grid. Species abundance and environmental data are thus de-correlated, but their original spatial structure is preserved. The random translation procedure is repeated k times (e.g. 999 times), and each time, the SSEF (i.e. fraction [b]) is recomputed from the randomised environment ([b] = [ab] - [a]). The observed SSEF is then tested by comparing its value to the k null-values obtained after translations (see below).

The second method, MSR, uses information on the spatial relations among sampled points, in a similar way as it is done with MEM, to re-create artificial variables (abundance or environmental) displaying spatial structures that are very similar to the original ones. More specifically, the method first uses a linear combination of MEM variables to capture the spatial patterns of the environmental variables. Secondly, it uses the detected spatial structures in a conditional randomisation procedure that maintains the original structure of the data, allowing testing a relation with another set of variables (here, the response variable(s)) while maintaining a correct type I error rate (see Wagner and Dray 2015 for details). MSR have the considerable advantage of being applicable to any type of sampling design, unlike TT that are restricted to regular designs. It has been shown to be a powerful method of randomisation preserving the spatial properties of univariate or multivariate data at all spatial scales, regardless of the sample size or type of sampling design (Wagner and Dray 2015) (see examples of environmental maps produced by TT and MSR in Fig. 2, step 3). As for TT, the MSR allows generating a distribution of constrained-null SSEF values, which is then compared to the observed SSEF. In both randomisation procedures (the TT and the MSR), the *p*-value is computed as the proportion of null values of SSEF equal to or larger than the true unpermuted value, including the latter, for a one-tailed test in the upper tail (Hope 1968, Legendre and Legendre 2012).

The procedure also addresses negative adjusted SSEF differently depending on the origin of the negative value. 1) These negative SSEF can arise from the R^2 adjustment procedure itself, and are then associated to explanatory variables that explain less of the response dataset than expected by chance for random normal deviates (Legendre and Legendre 2012, Borcard et al. 2018). In this case, the unadjusted SSEF is positive and the negative value can be considered as a zero. In this case, the *p*-value is computed as mentioned above. 2) However, a negative adjusted SSEF can also bear a real ecological meaning, in which case it cannot be considered as a zero and it should be tested too. This type of negative SSEF can arise when two explanatory variables are negatively correlated to one another and are both strongly correlated to the response variable (Legendre and Legendre 2012). This situation is differentiable from the first one because the unadjusted SSEF is also negative. In this case, our procedure computes the *p*-value as the proportion of null values of SSEF equal to or smaller than the true value for a one-tailed test in the lower tail. 3) A last source of negative SSEF may arise when *suppressor* variables are present (i.e. one variable that is not or nearly not correlated to the response variable, but is correlated to another explanatory variable, itself correlated to the response variable; see details in Azen and Budescu 2003, Beckstead 2012). Since the unadjusted SSEF resulting from a suppression effect is also negative, the SSEF testing procedure has no mean to differentiate it from a negative SSEF having actual ecological meaning. Nevertheless, suppression effects can easily be handled and avoided prior to performing the variation partitioning by (1) testing the significance of both explanatory components, and by (2) selecting an appropriate subset of MEM variables prior to performing the partitioning. We detailed and illustrated this in Appendix S2.

It is worth mentioning that, although this study focuses on the SSEF, both TT and MSR could also be used to test the global environmental component (i.e., fraction [ab] in Fig. 1; see Vleminckx et al. 2017), as they both correct the risk of type I error rate inflation caused by spatial autocorrelation that hinders the use of a classical permutation scheme for this fraction.

II.2. Simulations to calculate the statistical performances of the SSEF testing procedure

We carried out computer simulations of tree distributions in a forest environment to investigate the performance of our new SSEF testing procedure, and thereby infer a possible ISD. To do so, we used environmental variables from two existing forest datasets either structured at different spatial scales, or spatially randomised, that we combined to several types of tree distribution patterns. The latter corresponded to individuals either clustered at different scales (mimicking a dispersal limitation for example), or distributed at random. The combination of these distribution patterns to the different environmental variables produced a wide range of simulation scenarios (hereafter, SS). Each scenario was generated 1000 times, and each time a variation partitioning was performed, provided that both the global environmental and spatial models were significant, as explained earlier. This allowed calculating the type I error rate and statistical power of the testing procedure. In the next sections, we first present the environmental variables and the properties of the types of tree distributions used in each SS. We then describe the sampling design and the procedures used to generate the spatial descriptors (MEM) and calculate the statistical performance of the SSEF testing procedure.

II.2.1. Environmental data

We used environmental data from two 50-ha forest areas divided into 1250 quadrats of 20 × 20 m, one area located on Barro Colorado Island (BCI, Panama, Condit et al. 2012), and the other in Korup National Park (KNP, Cameroon, Chuyong et al. 2004). In each of these two areas, two variables showing contrasted spatial patterns were chosen: Elevation in BCI and KNP displayed coarse grained spatial structures (further referred to as broad-scale patterns), while topographical slope was spatially correlated at a relatively much finer scale (fine-scale patterns). We also added four supplementary environmental variables artificially created by spatially randomising the elevation and topographical slope values of BCI and KNP, yielding a total of eight environmental variables to be used as explanatory variables of the simulated tree abundances (Fig. S1). We then selected a subset of three variables displaying contrasted spatial structures (Fig. 2, step 1): the elevation in BCI (broad-scale spatial structure), the topographical slope in KNP (fine-scale), and the spatially randomised elevation values in BCI (no spatial structure). Each of these three variables was used to filter the simulated tree distributions (see sections below) in order to mimic a habitat filtering. Standardised values of the environmental variables are available in Table S1. Supplementary Fig. S1 shows heat maps of the eight environmental variables in the 1250 quadrats.

II.2.2. Simulating homogeneous tree species distributions

Ten types of tree distribution patterns were simulated over each of the three 50-ha environmental maps. In distribution patterns 1 to 9, we used a spatially explicit model developed by Plotkin *et al.* (2000) to generate artificial, yet realistic, tree distributions following a Poisson cluster process (PCP). The first step of this process is to randomly distribute “parents” with density ρ over one of the three selected environmental maps (see previous section). Each parent produces a number of offsprings that follows a Poisson distribution of mean m. The position of each offspring around the parent is then derived from a radially symmetric Gaussian distribution of variance s^2 (thus here, the PCP corresponds more exactly to a *homogeneous Thomas process*, Potts et al. 2004). This procedure therefore produces clumps (or aggregates) of trees for which the number and width are controlled by parameters ρ and s^2, respectively. Distribution patterns 1 to 9 corresponded to the combinations of three s^2 and three ρ values, chosen to reproduce contrasted types of plant aggregates (Fig. 2, step 2). Additional details about individual sampling are provided in Appendix S1: Section 6. Distribution pattern 10 consisted in simulating a spatially randomised tree distribution (no spatial structure; Fig. 2, step 2) by generating a number of random X-Y coordinates (using a random uniform distribution) that ensured obtaining an intermediate number of individuals (400 ± 50) compared to distribution patterns 1 to 9.

II.2.3. Simulation scenarios

A simulation scenario, hereafter SS, corresponded to one of the ten distribution patterns described in the previous section, that either remained independent from the environment (hereafter, *generalist* population), or that was filtered by one of the three environmental variables in Fig. 2 (*specialist* population). In total, there were therefore ten types of generalist populations (corresponding to the ten different types of distribution patterns) + three types of environmental filters × ten types of specialist distributions (with the same spatial properties as the generalist ones) = 40 SS. A specialist population was created by first generating a generalist population. The probability to keep each individual then followed a Gaussian probability density function that depended on the value of the filtering environmental variable in its quadrat (see details in Appendix S1: Section 4). All eight environmental variables (Fig. S1) were used as explanatory variables when performing a variation partitioning on generalist and specialist populations, but only one was used as a filter for specialist populations. Table 1 summarises the different SS (properties of the environmental filter, statistical performance tested, and underlying ecological question). Supplementary Table S2 provides further details for each SS.

Besides these 40 SS (further referred to as “main scenarios”), three additional SS were generated. The latter adopted a different approach to simulate complementary situations, using linear combinations of MEM variables to which a random noise was added to create spatially structured species abundances and environmental variables (see details in Appendix S1: Section 5). The first two additional SS corresponded to extreme cases in which (i) the environment and the abundance displayed totally independent spatial patterns, and (ii) the environment was entirely spatially structured and totally correlated to the spatial structure of the abundance. The third additional SS corresponded to a response variable strongly correlated to an environmental variable, itself negatively correlated to three MEM variables positively correlated to the response variable. This SS therefore simulated negative SSEF corresponding to an ISD, hence allowing testing the statistical performances of our SSEF testing procedure for this type of situation.

II.2.4. Sampling design

Once all SS were constructed, the environmental and abundance data were sampled in 153 regularly-spaced quadrats among the 1250 quadrats of the 50-ha study area (12.24% of the whole area’s surface, 6.12 ha; see Fig. 2, steps 2 and 3), a realistic sampling effort that fits the range of sampling sizes found in the literature (e.g., Kadavul and Parthasarathy 1999, Condit et al. 2004, Zent and Zent 2004, Biwolé et al. 2015, Muledi et al. 2017). These sampled quadrats were regularly disposed according to a 9 by 17 grid-like structure (see Fig. 2, step 2), a spatial configuration that allowed performing torus-translations. Adjacent sampled quadrats were distant of 60 m both along the X and Y axes of the grid. Supplementary details about the sampling of individuals are provided in Appendix S1: Section 6.

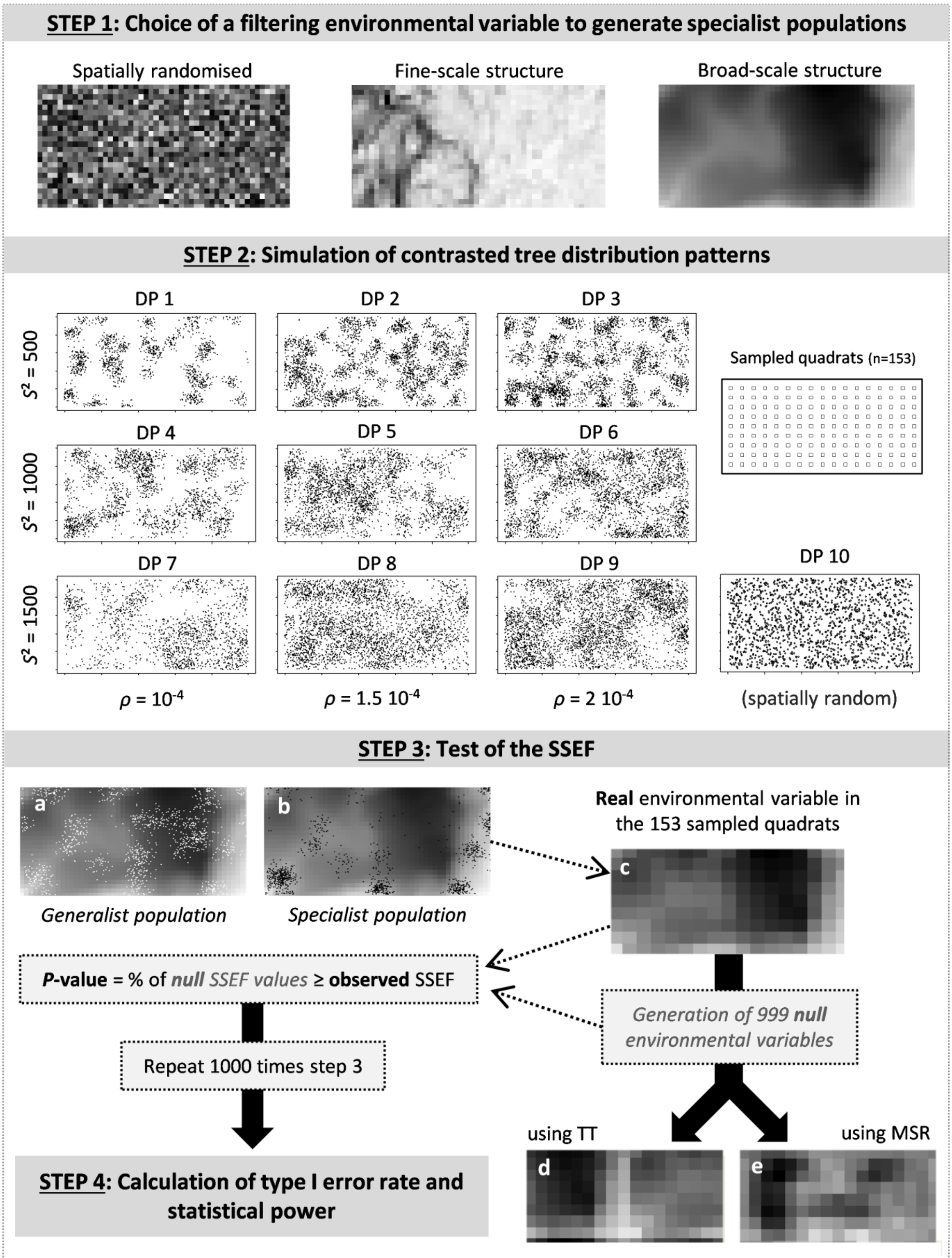

STEP 1: Choice of a filtering environmental variable to generate specialist populations
Spatially randomised
Fine-scale structure
Broad-scale structure
STEP 2: Simulation of contrasted tree distribution patterns
DP 1
DP 2
DP 3
S² = 500
DP 4
DP 5
DP 6
S² = 1000
DP 7
DP 8
DP 9
S² = 1500
DP 10
Sampled quadrats (n=153)
ρ = 10-4
ρ = 1.5 10-4
ρ = 2 10-4
(spatially random)
STEP 3: Test of the SSEF
a
b
Generalist population
Specialist population
Real environmental variable in the 153 sampled quadrats
c
P-value = % of null SSEF values ≥ observed SSEF
Generation of 999 null environmental variables
Repeat 1000 times step 3
using TT
using MSR
d
e
STEP 4: Calculation of type I error rate and statistical power

Figure 2: Schematic description of the procedure used to evaluate the type I error rate and power of the SSEF test. Step 1: heat maps showing the spatial structure of the three environmental variables (broad-scale, fine-scale, and spatially randomised) used for filtering tree abundances.

Step 2: examples of artificial tree distribution patterns (DP) obtained (i) by using a Poisson cluster process, for three different values of s^2 and ρ (DP 1 to 9), and (ii) by completely randomising x and y coordinates (spatially random distribution, DP 10). The disposition of the 153 sampled quadrats is also shown on the right. Step 3: examples of simulated generalist (a) and specialist (b) tree distributions (using a simulation of DP 2) over one of the three environmental maps. The specialist population was negatively correlated to the environmental variable. (c) Heat map showing the spatial structure of the environmental variable in the 153 sampled quadrats (artificially joined here for simplicity). From this map, null environmental variables were generated (999 times) using either torus-translations (TT) or Moran Spectral Randomisations (MSR), in order to test the SSEF (maps d and e show two examples of environmental variables generated by TT or MSR). The variation partitioning was conducted if, and only if both the global model of the environment and space were significant. The SSEF test was performed only if the environment was significantly spatially structured (see *Description of the SSEF testing procedure*). Step 4: Calculation of either the type I error rate or the statistical power (depending on the simulation scenario, see Table 1) associated to the test of the SSEF, based on 1000 repetitions of step 3.

II.2.5. Spatial variables

A distance-based spatial weighting matrix (db-MEM, Dray et al. 2006) was used to generate the MEM variables from the spatial coordinates of the 153 subsampled quadrats. The connectivity matrix was based on the minimum distance that kept all these quadrats connected (i.e. the largest edge of the minimum spanning tree), and the resulting links were weighted by 1 - $(D_{ij}/4t)^2$, where D_{ij} is the euclidean distance between sampled units *i* and *j*, and *t* the threshold connexion distance (see details in Dray et al. 2006). The arbitrary choice of this spatial weighting matrix was motivated by the fact that db-MEM are well-suited for regular sampling designs (Bauman et al. 2018b), and because the benefit of optimising the spatial weighting matrix is small in comparison with the cost of statistical power for those sampling designs (Dray et al. 2006, Bauman et al. 2018b). We only considered the 58 MEM variables associated to positive eigenvalues and hence corresponding to positively autocorrelated patterns, as the spatially structured simulated populations (distribution patterns) all displayed positively autocorrelated patterns.

Table 1: Characteristics of the 40 main simulation scenarios (SS) used in our study. The first column refers to the range of scenarios described. Column 2 indicates whether no environmental filtering is simulated (SS 1 to 10) or if there is an environmental filtering by a spatially randomised variable (SS 11 to 20) or by a variable displaying fine-scale (SS 21 to 30) or broad-scale spatial structure (SS 31 to 40) (See Fig. 2, step 1). Column 3: Type of tree distribution pattern (DP; see Fig. 2, step 2): aggregated (DP 1 to 9) or spatially random (DP 10). Column 4 indicates the type of statistical performance investigated: type I error rate if the SS does not model an induced spatial dependence effet (ISD), or statistical power if it does model an ISD.

SS	Environmental filter	DP	Performance
1 to 10	None	Aggregated (DP 1 to 9) Random (DP 10)	Type I error rate
11 to 20	Spatially Random	Aggregated (DP 1 to 9) Random (DP 10)	
21 to 30	Structured (fine-scale)	Aggregated (DP 1 to 9) Random (DP 10)	Statistical power
31 to 40	Structured (broad-scale)	Aggregated (DP 1 to 9) Random (DP 10)	

II.2.6. Statistical performance of the SSEF testing procedure

Each of the 43 SS was replicated 1000 times, and each time the SSEF testing procedure was performed on the basis of 999 TT or MSR (providing that the global environmental and spatial models were both significant). The statistical performance of the two constrained randomisation procedures (TT and MSR) to test the adjusted SSEF was assessed as the proportion of significant *p*-values among the 1000 replicated simulations, and either corresponded to the type I error rate or to the statistical power, depending on the SS (see Table 1). A schematic overview of the different steps of our simulation procedure for the 40 main SS is presented in Fig. 2: the three environmental maps used as filters to generate specialist populations (step 1 on the figure), the ten types of simulated tree distribution patterns (step 2), the test of the SSEF (step 3) and the calculation of type I error rate and statistical power associated to this test (step 4). Note that all the simulations presented in this study are performed on univariate response variables (i.e. the abundance of one single species). The SSEF testing procedure is however equally useful in a multivariate framework (e.g. community data; see Appendix S4 for an illustration).

All simulations were performed in the R statistical environment (R Development Core Team 2018), using packages *car* (Fox and Weisberg 2011), *adespatial* (Dray et al. 2018), *spdep* (Bivand and Piras 2015), *splancs* (Rowlingson and Diggle 2015), *tripack* (Renka 2013), and *vegan* (Oksanen et al. 2015). The R script used to perform these simulations is provided in Appendix S3. A user-friendly R function to test the SSEF of a variation partitioning – *envspace.test* – was implemented in the package *adespatial* and is detailed in an R tutorial (Appendix S4).

III. Results

Overall, we found that the type I error rate associated with the test of the SSEF (SS 1 to 20, see Table 1) was always below the commonly accepted threshold of 0.05 (Fig. 3a and 3b). Furthermore, the statistical power obtained when the abundance was filtered using the fine-scale environmental variable was considerably low compared to when it was filtered by the broad-scale environmental one (Fig. 3c and 3d). The frequency at which each MEM variable was selected to model the spatial patterns of the different SS are presented in supplementary Table S3. The latter indicates that the spatial patterns resulting from the combination of the different distribution patterns and environmental variables encompassed all broad, intermediate, and fine spatial scales.

III.1. Type I error rate

The type I error rate associated with the test of the SSEF never exceeded 0.05 (Fig. 3a and 3b), and so when using either TT or MSR to generate null SSEF values. The maximum type I error value reached 0.037 for distribution pattern 6 (numerous and medium size tree aggregates; Fig. 2, step 2) when simulating generalist tree populations (SS 1 to 10; Fig. 3a). Type I error rate values were always higher for generalist populations compared to when the abundance was filtered using a spatially randomised environmental variable (SS 11 to 20; Fig. 3b). More precisely, type I error rate reached, on average, 0.023 ± 0.012 (mean and standard deviation of type I error rate values obtained using TT and MSR tests) among generalist populations, while it never exceeded 0.004 in SS 11 to 20 (spatially randomised environmental filter; mean and standard deviation of 0.0017 and 0.0014, respectively).

When simulating the particular additional scenario where the filtering environment was spatially structured and where both the global environmental and spatial components (i.e. fractions [ab] and [bc], see Fig. 1) were significant but had no common variation (i.e. no SSEF expected), the SSEF was close to zero and we did not observe any type I error (see additional scenario 1 in Table S4). No excess of type I error rate (0.007) was found in the additional scenario addressing the presence of negative adjusted R^2 corresponding to an ISD (additional scenario 3; Table S4).

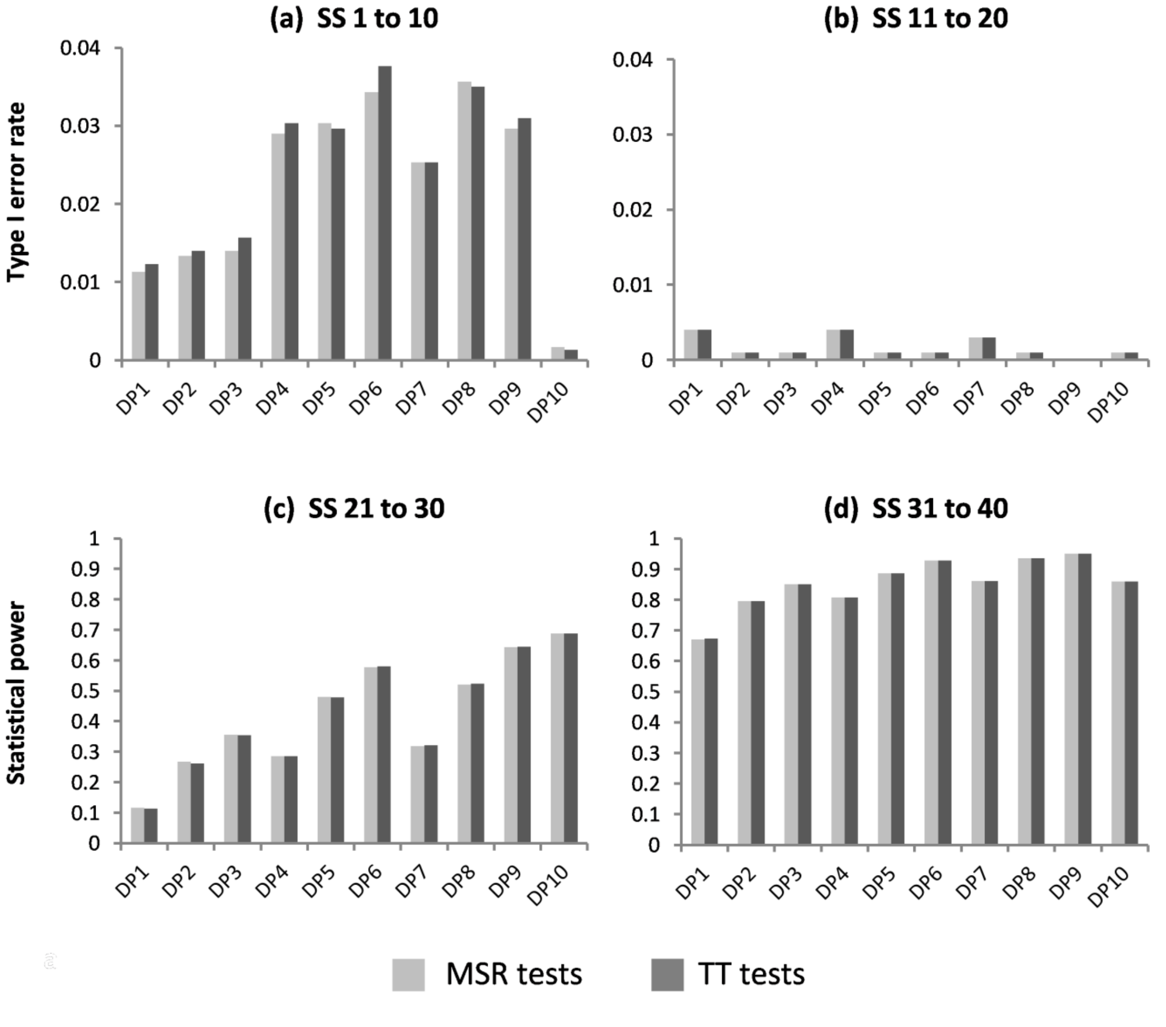


Figure 3: Type I error rate (a and b) and statistical power (c and d) associated to the test of the SSEF using MSR and TT tests (light and dark grey histograms, respectively). Tests of significance were carried out using a 0.05 significance level. (a) and (b): Type I error rate calculated using simulation scenarios (SS) 1 to 10 (no environmental filtering; see Table 1) and SS 11 to 20 (abundance filtered by the spatially randomised environmental variable), respectively. (c) and (d): statistical power obtained when the abundance was filtered using the fine-scale (SS 21 to 30) and the broad-scale (SS 31 to 40) environmental variable, respectively.

III.2. Statistical power

For all distribution patterns (1 to 10; Fig. 2, step 2), the statistical power associated to the SSEF test (SS 21 to 40, see Table 1) was substantially lower when using the fine-scale environmental filter (SS 21 to 30; Fig.3c) compared to when using the broad-scale one (SS 31 to 40; Fig. 3d). More specifically, it reached, on average, 0.426 ± 0.181 (mean and standard deviation of power values obtained using TT and MSR tests) across SS 21 to 30 (fine-scale environmental filter), with a minimum and a maximum value of 0.113 in distribution pattern 1 (low number of small tree aggregates; see Fig. 2, step 2) and 0.688 in distribution pattern 10 (spatially randomised tree distribution), respectively. In SS 31 to 40 (broad-scale environmental filter), however, power values reached, on average, 0.855 ± 0.080, with a minimum and a maximum value of 0.671 in distribution pattern 1 and 0.950 in distribution pattern 9 (numerous and large tree aggregates), respectively. Finally, high power values (> 0.970) were observed in the particular additional scenario where the global environmental effect (fraction [ab], see Fig. 1) was entirely comprised in the global spatial effect (fraction [bc]) (additional scenario 2 in Table S4). In additional scenario 3 (modeling ISD patterns with negative adjusted SSEF), the statistical power reached 0.784.

Mean and standard deviation values (calculated over the 1000 replicated simulations) of the SSEF in each simulation scenario are detailed in Table S5.

IV. Discussion

The variation partitioning is the most widely used analysis to assess the relative and shared effects of environmental heterogeneity and spatial predictors on species distributions (Borcard et al. 1992, Peres-Neto et al. 2006, Peres-Neto and Legendre 2010). However, no specific procedure has been proposed to test the shared effect of space and environment (the SSEF). This is a major gap as the SSEF is expected to reflect an induced spatial dependence effect (ISD) when one is present and when the sampling design allows detecting it. The absence of a reliable test therefore jeopardises any ecological interpretation of this fraction. Hence the objective of the present study, which was to introduce a testing procedure for the SSEF. We found that the procedure presented type I error rates < 0.05, indicating a very small risk to wrongly detect ISD when there is none. Efficient statistical power values (> 0.8) were obtained when simulating large and numerous tree aggregates filtered by an environmental variable displaying a broad-scale spatial structure. The lower power values (0.113 to 0.688) observed with the fine-scale environmental filter were most likely related to a limited power of MEM at scales too close or inferior to the fine-scale spatial resolution of the sampling design, as we discussed below. The SSEF values obtained through our simulations always fitted realistic and expected values. The mean SSEF values of the generalist populations reached 0.009 (min/max: 0.000/0.022) across SS 1 to 10, which was within the range of small values expected from species unrelated to the environment. The mean SSEF values of the simulated specialist populations reached 0.129 (min/max: 0.041/0.198) and 0.289 (min/max: 0.117/0.461) when the filtering environment displayed fine-scale and broad-scale spatial structures, respectively. These correspond to the range of values obtained with real datasets (e.g., Legendre et al. 2009, Chang et al. 2013, Punchi-Manage et al. 2014, Prada and Stevenson 2016). It is worth reminding that, although the SSEF testing procedure has been presented here on univariate response variables, the procedure can equally be used for multivariate response data (see illustration in Appendix S4).

It is also worth mentioning that, although it was not the primary focus of the study, tests based on both TT and MSR provided type I error rates below the significance level (0.05) as well as high statistical power values for the global environmental component (fraction [ab]; not shown).

IV.1. The TT and MSR tests prevent detecting spurious induced spatial dependence effects

The risk of detecting a significant species-environment association when the environment does not influence a species distribution but when both the species and the environment are spatially structured is a well-known issue in ecology (Dutilleul 1993, Legendre et al. 2002). The latter issue arises because aggregated populations and environmental structures may overlap by chance and inflate the SSEF, which in turn may enhance the risk of detecting a spurious ISD. However, we demonstrated that the TT and MSR tests of the SSEF provided levels of type I error rate below the usual threshold of 0.05 in a wide variety of scenarios (Fig. 3a and 3b). These tests were thus well able to avoid the detection of a false ISD when no environmental filtering was simulated (Fig. 3a), or when the filtering environment was not spatially structured (Fig. 3b) or was structured independently from the response variable (Table S4), and so regardless of the type of spatial patterns present in the tree distribution. Type I error rate was particularly low in the second case (Fig. 3b), but also when the abundance was spatially randomised while the environment was not structured (Fig. 3a; SS 10), which was expected since we forbid the SSEF test to be performed if the environment or the abundance was not spatially structured, to avoid unnecessary risks of false positives.

There is a situation, however, in which spurious correlations between environmental and species data are expected, that is, when a coarse gradient, or spatial trend, is present in both the environment and the species (Borcard et al. 2004). It has been advocated that spatial trends should be removed prior to using PCNM (i.e. detrending), to avoid using spatial predictors to model the trends while they could be used to model finer and more complex patterns (Borcard and Legendre 2002, Borcard et al. 2004). The statistical power and accuracy related to detrending the data or not should however be evaluated in the broader framework of MEM, given the flexibility of the method and recent advances increasing its power and accuracy (Bauman et al. 2018a, b). Regardless of whether a detrending is used or not, the presence of a trend should be tested and explicitly considered in the interpretation of a significant SSEF.

IV.2. The power to detect induced spatial dependence effects depends on the scale of the environmental filter and on sampling design characteristics

Our results indicated that the statistical power of the SSEF testing procedure (using either TT or MSR) to detect an ISD effect greatly depended on the scale of the spatial structures of the filtering environmental variable. Note that 'scale' here exclusively refers to the spatial characteristic 'focus', *sensu* Scheiner (2011), that is, the spatial area covered by the aggregated sampling units (e.g. quadrats) of a spatial pattern. Good power values (considering values higher than 0.7 and reaching up to 0.950) were obtained when trees were filtered by the broad-scale environmental variable, and provided that they were distributed in relatively large and numerous aggregates, or randomly distributed (distribution patterns 2 to 10 in Fig. 3d). However, when the fine-scale environmental filtering variable was used, only four types of tree distribution patterns, characterised by relatively large and numerous aggregates or random distribution (distribution patterns 5, 6, 8, 9 and 10), provided power values close to or higher than 0.5, while lower power values were obtained for the remaining six tree distribution patterns (down to 0.113 in distribution pattern 1, see Fig. 3c).

These lower power values most likely arose from a limited power of MEM at scales too close or inferior to the fine-scale spatial resolution of the sampling design (60 m). Indeed, the MEM method has already been shown to display low statistical power when the spatial eigenvectors associated to very small eigenvalues (absolute value) are needed to model the spatial pattern of the response dataset (Bauman et al. 2018a). This was the case here, as the eigenvectors supposed to capture the patterns filtered by the environmental variable structured at fine spatial scale were among the last MEM variables associated to positive eigenvalues. The latter variables are those modelling the finest positively autocorrelated spatial patterns detectable by the sampling design. As a consequence, our method for the test of the SSEF is expected to provide high statistical power as long as the spatial pattern of the ecological process responsible for an ISD is not exclusively explained by the MEM variables associated to the smallest eigenvalues (in absolute value). This, in turn, depends simultaneously on the sampling design characteristics (e.g., grain and spatial extent of the study, the "sampling interval" *sensu* Legendre and Legendre 2012, p. 786; see their Fig. 13) and on the scale of the spatial pattern of the process responsible for the ISD. It is therefore worth mentioning that a non-significant SSEF only informs that no evidence of an ISD could be found.

When using any of the structured environmental filters, the statistical power of the SSEF test (using either TT or MSR) increased with the number of clumps (ρ), which was expected since ρ controlled for the total number of individuals eventually sampled in the 153 quadrats. For similar ρ values, the power also increased with the size of the clumps (s^2) (see Fig. 3b and 3c). The same phenomenon has been observed by Itoh et al. (2010) for other species-environment association statistics. This increase of power with s^2 presumably arises from a higher chance of overlap between a few large tree clusters and broad-scale environmental structures, than between a more complex patchwork of abundant and relatively smaller tree clusters and environmental patterns. The same phenomenon is likely to explain the relatively higher type I error rates for generalist populations (SS 1 to 10) compared to SS 11 to 20 (where no spatial structures were present in the filtering environment). It is also likely to results from the fact that large dispersal capacities and population size increased the chance for plants to be located on more various environmental conditions than when tree distributions are highly clumped, thereby reducing the variance of the SSEF after the randomisation procedure used in the TT or MSR tests.

IV.3. Spatial variable selection and variation partitioning

When constructing the spatial predictors used in the variation partitioning, the choice of the spatial weighting matrix used to generate the MEM variables did not matter much in our case, as we performed our simulations on a regular grid (Dray et al. 2006). We therefore used the same spatial weighting matrix to test a spatial signal in both the response variable and the environmental dataset. However, when the sampling design is irregular, which is often the case with real datasets, the selection of the spatial weighting matrix should be optimised separately for the response and the environmental data to ensure a maximal statistical power and accuracy of MEM for both datasets (Bauman et al. 2018b; see Appendix S1: Section 3 and illustration in Appendix S4).

In a previous study, Gilbert and Bennett (2010) highlighted a problem of accuracy related to the MEM used in variation partitioning analyses. They also showed that the fractions associated to a spatial effect (i.e. [b], [c], and [bc]) were overestimated, even when the environment had no spatial structure. However, as illustrated by our results, our SSEF testing procedure displayed a correct type I error rate throughout the wide range of simulation scenarios that were used. This discrepancy is likely to arise (1) from the fact that the authors considered the value of the SSEF even though the environment was not spatially structured, and (2) from the absence of a test for the SSEF. In addition, to evaluate accuracy, the authors compared what they called the “true variation” of the different partitioning fractions (i.e. the reference values at the level of a complete grid of 129 × 129 cells, that is, 16641 cells) to the values obtained from different sampling designs of very small samples with respect to the complete grid (i.e. either 64 or 256 cells). It seems therefore that the lack of accuracy that they attribute to MEM may have simply arisen from the sampling procedure. This may be confirmed by two recent studies in which the power and spatial R^2 estimation accuracy of MEM have been shown to be high in a wide variety of simulation scenarios, when considering the reference values of variation on the sampled cells instead of on a much bigger grid (Bauman et al. 2018a, b). It is also worth mentioning that using distance-based MEM has recently been shown to yield poor results in terms of power and accuracy for irregular sampling designs, compared with the more powerful and accurate graph-based MEM (Bauman et al. 2018b). The use of distance-based MEM in studies based on irregular sampling designs is therefore expected to yield inaccurate estimations of the fractions of the variation partitioning. In those cases, the highest accuracy of MEM is reached with an optimisation of the selection of the spatial weighting matrix combined with the forward selection with double stopping criterion (Bauman et al. 2018a, b). These recent studies, together with our results, strongly suggest that variation partitioning can be reliably performed provided that: (i) preliminary conditions are met (significance of both the global environmental and the global spatial models), (ii) an optimal set of MEM variables is selected among multiple spatial weighting matrices (see Bauman et al. 2018b and Appendix S4), and (iii) the SSEF is only tested when the environment displays a significant spatial structure.

IV.4. Conclusion

In this study, we showed that our testing procedure did not detect any ISD when there was none, that is, when the environment was not spatially structured or when no environmental filtering occurred (Table S3). We also showed that the procedure was able to detect a significant SSEF when an ISD had been simulated, and so in many contrasting simulation scenarios (provided that the scale of the spatial patterns was not too close or inferior to the fine-scale spatial resolution of the sampling design). In the SSEF testing procedure, both TT and MSR – when used with a well-selected subset of MEM variables (see Appendix S2) – were able to detect an ISD related to negative SSEF while avoiding confusion with suppression effects and negative value obtained by the R^2 adjustment. In addition, both TT and MSR can be used to test the global environmental component while controlling for spatial autocorrelation issues, hence avoiding the type I error rate inflation of classical permutation tests. It is worth mentioning that in real data studies, a significant SSEF may also be caused by spatially structured unmeasured environmental variables that are correlated to one or several of the measured structured environmental variables (Borcard and Legendre 1994, Peres-Neto and Legendre 2010). This, however, is also true for fraction [ab] and fraction [a], as an unmeasured variable could always be correlated to both the response data and the spatially structured or not spatially structured portion of variation of a measured explanatory variable. Inferring causal relations should always be done with caution, and a priori hypotheses and ecological theory or knowledge should underlie the choice of the environmental variables measured and included in the variation partitioning to allow strong interpretations of the SSEF and the other fractions of the variation partitioning analysis (McIntire and Fajardo 2009).

Declarations

JV and DB contributed equally to the study. DB is a PhD candidate currently funded by the *Fund for Scientific Research of Belgium* (F.R.S.-FNRS). JV is currently funded by the labEX CEBA (*Centre d'Étude de la Biodiversité Amazonienne*, managed by CNRS, France) and was funded by the *Belgian Fund for Training to Research in Industry and Agriculture* (FRIA) when the study was designed. We are very grateful to Pierre Legendre, Daniel Borcard, Cajo ter Braak, and an anonymous reviewer for their constructive comments and suggestions.

Supporting information

See *Annexes* at the end of the thesis.

References

Anderson, M. J. and Legendre, P. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. – J. Stat. Comput. Sim. 62(3): 271–303.

Azen, R. and Budescu, D.V. 2003. The dominance analysis approach for comparing predictors in multiple regression. – Psychol. Methods 8(2): 129–148.

Baldeck, C. A. 2013. Soil resources and topography shape local tree community structure in tropical forests. – P. R. Soc. B. 280: 20122532.

Bauman, D. et al. 2016. Multiscale assemblage of an ectomycorrhizal fungal community: the influence of host functional traits and soil properties in a 10-ha miombo forest. – FEMS Microbiol. Ecol. 92(10): fiw151.

Bauman, D. et al. 2018a. Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors. – Ecography 41(10): 1638–1649.

Bauman, D. et al. 2018b. Optimizing the choice of a spatial weighting matrix in eigenvector-based methods. – Ecology: doi: 10.1002/ecy.2469.

Beale, C. M. et al. 2010. Regression analysis of spatial data. – Ecol. Lett. 13: 246–264.

Beckstead, J. 2012. Isolating and examining sources of suppression and multicollinearity in multiple linear regression. – Multivariate Behav. Res. 47(2): 224–246.

Bivand, R. and Piras, G. 2015. Comparing implementations of estimation methods for spatial econometrics. – J. Stat. Softw. 63: 1–36.

Biwolé, A. B. et al. 2015. New data on the recent history of the littoral forests of southern Cameroon: an insight into the role of historical human disturbances on the current forest composition. – Plant Ecol. Evol. 148: 19–28.

Blanchet, F. G. et al. 2008. Forward selection of explanatory variables. – Ecology 89: 2623–2632.

Borcard, D. et al. 1992. Partialling out the spatial component of ecological variation. – Ecology 73: 1045–1055.

Borcard, D. and Legendre, P. 1994. Environmental control and spatial structure in ecological communities an example using Oribatid mites (Acari, Oribatei). Environmental and ecological statistics 1: 37–53.

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(7): 1826–1832.

Borcard, D. et al. 2018. Numerical ecology with R, second edition. Use R! Series, Springer International Publishing, Cham, Switzerland.

Chang, L.W. et al. 2013. Better environmental data may reverse conclusions about niche- and dispersal-based processes in community assembly. – Ecology 94(10): 2145–2151.

Chuyong, G. B. et al. 2004. Korup Forest Dynamics Plot, Cameroon. In Losos E.C. and Leigh, E.G. Jr. eds. Forest Diversity and Dynamism: Findings from a Large-Scale Plot Network, pp. 506-516. University of Chicago Press, Chicago.

Chuyong, G. B. et al. 2011. Habitat specificity and diversity of tree species in an African wet tropical forest. – Plant Ecol. 213: 1363–1374.

Condit, R. et al. 2004. Tropical forest dynamics across a rainfall gradient and the impact of an El Niño dry season. – J. Trop. Ecol. 20: 51–72.

Condit, R. et al. 2012. Barro Colorado Forest Census Plot Data, 2012 Version. Center for Tropical Forest Science Databases.

Diniz-Filho J. A. F. and Bini L. M. 2005. Modelling geographical patterns in species richness using eigenvector-based spatial filters. – Global Ecol. Biogeogr. 14: 177–185.

Dray, S. et al. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). – Ecol. Model. 196: 483–493.

Dray, S. et al. 2018. adespatial: Multivariate Multiscale Spatial Analysis. R package version 0.2-0.

Dutilleul, P. 1993. Modifying the t test for assessing the correlation between two spatial processes. – Biometrics 49: 305–314.

Fortin and Dale 2005. Spatial Analysis: A Guide for Ecologists. Cambridge Unversity Press, Cambridge, UK.

Fox, J. and Weisberg, S. 2011. car: An R Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage.

Griffith D. and Peres-Neto P. 2006. Spatial modeling in Ecology: the flexibility of eigenfunction spatial analyses. – Ecology 87:2603 –2613.

Harms, K. E. et al. 2001. Habitat association of trees and shrubs in a 50-ha neotropical forest plot. – J. Ecol. 89: 947–959.

Hope, A.C.A. 1968. A simplified Monte Carlo significance test procedure. J. Roy. Statist. Soc. Ser. B 30: 582-598.

Itoh, A. et al. 2010. Comparison of statistical tests for habitat associations in tropical forests: a case study of sympatric dipterocarp trees in a Bornean forest. – Forest Ecol. Manag. 259: 323–332.

John, R. et al. 2007. Soil nutrients influence spatial distributions of tropical tree species. – P. Natl. Acad. Sci. USA. 104: 864–869.

Kadavul, K. and Parthasarathy, N. 1999. Plant biodiversity and conservation of tropical semi-evergreen forest in the Shervarayan hills of Eastern Ghats, India. – Biodivers. Conserv. 8: 419–437.

Lan, G. et al. 2011. Topography-related spatial distribution of dominant tree species in a tropical seasonal rain forest in China. – Forest Ecol. Manag. 262: 1507–1513.

Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? – Ecology 74: 1659–1673.

Legendre, P. et al. 2002. The consequences of spatial structure for the design and analysis of ecological field surveys. – Ecography 25: 601–615.

Legendre, P. et al. 2009. Partitioning beta diversity in a subtropical broad-leaved forest of China. – Ecology 90: 663–674.

Legendre, P. and Legendre, L. 2012. Numerical Ecology, 3rd English edn. Elsevier Science BV, Amsterdam.

McIntire E. J. B. and Fajardo A. 2009. Beyond description: the active and effective way to infer processes from spatial patterns. – Ecology 90: 46–56.

Moran, P. A. P. 1953. The statistical analysis of the Canadian lynx cycle. II. Synchronization and meteorology. – Aust. J. Zool. 1: 291–298.

Muledi J. I. et al. 2017. Fine-scale habitats influence tree species assemblage in a Miombo Forest. – J. Plant Ecol. 10: 958–959.

Noguchi, H. et al. 2007. Habitat divergence in sympatric Fagaceae tree species of a tropical montane forest in northern Thailand. – J. Trop. Ecol. 23: 549–558.

Oksanen, J. et al. 2015. vegan: Community Ecology Package. R package version 2.2-1.

Peres-Neto, P. R. et al. 2006. Variation partitioning of species data matrices: estimation and comparison of fractions. – Ecology 87: 2614–2625.

Peres-Neto, P. R. and Legendre, P. 2010. Estimating and controlling for spatial structure in the study of ecological communities. – Global Ecol. Biogeogr. 2: 174–184.

Plotkin, J. B. et al. 2000. Species-area Curves, Spatial Aggregation, and Habitat Specialization in Tropical Forests. – J. Theor. Biol. 207: 81–99.

Potts, M. D. et al. 2004. Habitat heterogeneity and niche structure of trees in two tropical rain forests. – Oecologia 139: 446–453.

Prada, C. M. and Stevenson, P. R. 2016. Plant composition associated with environmental gradients in tropical montane forests. – Biotropica 48(5): 568–576.

Punchi-Manage, R. et al. 2014. Effect of spatial processes and topography on structuring species assemblages in a Sri Lankan dipterocarp forest. – Ecology 95(2): 376–386.

R Development Core Team 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org

Ray-Mukherjee, J. et al. 2014. Using commonality analysis in multiple regressions: A tool to decompose regression effects in the face of multicollinearity. Methods Ecol. Evol. 5(4), 320–328.

Renka, J. 2013. R functions by Albrecht Gebhardt. With contributions from S. Eglen, S. Zuyev and D. White. tripack: Triangulation of irregularly spaced data. R package version 1.3-6. http://CRAN.R-project.org/package=tripack

Rowlingson, B. and Diggle, P. 2015. splancs: Spatial and space-time point pattern analysis. R package version 2.01-37.

Scheiner, S. M. 2011. Musings on the Acropolis: Terminology for biogeography. - Front. Biogeogr. 3: 62–70.

Seidler, T. G. and Plotkin, J. B. 2006. Seed dispersal and spatial pattern in tropical trees. – Plos Biol. 4: 2132–2137.

Upton, G.J.G. and Fingleton, B. 1985. Spatial data analysis by example. Vol. 1: Point pattern and quantitative data. Wiley, NewYork.

Vleminckx, J. et al. 2014. Soil charcoal to assess the impacts of past human disturbances on tropical forests. – PLoS ONE, 9, e108121.

Vleminckx, J. et al. 2015. Impact of fine-scale edaphic heterogeneity on tree species assembly in a central African rainforest. – J. Veg. Sci. 26: 134–144.

Vleminckx, J. et al. 2017. The influence of spatially structured soil properties on tree community assemblages at a landscape scale in the tropical forests of southern Cameroon. – J. Ecol. 105: 354–366.

Wagner, H. H. 2003. Spatial covariance in plant communities: integrating ordination, variogram modelling, and variance testing. – Ecology 84: 1045–1057.

Wagner, H. H. 2004. Direct multi-scale ordination with canonical corespondence analysis. – Ecology 85: 342–351.

Wagner, H. H. and Dray, S. 2015. Generating spatially constrained null models for irregularly spaced data using Moran spectral randomization methods. – Methods Ecol. Evol. 6: 1169–1178.

Zent, E. L. and Zent, S. 2004. Floristic composition, structure, and diversity of four forest plots in the Sierra Maigualida, Venezuelan Guayana. – Biodivers. Conserv. 13: 2453–2483.

Chapitre VI

Combining complementary spatial methods to address community assembly processes: A study case from a temperate forest

David Bauman, Roxane Beyns, Thomas Drouet

Unpublished

Le Chapitre V a montré qu'une nouvelle procédure de test de la fraction spatiale et environnementale conjointe de la partition de variation présentait un risque de faux positif très faible et une puissance statistique élevée. Il a également montré que la puissance du test pouvait varier en fonction du rapport de l'échelle du pattern spatial concerné avec la proximité de l'échelle de la résolution spatiale minimale du design d'échantillonnage.

Dans le Chapitre VI, cette méthode de test d'un processus lié à l'environnement et spatialement structurant, de même que les avancées méthodologiques des Chapitres III et IV, seront intégrées dans un cas d'étude visant à déterminer les processus d'assemblage d'une communauté d'arbres de forêt tempérée échantillonnée dans le cadre de la thèse. Ce chapitre illustrera également l'intérêt de l'utilisation parallèle d'approches spatiales complémentaires.

Combining complementary spatial methods to address community assembly processes: A study case from a temperate forest

Abstract

Combining complementary advanced statistical methods on real datasets is essential to understand the mechanisms of plant community assembly. The stabilising niche differences and relative fitness differences concepts of the contemporary coexistence theory has been shown to be a promising framework to reach key insights into plant community assembly mechanisms. However, precise spatially-explicit community approaches are needed to achieve this goal. Here, we propose an original and powerful multiple approach to understand the mechanisms of assembly of a tree community located in a temperate forest of Western Europe (Belgium) by combining two complementary spatial analyses. First, species abundances were analysed through a variation partitioning combining a high number of soil, light, and topographical variables to advanced spatial predictors (Moran's eigenvector maps) allowing the detection of coarse to complex spatial structures, ranging from 416 to 12.5 m. Second, a bivariate spatial point pattern analysis was used to characterise the interspecific spatial association of all species pairs in the immediate vicinity of individuals, hence considering a complementary scale to that of the variation partitioning (< 12.5 m). The correlations of key leaf functional traits among the significantly-associated species pairs were tested with Mantel tests to assess the prevalence of stabilising niche difference or relative fitness differences. The community composition displayed several contrasting spatial patterns; some patterns were highly related to chemical and physical soil variables, and to elevation, which acted either as stabilising niche differences or translated relative fitness differences. Other patterns were explained by the density of individuals (i.e. competition intensity), and some patterns remained unexplained. A high proportion of the species pairs displayed positive or negative patterns of association within a neighbourhood of 3 and 10 m radius. At 3 m radius, the interspecific association pattern (24% of the species pairs) was negatively related to the leaf area similarity, suggesting the differences in this trait acted as a stabilising niche difference. At 10 m radius, the interspecific spatial association pattern was much stronger (71% of the species pairs), but could not be explained by the functional traits measured. This study illustrated how combining eigenvector-based methods and spatial point pattern analysis can provide deeper and complementary insights into the mechanisms of plant community assembly in a spatially-explicit manner.

Keywords

community ecology, contemporary coexistence theory, multiple complementary methods, bivariate spatial point pattern analysis (SPPA), habitat filtering, functional traits, Moran's eigenvector maps (MEM), relative fitness differences, soil, stabilising niche differences, variation partitioning

I. Introduction

Community assembly is most often considered through the lens of the multiple filters metaphor (Kraft et al. 2015). A local community is formed from a regional pool of species that must pass a first filter of chance and dispersal to get to the space of the local community. Then an abiotic and a biotic filter (i.e. habitat filtering, when considered together) determine the species possessing a set of functional traits suitable enough to establish and persist in the local community (HilleRisLambers et al. 2012).

However, the species themselves can also have an impact on both their abiotic and their biotic environments (Chesson 2000, Lavorel and Garnier 2002, Chase and Leibold 2003, Suding et al. 2008), which is sometimes not taken into account by the filter metaphor (HilleRisLambers et al. 2012). Contemporary coexistence theory has recently been proposed as a more nuanced framework to address the mechanisms of community assembly by emphasising the roles of stabilising niche and relative fitness differences (HilleRisLambers et al. 2012). Stabilising niche differences are differences that cause a species to suppress itself more than it suppresses another species it competes with in a negative frequency-dependent manner (Chesson 2000). This effect can arise among species from differences in their impact on or response to limiting factors such as ressources, consummers, mutualistic organisms, or to host-specific enemies. It fosters coexistence and diversity in the community. Relative fitness differences, on the other hand, are differences between species that predict the outcome of negative interactions (i.e. competition) in the absence of stabilising mechanisms. They are the consequence of a different environmentally-driven fecundity, or different abilities to pre-empt a resource or tolerate herbivory or pathogens, and result in the competitive exclusion of one or several species (Chesson 2000, HilleRisLambers et al. 2012).

The detection of such coexistence mechanisms on plant community assembly requires the use of advanced spatial analyses. Indeed, the ecological processes driving community assembly are intrinsically spatial (Legendre and Legendre 2012), and usually generate spatially-structured signals at specific scales in communities (Borcard et al. 2018). Therefore, spatially-explicit approaches are needed when aiming at fine insights into these processes. The latter can be studied through the community composition (e.g. species' abundances; e.g. Dray et al. 2012, Bauman et al. 2016, Vleminckx et al. 2017) or through functional traits (e.g. Cornwell and Ackerly 2009, Blonder et al. 2017). While the former way has the advantage to reflect the effects of both stabilising niche differences and relative fitness differences in the way these are affected by the abiotic conditions and species interactions during community assembly (HilleRisLambers et al. 2012), studying traits provides a means to directly study the relations between function and process (e.g. Pavoine et al. 2011, Malhi et al. 2017).Spatial eigenvector-based methods, such as Moran's eigenvector maps (MEM; Dray et al. 2006) – the generalization of the principal coordinates of neighbour matrices (PCNM) (Borcard and Legendre 2002) –, are flexible, powerful, and accurate statistical tools to highlight coarse to complex multiscale spatial patterns in communities (Griffith and Peres-Neto 2006, Bauman et al. 2018a, 2018b). When used together with a set of well-chosen relevant environmental variables in a variation partitioning framework (hereafter, VP), MEM can help modelling the total variation of a community composition through additive pure and shared effects of the environmental and the spatial variables (see Fig. 2). If backed-up by solid hypotheses, these results can help inferring ecological processes underlying the explanatory fractions of the VP (McIntire and Fajardo 2009). The shared space-environment fraction has recently been shown to successfully model induced spatial dependences (Bauman et al. 2018c), that is, the reflection of a spatially-structured environmental variable in the community (Legendre and Legendre 2012). The latter can result either from stabilising effects or from a relative fitness difference acting differently in the different spatially-structured patches of environmental values. On the other hand, a pure spatial signal indicates that some spatial patterns remain in the community after all effects of the environment has been removed. Such pure spatial patterns are generally expected to either reflect the signature of an unmeasured influent environmental variable, or to be caused by limited dispersion (Legendre and Legendre 2012). Mapping these residual patterns can help generating hypotheses about potential underlying processes (McIntire and Fajardo 2009).

Yet, this analytical approach based on species abundances in quadrats (hereafter, quadrat-based approach) has its limitations. For instance, even for the best sampling design configuration (i.e. contiguous sampled quadrats), any spatial pattern occurring below the grain of the study (i.e. the dimension of a quadrat) will remain undetectable, and so will the processes influencing community assembly in the inmediate vicinity of individuals. This approach can thus be viewed as “hyperopic”.

The ecological processes responsible for community assembly have also widely been studied through spatial point pattern analysis (SPPA), an individual-based statistical approach consisting in studying patterns of ecological objects (here, tree individuals, assimilated to points) comprised in an observation window (Wiegand and Moloney 2014, Velázquez et al. 2016). Unlike the quadrat-based approaches, SPPA therefore studies the unique location of individuals, to which qualitative marks – such as species name – or information of environmental variables can be added. Patterns of spatial association among points can be characterised at a range of distances by different summary statistics, such as the *K*-function of Ripley (1981). Null distributions of points can then be generated on the basis of different parameters estimated from the true point pattern, and using theoretical point processes mimicking different ecological processes. The calculation of the summary statistics associated to such simulated point patterns allows building simulation enveloppes of the statistics distribution under the null hypothesis that the real point pattern has been generated by one or several processes. This individual-based approach has proved useful to detect potential signatures of different types of ecological processes in plant communities (e.g. Wiegand et al. 2007, Wang et al. 2010, Lin et al. 2011, Velázquez et al. 2015). Yet, the SPPA also has its limitations, notably to detect and infer processes at broad spatial scales as the estimator of the *K*-function has increased bias as the distance range at which the point pattern is characterized increases (Baddeley and Turner 2005). This second spatially-explicit approach can therefore – in comparison with the quadrat-based approach – be viewed as “myopic”.

Using complementary data analysis approaches has been shown to be crucial when assessing the role of different ecological processes (De’ath 2002, Meynard et al. 2013, Vellend et al. 2014, Aiello-Lammens et al. 2017, D’Amen et al. 2017). Therefore, the combination of the intrinsic “hyperopic” and "myopic" characteristics of the quadrat-based and individual-based approaches, respectively, is expected to offer a much more complete spatially-explicit insight of the processes responsible for plant community assembly than solely considering one of them.

In this study, we illustrated an original and powerful analytical approach on the tree community of a European temperate forest to assess and highlight the role of different coexistence mechanisms as driver of the community assembly. To do so, we first performed a variation partitioning (VP) of the community composition using spatial eigenvectors and a high number of ecologically-relevant environmental variables to uncover the processes influencing the tree community at spatial scales ranging from 12.5 to ~400 m. Then, a bivariate SPPA approach of all species pairs combined with functional trait data to evaluate potential stabilising niche and relative fitness differences in the direct vicinity of individuals (i.e. < 10 m). This multiple approach allowed providing a spatially and ecologically complementary view of the processes responsible for the community assembly.

II. Material and methods

II.1. Study site

The study was carried out in a temperate forest of 2.37 ha located on the north hillside of a calcareous hill (Le Moessia) in Treignes, Viroinval (Belgium) (50°5'31'' - 50°5'37'' N, 4°40'27'' - 4°40'47'' E). The site is located on the occidental portion of the Calestienne, a geomorphological landscape caracterised by a succession of calcareous hills and schistose valleys. Soils have developed on a calcareous bedrock dated from Middle Devonian. The mean annual temperature and rainfall are 8.6 °C and 781 mm, respectively. The topography is characterised by both low and steep slopes (9 to 52°) and an altitude range of 33 m (161 to 194 m a.s.l.) (see supplementary Table S2). The traditional extensive sheep-grazing practice in this region kept the site unforested until ~1960, when this practice was abandoned. A natural forest recolonisation of the site occurred then, and the forest was left untouched, most likely as a consequence of its difficult access and steep slopes. A permanent plot was established in 2014 to study temperate forest community assembly, functions, and dynamics.

II.2. Tree community census

The forest plot was subdivided into 143 quadrats of 12.5 × 12.5 m within which all trees and lianas above 1.3 m high were mapped, identified to the species, and measured for diameter at breast height (1.3 m; dbh) between October 2014 and September 2015 (see Fig. S1). The mapping was performed using a Differential Global Positioning System (Trimble® GPS Pathfinder® ProXRT receiver) with a post-treatment correction based on RINEX files from four ground-based reference stations (Onhaye, Mariembourg, Tellin, Charleroi) of the Walcors network, allowing an average horizontal and vertical precision of 0.9 ± 0.6 m and 1.1 ± 0.7 m, respectively. A degree of spatial accuracy < 0.5 m, < 1 m, and < 2 m was achieved for 31%, 72%, and 96% of the individuals. The dbh of multiple stem trees was computed as $2 \times \sqrt{(\sum r^2)}$, where r was the radius. Table S1 describes the composition and structure of the living tree community (supporting information). The tree community is composed of 9059 individuals distributed among 31 species, belonging to 16 and 13 families and orders, respectively. Dead individuals (1304 occurrences; Fig. 1a) will not be considered in this study, so that further uses of the term “community” will refer to the 7755 living individuals (Fig. 1b, c). The community is dominated in abundance by *Crataegus monogyna* (Rosaceae; 31% of the total abundance), *Corylus avellana* (Betulaceae; 20%), *Prunus spinosa* (Rosaceae; 9%), *Quercus robur* (Fagaceae; 9%), and *Cornus sanguinea* (Cornaceae; 6%) (Fig. 1b), while the species displaying the five highest basal areas are *Q. robur* (32% of the total basal area), *Pinus sylvestris* (Pinaceae; 20%), *Betula pendula* (Betulaceae; 15%), *C. monogyna* (9%), and *Carpinus betulus* (Betulaceae; 7%) (Fig. 1c; see contrasting colour dominances between Fig. 1b and c, when considering basal area explicitly).

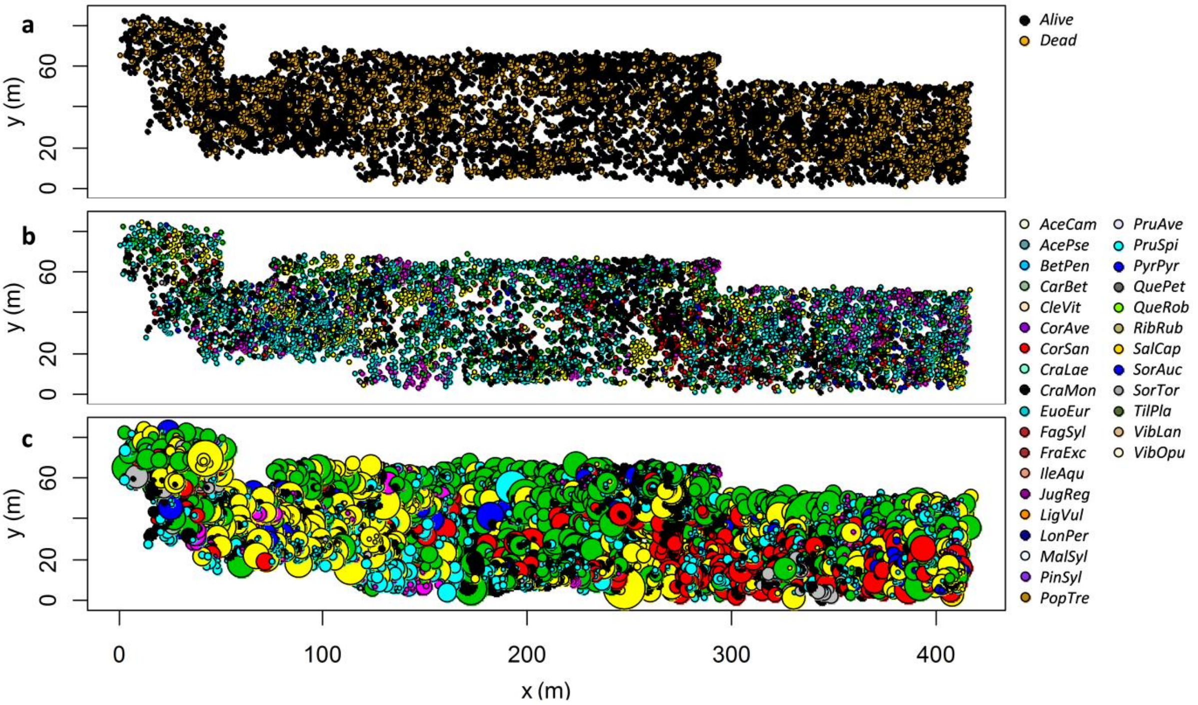


Figure 1: Mapped individuals of the Moesia forest (Treignes, Belgium). a: Living and dead individuals (black and gold circles, respectively); 9059 individuals. b, c: Colour-representation of the distribution of the 31 species (living individuals only); 7755 individuals. c: Circle sizes are proportional to the individuals' dbh. The colour dominance contrast between maps b and c highlights how different species appear to be dominant according to the abundance (b) or the dbh (c). The legend gives the acronym of the woody species (see Table S1 for corresponding full names).

The density and structure of the quadrats varied strongly within the plot, with a minimum and maximum density of 1024 and 8256 individuals per hectare, and some quadrats composed of a high number of small individuals while other were composed of a smaller number of bigger specimens (see varying point density on Fig. 1b, and unequal distribution of small and big individuals on Fig. 1c). In this study, we only considered the 16 species represented by a minimum of 40 individuals (see bold species in Table S1) to avoid biases related to underrepresented species. These 16 species account for 98.5% and 99.3% of all individuals and of the global basal area of the forest, respectively (bold species in Table S1).

II.3. Environmental variables: soil, light, and topography

Five cores (20-cm depth) were sampled (one at the centre of the quadrat and four on the two diagonals at mid- distance from the centre) within all 143 quadrats (see Fig. S1). All sampling as well as the measurement of 24 soil variables were conducted according to conventional protocols (Pansu and Gautheyrou 2007), briefly described in Appendix S1: Section 1. Measured soil variables were a stoniness index, soil texture (clay, silt, and sand), the red, green, and blue components of the soil colour ($col_{Red,\ Green,\ Blue}$), a carbonate index, pH-H_2O, electrical conductivity (EC), organic matter content (OM), effective cation exchange capacity (CEC), nitrate (NO_3^-), ammonium (NH_4^+), and total inorganic nitrogen (N_{total}) contents, and plant-available element contents (P, Ca, Mg, K, Al, Fe, Mn, Zn, Ba). The canopy cover (CaCo) and canopy openness – two accurate and robust complementary light intensity indices – were computed from hemispheric photographs of the canopy taken from the center of each quadrat, using the Gap Light Analysis Mobile Application software (Tichý 2016). We considered height topographical variables covering all four types of terrain information (i.e. slope, orientation, curvature, and terrain variability; Wilson et al. 2007): altitude, slope, aspect (i.e. slope orientation), convexity, terrain ruggedness index (TRI), topographic position index (TPI; Weiss 2001), roughness, and flow direction.

Soil physical and chemical properties are highly heterogeneous, despite the small extent of the site (e.g. range of pH_{H2O} of 6.1 to 8.3, clay proportion of 9 to 45%, and nitrate content of 6 to 131 µg/g, see details in supplementary Table S2). Light availability also presents a strong variability (CaCo between 42 and 98%). Supplementary Fig. S2 illustrates the relations among environmental variables on the first two axes of a principal component analysis, while supplementary Table S3 presents Pearson correlations among the variables.

II.4. Functional traits

Fresh mature leaves from 10 individuals located in contrasted zones of the plot were collected for all species. They were scanned and the leaf area (LA) was computed using the Image J software. Then, the leaves were weighted after they were oven-dried at 65° C for two days, and the specific leaf area (SLA) was computed as the ratio between the leaf area and the leaf dry mass, following Pérez-Harguindeguy et al. (2013). The LA and SLA were used as proxy of the water use/leaf energy, and resource use strategies, respectively (Wright et al. 2004, Díaz et al. 2015), in the bivariate spatial point pattern analysis (see *Data analysis*).

II.5. Data analysis

II.5.1. Quadrat-based approach

II.5.1.1. Environmental variable selection

The environmental variables were Box-Cox transformed to bring them closer to a normal distribution and stabilise variances. A two-step procedure was used to select the most suited subset of environmental variables to be integrated in the VP. First, we excluded one variable of each pair of variables displaying a Pearson correlation ≥ 0.7 (see Table S2) in order to avoid collinearity issues

(Dormann et al. 2013). Then, a genetic algorithm developed by Calcagno and Mazancourt (2010) (package *glmulti*) (i.e. an Information Criterion-based selection approach) was adopted to select the final subset of environmental variables (**E**) (see details in Appendix S1: Section 2).

II.5.1.2. Spatial predictors

Spatial predictors were generated to capture the spatial patterns of organization of the community at multiple scales. The latter were generated using Moran's eigenvector maps (MEM) (Dray et al. 2006). MEM consist in diagonalising a doubly-centered spatial weighting matrix (SWM) built from the spatial coordinates of the sampled units, and in generating the eigenvectors that maximise the Moran's index of spatial autocorrelation (see details in Dray et al. 2006). The spatial eigenvectors (or MEM variables) can then directly be used as spatial predictors in regression or canonical analyses to highlight coarse or complex patterns at a broad range of spatial scales. We optimised the selection of the SWM and of a subset of spatial predictors following Bauman et al. (2018a, 2018b) (function *listw.select* of the package adespatial; Dray et al. 2018) (see details and parameters in Appendix S1: Section 3). The optimised subset had 30 spatial predictors and corresponded to a SWM built from a Gabriel's graph with unweighted links. The spatial predictors were subdivided on a visual basis into a set of broad-scale and a set of fine-scale MEM variables (MEM variable 1 to 18, and 19 to 30, respectively). No negative spatial autocorrelation was detected in the community data (i.e. non-significant global tests of the MEM variables associated to negative eigenvalues), so that all the selected MEM variables modelled positively autocorrelated patterns. The selected SWM allowed the detection of spatial patterns ranging between 12.5 (i.e. the grain of the study) and ~400 m (i.e. the extent of the study).

II.5.1.3. Variation partitioning

The community data (i.e. species' abundances) was Hellinger-transformed prior to performing the VP, following Legendre and Gallagher (2001). Then, we performed a VP of the Hellinger-transformed community matrix (**Y**) into three components (i.e. matrices): the selected environmental variables (**E**), the broad-scale MEM variables ($\mathbf{MEM_B}$), and the fine-scale MEM variables ($\mathbf{MEM_F}$) (Fig. 2). The VP consisted in a serie of complete and partial redundancy analyses (RDA; Rao 1964) (see right portion of Fig. 2). The resulting canonical coefficients of determination (R^2) were adjusted using Peres-Neto et al.'s correction (2006) (hereafter, R^2). Besides the global environmental model ([adfg] on Fig. 2), the adjusted R^2 resulting from an actual model were tested by permutation of the model residuals (Anderson and Legendre 1999). The global environmental model as well as the shared space-environment fractions (SSEF) related to broad, fine, and both scales (i.e. [d], [f], and [g], respectively, in Fig. 2) were tested using Moran spectral randomisations (MSR; Wagner and Dray 2015), following (Bauman et al. 2018c; see details in Appendix S1: Section 4). These fractions are critical as they are expected to be the ones revealing potential induced spatial dependencies of **E** on **Y**.

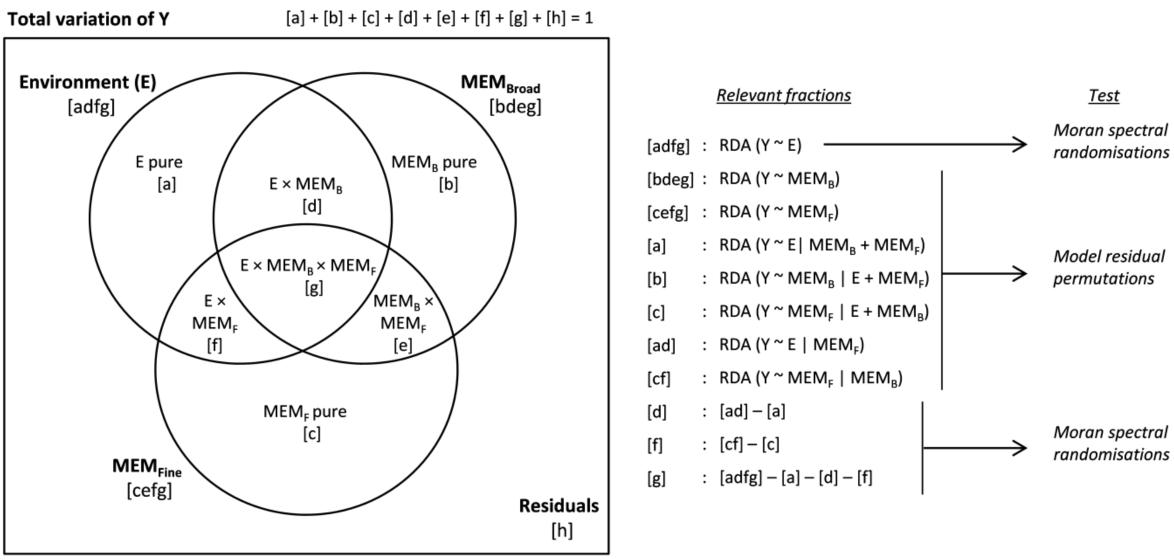

Total variation of Y
[a] + [b] + [c] + [d] + [e] + [f] + [g] + [h] = 1
Environment (E)
[adfg]
MEM$_{Broad}$
[bdeg]
E pure
[a]
E × MEM$_B$
[d]
MEM$_B$ pure
[b]
E × MEM$_B$ × MEM$_F$
[g]
E × MEM$_F$
[f]
MEM$_B$ × MEM$_F$
[e]
MEM$_F$ pure
[c]
MEM$_{Fine}$
[cefg]
Residuals
[h]
Relevant fractions
Test
[adfg] : RDA (Y ~ E)
Moran spectral randomisations
[bdeg] : RDA (Y ~ MEM$_B$)
[cefg] : RDA (Y ~ MEM$_F$)
[a] : RDA (Y ~ E| MEM$_B$ + MEM$_F$)
[b] : RDA (Y ~ MEM$_B$ | E + MEM$_F$)
[c] : RDA (Y ~ MEM$_F$ | E + MEM$_B$)
[ad] : RDA (Y ~ E | MEM$_F$)
[cf] : RDA (Y ~ MEM$_F$ | MEM$_B$)
Model residual permutations
[d] : [ad] – [a]
[f] : [cf] – [c]
[g] : [adfg] – [a] – [d] – [f]
Moran spectral randomisations

Figure 2: Variation partitioning of the Hellinger-transformed community data (Y) between the environmental (E), broad-scale MEM (MEM$_B$), and fine-scale MEM (MEM$_F$) variables. The fractions of the VP (square brackets) represent the adjusted *R*^{2} values (Peres-Neto et al. 2006). The right portion of the figure details the main complete and partial-RDAs as well as the fractions subtractions performed to obtain the corresponding R^2 values. The explanatory components after | are conditions (partialled out from Y in the analysis). The fractions are tested either by the classical procedure of model residual permutation or using Moran spectral randomisations, following Bauman et al. (2018c; see details in Appendix S1: Section 4).

To quantify the spatial scale of the patterns of **Y** related, or unrelated, to patterns in **X** ([d], [f], [g], and [b], [c], respectively), we used the spatial range resulting from the fit of spherical or Gaussian semivariogram models to the empirical semivariograms of the pattern. To uncover the precise relevant environmental variables behind the shared space$_{Broad}$-environment fraction (SSEF$_B$) detected (see *Results*), we proceeded as follows. First, the constrained axes of the global RDA of **Y** against **X** were tested by 999 permutations, providing a set of significant constrained axes. A second RDA of this set of axes against **MEM$_B$** while controlling the effect of **MEM$_F$** was performed. The significance of the constrained axes of this partial-RDA was tested by permutations, and a multiple linear regression model of each significant axis was conducted against **X** to identify the environmental variables significantly related to the spatial patterns of SSEF$_B$.

The VP combined with MEM allows mapping the spatial patterns that remain in **Y** after the effect of **E** was removed. These mapped residual patterns can then theoretically serve as a basis to formulate hypotheses about possible underlying processes (McIntire and Fajardo 2009). While these patterns generally remain unexplained, we further explored the potential role of biotic processes and dispersal limitation as underlying processes. To explain the significant patterns of **Y** unrelated to **E** ([b] and [c], on Fig. 2) (see *Results*), we first used two biotic explanatory variables: (1) local density (i.e. the number of individuals per quadrat), and (2) the total basal area of the quadrat. These variables were used as proxies of the competition intensity.

II.5.2. Individual-based approach

II.5.2.1. Unexplained spatial patterns of the VP and dispersal limitation

Besides the three biotic variables, we used a univariate spatial point pattern analysis (SPPA) approach to compare the spatial scale of the residual patterns in the VP to that of the estimated dispersal limitation range of each species. This approach was used to evaluate whether dispersal was likely to be an ecological process underlying some of the unexplained patterns. To do so, we fitted an inhomogeneous Thomas point process (i.e. a Cox process) to the observed pattern of each species (Baddeley et al. 2015; see details in Appendix S1: Section 5). From the fitted point process of each species, we used the estimated standard deviation of random displacement of a point from the centre of its cluster, multiplied by two, as the estimated scale of a dispersal-limited cluster.

II.5.2.2. Interspecific spatial association patterns

In order to investigate potential stabilising niche differences or relative fitness differences at very local scales (< 12.5 m), we tested for overdispersion or underdispersion of key functional traits among pairs of species significantly closer or further apart from one another than expected by chance. The underlying hypotheses are that a stabilising niche difference would cause geographically closer species to display dissimilar trait values (i.e. trait divergence), while relative fitness differences could generate the same signal or rather only allow species with similar traits to co-occur (i.e. trait convergence) (HilleRisLambers et al. 2012, Kraft et al. 2015). Since the VP only allowed exploring indices of such coexistence mechanisms beyond 12.5 m, we used a bivariate SPPA approach to focus on two local scales (i.e. 3 and 10 m), hence providing a complementary spatially-explicit approach of the processes acting on the community assembly.

The bivariate analysis followed the approach of Wiegand et al. (2007, 2012) to highlight interspecific spatial association patterns. First, the interspecific association patterns of all pairs of species i and j were quantified at the two scales of interest using two bivariate cumulative summary statistics: the *K*-function $K_{ij}(r)$ and the nearest-neighbour distribution function $D_{ij}(r)$. While $K_{ij}(r)$ provides a measure of how individuals of species j are distributed around the individuals of species i in a neighbourhood of radius r, $D_{ij}(r)$ gives the probability that individuals of species i have no neighbours of species j in a radius r. Global envelopes of simulations were generated by 199 permutations for each species pairs to create null distributions of the summary statistics under the hypothesis that the species were distributed independently from one another (Myllymäki et al. 2017). The pairs whose $K_{ij}(r)$ and/or $D_{ij}(r)$ fell outside of the global envelopes at the distance of interest were considered significant. For those pairs, we then centered the observed $K_{ij}(r)$ and $D_{ij}(r)$ by subtracting their expected values under the null model (see details in Velázquez et al. 2015), and log-transformed $K_{ij}(r)$ so that both summary statistics would departure from the null model in the same manner (Wiegand et al. 2007a, 2012, Wiegand and Moloney 2014). The significant pairs of species were then positioned on two axes consisting of the two centered and log-transformed summary statistics (Fig. 6):

$$\hat{P}(r) = \hat{D}_{ij}(r) - (1 - \exp\,(-\lambda_2 \pi r^2))$$
$$\hat{M}(r) = \ln\left(\hat{K}_{ij}(r)\right) - \ln\,(\pi r^2)$$

where the hat symbol corresponds to the observed value of the summary statistics, and i and j are the species of a given pair. Species pairs in the upper-right quadrant of the plot have a positive interspecific spatial association, since individuals of species j occur more often within radius r of species i than expected by chance [$\widehat{M}(r) > 0$], and individuals of species i is surrounded by more indiviuals of species j than expected under the null hypothesis [$\widehat{P}(r) > 0$]. On the contrary, species pairs in the lower-left quadrant display a negative spatial association (i.e. segregation) [$\widehat{M}(r)$ and $\widehat{P}(r) < 0$]. Species pairs in the upper-left quadrant display a partial overlap [$\widehat{M}(r) > 0$ and $\widehat{P}(r) < 0$], in which some individuals of species i have more neighbours of species j than expected under the null model, while others have less. Species pairs in the lower-right quadrant display the opposite partial overlap situations, but tend to be rare (Wiegand et al. 2007a, 2012, Velázquez et al. 2015).

We then calculated an index of "interspecific spatial association" (hereafter, A_{ij}) for the two distances of interest separately, as the coordinates of the significant species pairs on the first axis of a PCA of $\widehat{P}(r)$ and $\widehat{M}(r)$ (Velázquez et al. 2015). The species pairs displaying a significant positive and negative spatial association correspond to the extreme positive and negative values of A_{ij}, while the species pairs of the two partial overlap quadrants correspond to the intermediate values.

Next, a functional similarity index was computed on the basis of each trait (i.e. LA and SLA) for the species pairs displaying a significant spatial association interaction. The indices were generated through a conversion of the Gower dissimilarity coefficient (Laliberté and Legendre 2010) into a similarity coefficient by subtracting it to 1 (Laliberté et al. 2014).

Finally, we tested whether the A_{ij} was related to the SLA and/or LA similarity matrices using simple Mantel tests based on Pearson's product-moment correlation (Legendre and Legendre 2012). The Mantel correlation r_M corresponds to the strength of the association between spatial interspecific association and trait similarity among the species pairs. A positive r_M would indicate that the trait differences translate into relative fitness differences, while a negative r_M suggest that the trait differences act as stabilising niche differences.

The significance of each Mantel test was tested by 999 permutations at the distances of 3 and 10 m (one-tailed test in the upper or lower tail, according to the sign of the Mantel correlation). We corrected the p-values for multiple tests according to the number of traits (i.e. two) using the Šidak correction (Šidák 1967) to avoid having an inflated type I error rate.

All data analyses were conducted in the R environment (v. 3.4.3) using the packages 'adespatial' (Dray et al. 2018), 'vegan' (Oksanen et al. 2017), 'spdep' (Bivand 2006), 'car' (Fox et al. 2018), 'gstat' (Pebesma 2004) , 'FD' (Laliberté et al. 2014), and 'spatstat' (Baddeley and Turner 2005).

III. Results

III.1. Quadrat-based approach

The community variation was significantly explained by the environment (*R*² = 0.16, *P* = 0.001), and was significantly spatially structured at both broad and fine scales (*R*² = 0.29 and 0.05, respectively, and *P* = 0.001 for both) (Fig. 3).

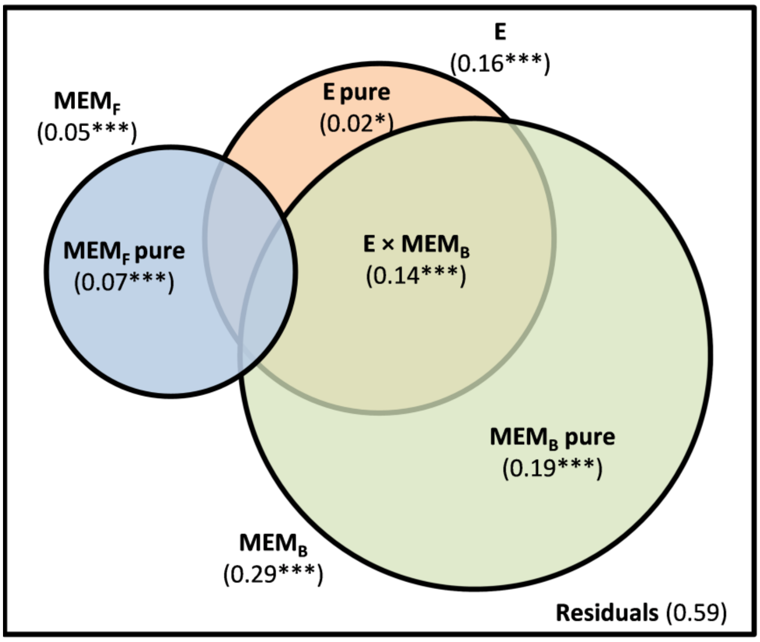


Figure 3: Venn diagram representation of the variation partitioning of the community composition into an environmental (E), broad-scale spatial (MEM_B), and fine-scale spatial components. The adjusted *R*² are presented between brackets. The non-significant fractions are not represented. *P*-values: *: <0.05; **: <0.01; *: <0.001.**

The results indicated that most of the environmental effect on **Y** was spatially structured at broad spatial scales (*R*² = 0.14, *P* = 0.001; **E** × **MEM_B** on Fig. 3). This broad-scale SSEF (hereafter, $SSEF_B$) presented two significant spatial patterns (Fig. 4), both of ~45 m. The first pattern (Fig. 4a) was well explained by the soil texture (sand content), nitrate and Zn content, as well as the elevation (*R*² of the four variables = 0.5), while the second pattern of community composition (Fig. 4b) was highly related to the the stoniness index, the elevation, and the soil NO_3^-, NH_4^+, and sand content (*R*² = 0.46).

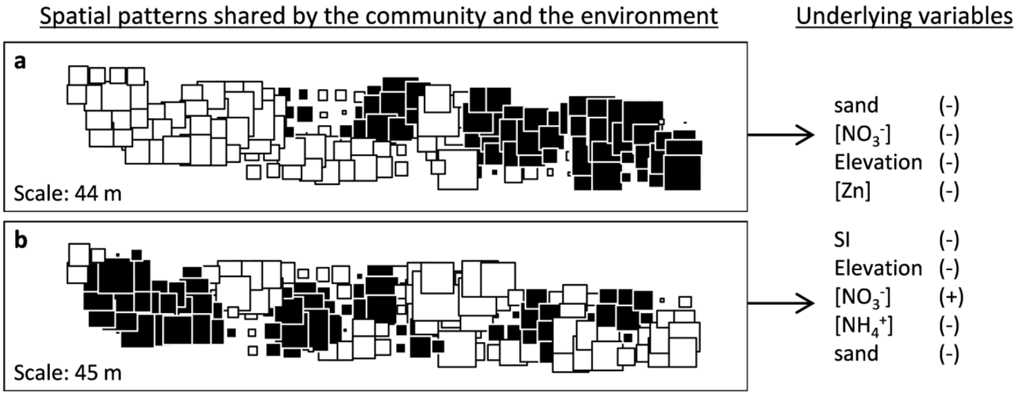


Figure 4: Map of the two significant broad-scale spatial patterns of Y related to spatial patterns of E, and underlying significant environmental variables. A first RDA of Y against E allowed retrieving three significant constrained axes. The partial-RDA of these three axes against MEM$_B$ while controlling for the variation related to MEM$_F$ displayed two significant constrained axes. a and b: Mapped representation of the coordinates of the quadrats on the first and second axes of the partial-RDA, respectively. Black and white squares are positive and negative coordinates, respectively, and square size is proportional to the absolute value of the coordinate. The (+) and (-) are the coefficient signs of the significant environmental variables (multiple linear regression of the RDA axis against E). Scale indicates the spatial scale of the pattern.

The three RDA triplots representing the significant constrained axes of the RDA of **Y** against **E** (not shown) highlighted the species most associated to contrasted values of these environmental variables, hence helping understanding the patterns of Fig. 4. *Corylus avellana* (2nd most abundant species) was mostly present on soils with high sand but low clay contents, high NO_3^-, and at higher elevations, while *Prunus spinosa* and *Crataegus monogyna* (3rd and 1st most abundant species, respectively) preferred soils of high clay content. However, *P. spinosa* differed from *C. monogyna* in that it occurred more at low sand proportions, and at low NO_3^- contents, while *C. monogyna* was more present at low soil NH_4^+ contents and low stoniness index. It also appeared that *Quercus robur* and *Carpinus betulus* (1st and 5th highest basal area species) were more abundant at higher elevations, high concentrations of NH_4, high stoniness index, but low clay contents. At the opposite, the presence of *Pinus sylvestris* (2nd most dominant species in basal area) was explained by a relatively higher clay content, but at lower NH_4^+ and stoniness index values. Finally, *Cornus sanguinea* tended to occure more at low elevation, and at low NH_4^+ and NO_3^- concentrations.

Four significant residual spatial patterns of broad scales (Fig. 5a, b) and fine scales (Fig. 5c, d) were detected, after the effect of the environment was partialled out. These broad and fine-scale pure spatial fractions explained 0.19 (P = 0.001) and 0.07 (P = 0.001) of the total variation in **Y**, and displayed spatial structures occurring at 116 m (Fig. 5a), 62 m (Fig. 5b), and 27 m (Fig. 5c, d). Potential underlying biotic processes were explored by regressing the RDA axes corresponding to these patterns against the density of individuals (i.e. intensity) and the total basal area of the quadrats (i.e. biotic variables). Local intensity was significantly negatively and positively related to two residual patterns that occurred at a scale of 62 and 27 m, respectively (Fig. 5b, d) (R^2 = 0.08, P < 0.001, and R^2 = 0.05, P = 0.01).

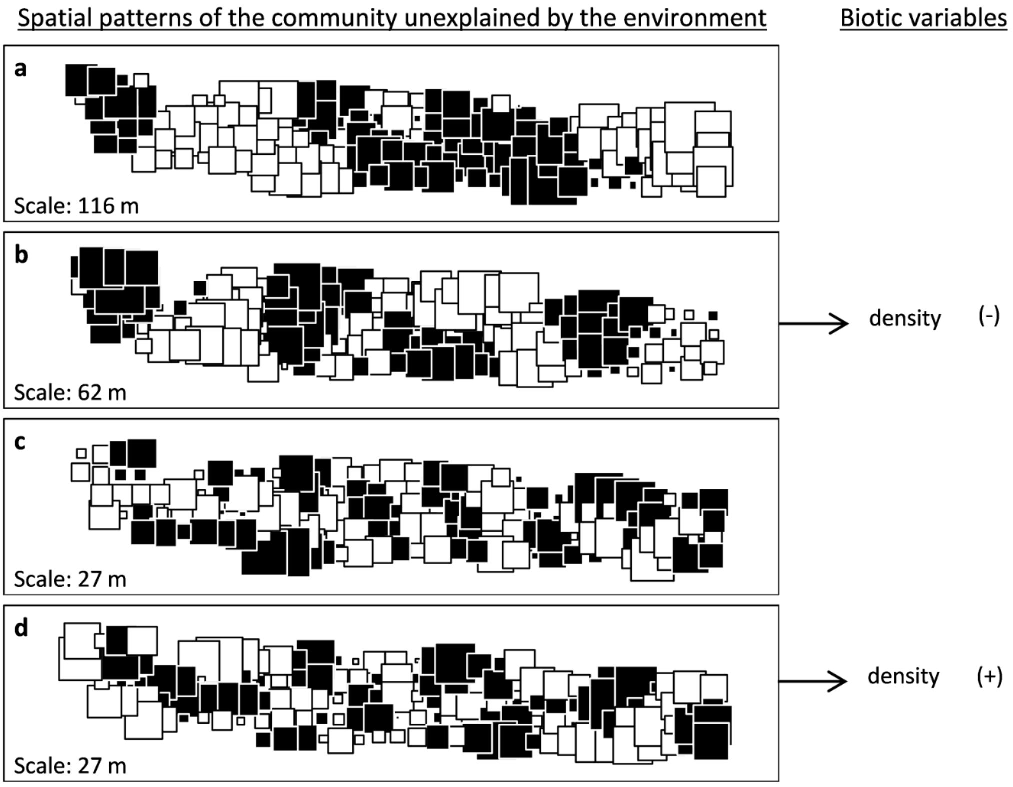


Figure 5: Map of the four significant broad-scale (a, b) and fine-scale (c, d) spatial patterns of Y unexplained by E, and the biotic variables that significantly related to them. 'Scale' indicates the spatial scale of the pattern. See legend of Fig. 4 for the meaning of the colour and size of the squares, as well as for the (+) and (-) signs.

An inhomogeneous Thomas point process was fitted to each species separately in order to retrieve the precise scales of the clusters remaining after removing the effect of the environmental variables and to compare them to those of the residual spatial patterns of the variation partitioning (Fig. 5). This spatial point pattern approach highlighted cluster scales below the scales of the unexplained patterns of the VP, that is, varying between 0.3 and 10 to 17.4 m for most species (see supplementary Table S4).

III.2. Individual-based approach

A total of 62 and 181 species pairs (i.e. 24 and 71% of the total number of pairs) displayed significant spatial association patterns within radius of 3 and 10 m, respectively (Fig. 6 and details in supplementary Table S5). For the two studied ranges, most species pairs displayed a negative association pattern, hence indicating that the presence of the two species within a radius of 3 or 10 m is more unlikely than expected by chance. Yet, seven and twelve species pairs displayed a positive spatial association pattern for the ranges of 3 and 10 m, respectively, meaning that the presence of one of the species makes the presence of the other more likely than expected under the hypothesis of an independent distribution.

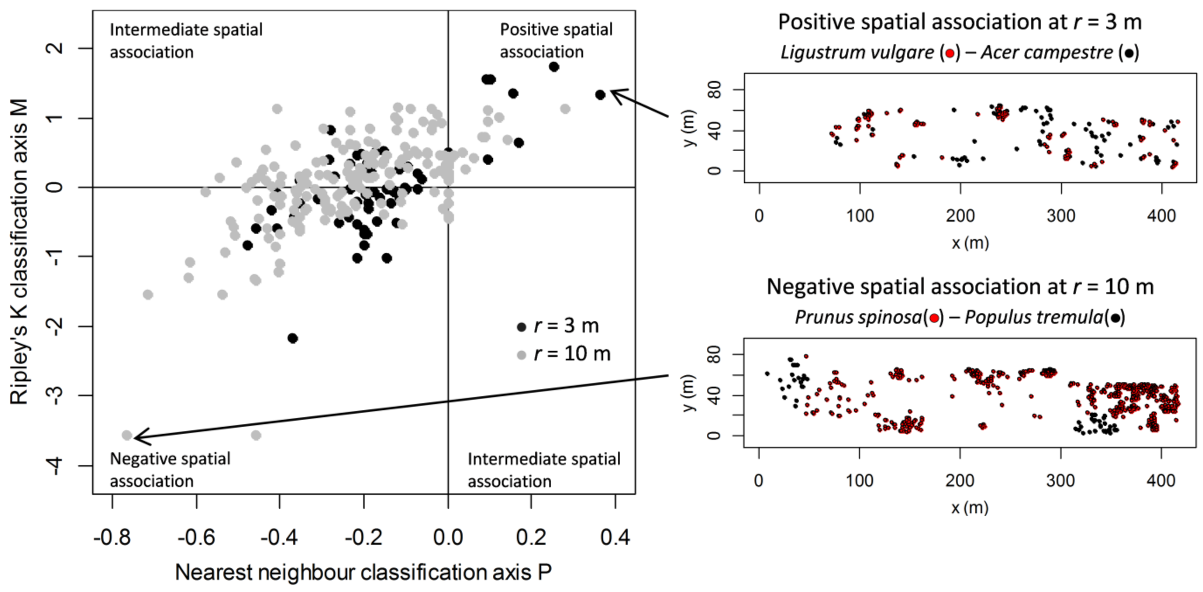


Figure 6: Classification of the interspecific spatial association patterns of the woody species pairs in the forest plot of Treignes (Belgium). Black and grey circles of the left-handed plot are the species pairs that fell out of the global simulation envelopes for at least one of the two summary statistics, at a neighbourhood of 3 and 10 m radius, respectively. The upper and lower right-handed maps illustrate positive and negative interspecific spatial associations, respectively.

The interspecific spatial association index, A_{ij}, captured 84% and 80% of the variability of the axes $\hat{P}(r)$ and $\widehat{M}(r)$ of Fig. 6 at the neighbourhoods of 3 and 10 m radius, respectively. At the spatial range of 3 m, A_{ij} was significantly explained by the LA similarity (corrected *P*-value: 0.033) with a high negative Mantel correlation (-0.68). The correlation with the SLA similarity was not significant (corrected *P*-value: 0.198) but still showed a similar trend (r_M = -0.30) (Table 1). Therefore, within 3 m, the species pairs that significantly occurred together had much more dissimilar LA than expected by chance, and the species possessing similar LA occurred significantly further apart than expected under a model of independent distribution.

Table 1: Results of the simple Mantel tests of the indices of interspecific spatial association against the matrices of trait similarity. *n*: number of species pairs showing a significant spatial association; LA: leaf area; SLA: specific leaf area. Corrected *P*-values were adjusted for two tests (i.e. number of traits) using the Sidak correction.

Range (m)	*n*	Trait	Mantel correlation	uncorrected *P*-value	corrected *P*-value
3	62	LA	-0.68	0.012	0.024
		SLA	-0.30	0.105	0.199
10	181	LA	-0.08	0.368	0.600
		SLA	0.03	0.363	0.594

IV. Discussion

In this study, we used an original analytical approach by combining a partitioning of the woody community variation between environmental and spatial predictors with the assessment of the functional determinants of the bivariate interspecific association patterns. This strategy allowed highlighting (1) complex multiscale patterns of variation in the community composition, either explained or unexplained by the environment, and (2) patterns of interspecific association well explained by functional similarity.

The VP approach allowed explaining 41% of the total variation in the community composition (**Y**). It showed that 16% of the variation in **Y** was explained by the environment (mostly by the different forms of plant-available N, Zn, soil texture, the stoniness index, and elevation). These environmental variables displayed spatial patterns of 45 m that were reflected in the community (Fig. 4), hence indicating that local communities differentiated at this scale over the corresponding environmental variables (induced spatial dependence). Such spatially-structured differentiation of local communities in reponse to heterogeneous environmental conditions may be the results of either stabilising niche differences or local competitive exclusion (HilleRisLambers et al. 2012). Indeed, stabilising niche differences may result from resource partitioning (for the forms of N, or different Zn requirement), or different tolerances to abiotic constraint such as soil stoniness and depth, or a hydric stress positively related to both the proportion of sand and elevation, that is, two of the environmental variables that explained the community composition patterns (Fig. 4). On the other hand, the same patterns related to the environment may also have arisen from varying competitive ability of species in different environment, hence leading locally to the exclusion of species whose fitness is too low in the local environmental conditions. A combination of these two coexistence mechanisms could also be responsible for the patterns, as stabilising niche differences and relative fitness differences have already been shown to often be related in a complex manner (see Chesson 2000, HilleRisLambers et al. 2012).
Manipulative experimental studies, such as *in situ* reciprocal transplant of seedlings (e.g. Cuma et al. 2018), could greatly help disentangling the respective roles of these two opposite processes and determine more precisely the actual community assembly mechanisms (e.g. Fargione et al. 2003, Hooper and Dukes 2010).

Quantitatively, an explanation of 16% of the variation in the community by the environment suggests a weaker structuring importance of the environment than previously highlighted in temperate forests (Gilbert and Lechowicz 2004), which is unexpected given the quantity of relevant ecological variables considered in our study. This may be related to the young age of the forest (~ 70 years), if the effect of abiotic filters has not had time to have a more strongly detectable effect, but it could also indicate an actual dominance of neutral dynamics at the studied extent of the forest.

However, the use of advanced spatial predictors (i.e. MEM) in the VP allowed highlighting complex spatial patterns remaining in **Y** after partialling out the effect of the environment (Fig. 5) and explaining 26% of the variation in **Y**. Two of these patterns (Fig. 5b, d) were significantly related to the density of individuals at scales of 62 and 27 m, hence suggesting a different ability to cope with highly competitive environments among the species, that is, a relative fitness difference in such biotic environment. The density, however, explained only a small portion of these patterns (< 10%), hence suggesting some other underlying process. The residual patterns of Fig. 5a, c, occurring at scales of 116 m and 27 m, could not be explained by any of the biotic variables.

Given the scale of the residuals spatial patterns (i.e. 27, 62, 116 m) and that of the dispersal clusters (i.e. mostly between 0.5 and 17 m) obtained through the inhomogeneous Thomas point process fitted to the point patterns of the species, dispersal limitation seems to be an unlikely ecological process to explain the patterns. However, the actual dimension of fine-scaled spatial patterns close to the minimal spatial resolution (i.e. 12.5 m, here) may be unreliable as a consequence of "aliasing effects" that bias the detection of a pattern towards a broader scale as a consequence of undersampling (Platt and Denman 1975). This bias arises from the fact that the number of quadrats that needs to be considered to detect a spatial pattern may correspond to a much broader scale than the actual scale of the pattern. Aliasing effects were estimated to occur whenever the real pattern is finer than four times the minimal spatial resolution (i.e. 50 m here; see details in Platt and Denman 1975). In the present case, aliasing effects could therefore have distorted the actual scale of the spatial patterns occurring below 50 m, so that the real scale of the residual patterns occurring at 27 m might actually be much finer; hence potentially matching the range of scales of the dispersal clusters (Table S4). A proper exploration of aliasing effect issues (through simulation and real data analysis) is still necessary, however, to further investigate this hypothesis.

A total of 36% of the total variation in **Y** was spatially organised, indicating a rather strong overall spatial structure in the community composition, although only about a half of it could be related to the environment. Yet, inspite of the broad range of spatial scales encompassed by the quadrat approach, neutral dynamics (i.e. ecological drift) seems to dominate among the processes structuring **Y** (64% of the variation not spatially organised). Using the spatial point pattern analysis to explore the spatial structures below the spatial resolution of the quadrat-approach, however, suggested otherwise.

The individual-based approach used to characterise the bivariate interspecific association patterns revealed that 24% of the species pairs occurred either closer or further away from one another than expected under a model of independent distribution at a neighbourhood of 3 m radius. This interspecific spatial association was related to a significant LA divergence (and a similar nearly significant pattern of the SLA). This means that species occurring close from one another had dissimilar LA, and that the species that possessed similar LA occurred further apart from one another than expected by chance. Leaf area can therefore be assumed to be related to a coexistence mechanisms acting very locally; differences of LA reflecting stabilising niche differences (Chesson 2000, HilleRisLambers et al. 2012). This trait is a functional trait mostly related to leaf energy and water balance (Cornelissen et al. 2003, Díaz et al. 2015), and is positively correlated to the species adult height, at the global scale (Díaz et al. 2015). It can therefore be seen as a proxy for the light-preemption dominance or for the water-uptake strategy. The positive spatial association of species that possess dissimilar LA may indicate a water and/or light partitioning allowing functionally complementary species to coexist at very short ranges, while a competitive exclusion would hinder the establishment of individuals possessing too similar LA (i.e. limiting similarity). Variables of light and water availability measured locally (i.e. 3 m) could allow verifying this hypothesis. However, other traits related to the general strategy of individuals to capture light or water should also be investigated. Traits such as height, leaf densities, canopy area, or rooting depth could help getting a better grasp at the multidimensional nature of the plant phenotype (Laughlin and Messier 2015). Further, individual-based measures of functional traits could bring more precise and specific informations, as intraspecific trait variability is known to mediate plant interactions (Schöb et al. 2013, Kraft et al. 2014), and trait values can be directly influenced by neighbour identity (Bowsher et al. 2017). The integration of ontogeny (e.g. dbh) could also provide key information, as competition and facilication among individuals as well as the degree of dependence upon abiotic variables has been shown to depend on this factor (e.g. Miriti 2006, Wiegand et al. 2007b, Valiente-Banuet and Verdú 2008).

We showed that an even stronger pattern of interspecific spatial association was found within 10 m radius around individuals; as such a pattern concerned 181 species pairs (i.e. 71% of all pairs). This result clairly indicates the influence of an important underlying ecological process. Giving that the scale of this pattern closely matches the scale of the minimal spatial resolution of environmental variable measurement, the inclusion of those variables may help further investigate the mechanisms causing a segregation of so many species pairs. The A_{ij} at the range of 10 m was not tested against the environmental variables for now, as using Mantel tests to do so would require as many tests as there are environmental variables. The resulting high number of tests would cause a severe drop of statistical power, as a consequence of the necessary *P*-value correction for multiple tests to maintain an overal correct type I error rate. Note that, although some studies simply performed numerous Mantel tests without bottering with *P*-value adjustment, doing so may have strongly inflated their false discovery rate. Another solution would be to use a bivariate inhomogenous *K*-functions fitted from an inhomogeneous Poisson or Cox point process including the environment as covariates (Velázquez et al. 2016), for instance, to construct the global simulation envelopes. Further investigating relevant key functional traits may also help unravelling key coexistence mechanisms acting at this range (HilleRisLambers et al. 2012).

The bivariate individual-based approach highlighted that 24% of the species pairs had a spatial distribution related to that of at least one other species within a range of 3 m radius. This proportion strongly increased to reach 71% at a range of only 10 m. This result indicates, when considered together with the relatively high degree of spatial structure of the total community composition variation (36%, see Fig. 3), that the tree community displays a strong spatial component overall – although the results of the two analyses are not directly sumable. The spatial dimension of the community encompasses a wide range of spatial scales, with spatial patterns beyond 100 m, others at 65 m, 45 m, and 25 m, while spatial patterns of pairwise interaction were present for nearly three quarters and a quarter of the species pairs within a range of 10 m and 3 m, respectively. While some of these patterns appeared to be potentially caused by induced spatial dependences mostly related to soil chemistry (plant-available forms of N, Zn) and physical variables (soil texture and stoniness), as well as elevation (scale of 45 m), a significant proportion of the community variation (i.e. 26%) displayed spatial patterns unrelated to the environmental conditions. Some of the latter were potentially related to a dispersal limitation (but see previous mention of aliasing effect issues prior to validating this hypothesis).

The two analytical approaches used here are instrincally different by the nature of the objects studied (i.e. abundances in quadrats or spatial locations of individuals), the nature of the test (i.e. raw data analysis versus correlations between similarity matrices), or even the explanatory variables (i.e. environmental variables or functional traits). However, the spatial patterns highlighted by these two approaches are independent and therefore provide complementary valuable information about the signature of processes in the community within a broad range of spatial scales. Even if some bivariate association patterns may occur within some broader spatial structures (e.g. one of the patterns of Fig. 4 or Fig. 5), it would still be complementary, as the structures highlighted by the SPPA occurred < 10 m and were undetectable by the quadrat approach. Regarding the nature of the studied objects, it is precisely a major strength of the combined approach aiming at using complementary methods to address a same question, and we can only think of advantages to use both abundances and individual locations to gain a more detailed insight into the community assembly mechanisms. Future methodological developments aiming at a more integrated use of such combined approaches could however allow a more unified ecological interpretation of the complementary results.

Our study thereby illustrated how complementary data analysis approaches are necessary to gain a more complete insight into the ecological processes contributing to the community assembly (see other examples in De'ath 2002, Meynard et al. 2013, Aiello-Lammens et al. 2017, D'Amen et al. 2017). Using only either the quadrat-based or the individual-based framework would have yielded an incomplete picture of both the magnitude of the spatial organisation of the community and the ecological processes contributing to the community assembly. Moreover, using the inhomogeneous Thomas process with the same environmental variables used in the VP allowed investigating a potential – and rarely tested – role of dispersal limitation to explain residual spatial patterns of the community composition variation. However, a proper simulation study would be needed to further investigate the statistical performances (e.g. power, accuracy) of this combined approach to detect a role of dispersal limitation among unexplained spatial patterns obtained from a quadrat-based approach (e.g. VP combined with MEM). The study also highlights the inherent inference limitations of empirical field studies performed alone, regardless of how accurate and relevant the environmental variables can be. Therefore, both methodological advances and original new combinations of complementary analytical approaches are still needed for the empirical study of plant communities to catch up with the promising theoretical framework offered by the contemporary coexistence theory. In addition, manipulative experiments *in situ* and/or demographic models combined to the present approach would probably be among the most promising ways to further distinguish between traits that stabilise coexistence and traits that drive competitive exclusion (HilleRisLambers et al. 2012), hence yielding a deeper insights into the precise mechanisms of plant community assembly.

Acknowledgment

We are very grateful to E. Wolff (IGEAT, ULB) for lending us the DGPS. A. Sotiaux, T. Grippa, C. Michellier, and S. Vanhuysse were of valuable help to help with the DGPS. The Trimble products were used in participation in the Trimble University Partnership Program. Many thanks to A. Van Baekel and K. Wart for their important help measuring functional traits. J. Vleminckx, A. Jacobs, and D. Hernandez are warmly thanked for their help on the field.

Supporting information

See *Annexes* at the end of the thesis.

References

Aiello-Lammens, M. E., J. A. Slingsby, C. Merow, H. K. Mollmann, D. Euston-Brown, C. S. Jones et al. 2017. Processes of community assembly in an environmentally heterogeneous, high biodiversity region. Ecography 40:561–576.

Anderson, M. J., and P. Legendre. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. Journal of Statistical Computation and Simulation 62:271–303.

Baddeley, A., E. Rubak, and R. Turner. 2015. Spatial point patterns: methodology and applications with R. Chapman & Hall/CRC.

Baddeley, A., and R. Turner. 2005. spatstat: An R package for Analyzing Spatial Point Patterns. Journal of Statistical Software 12:1–42.

Bauman, D., T. Drouet, S. Dray, and J. Vleminckx. 2018a. Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors. Ecography 41(10): 1638–1649.

Bauman, D., T. Drouet, M.-J. Fortin, and S. Dray. 2018b. Optimizing the choice of a spatial weighting matrix in eigenvector-based methods. Ecology: doi: 10.1002/ecy.2469.

Bauman, D., O. Raspé, P. Meerts, J. Degreef, J. I. Muledi, and T. Drouet. 2016. Multiscale assemblage of an ectomycorrhizal fungal community: the influence of host functional traits and soil properties in a 10-ha miombo forest. FEMS Microbiology Ecology 92:fiw151.

Bauman, D., J. Vleminckx, O. Hardy, and T. Drouet. 2018c. Testing and interpreting the shared space-environment fraction in variation partitioning analyses of ecological data. Oikos, doi: 10.1111/oik.05496.

Bivand, R. 2006. spdep: spatial dependence: weighting schemes, statistics and models. R package (version 0.6-13).

Blonder, B., N. Salinas, L. Patrick Bentley, A. Shenkin, P. O. Chambi Porroa, Y. Valdez Tejeira et al. 2017. Predicting trait-environment relationships for venation networks along an Andes-Amazon elevation gradient. Ecology 98:1239–1255.

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

Borcard, D., P. Legendre, and F. Gillet. 2018. Numerical ecology with R, second edition. Use R! Series. Springer International Publishing, Cham, Switzerland.

Bowsher, A. W., P. Shetty, B. L. Anacker, A. Siefert, S. Y. Strauss, and M. L. Friesen. 2017. Transcriptomic responses to conspecific and congeneric competition in co-occurring Trifolium. Journal of Ecology 105:602–615.

Calcagno, V., and C. De Mazancourt. 2010. glmulti: An R package for easy automated model selection with (Generalized) linear models. Journal of statistical software 34:1–29.

Chase, J. M., and M. A. Leibold. 2003. Ecological niches: linking classical and contemporary approaches. University of Chicago Press.

Chesson, P. 2000. Mechanisms of maintenance of species diversity. Annual Review of Ecology and Systematics 31:343–366.

Cornelissen, J. H. C., S. Lavorel, E. Garnier, S. Diaz, N. Buchmann, D. E. Gurvich et al. 2003. A handbook of protocols for standardised and easy measurement of plant functional traits worldwide. Australian Journal of Botany 51:335–380.

Cornwell, W. K., and D. D. Ackerly. 2009. Community assembly and shifts in plant trait distributions across an environmental gradient in coastal California. Ecological Monographs 79:109–126.

Cuma, F. M., D. Bauman, B. M. Bazirake, Y. Mleci, M. Kalenga, M. N. Shutcha et al. 2018. Edaphic specialisation in relation to termite mounds in Katanga (DR. Congo): a reciprocal transplant experiment with congeneric tree species. Journal of Vegetation Science: doi: 10.1111/jvs.12675.

D'Amen, M., H. K. Mod, N. J. Gotelli, and A. Guisan. 2017. Disentangling biotic interactions, environmental filters, and dispersal limitation as drivers of species co-occurrence. Ecography:1–11.

De'ath, G. 2002. Multivariate regression trees: A new technique for modeling species-environment relationships. Ecology 83:1105–1117.

Díaz, S., J. Kattge, J. H. C. Cornelissen, I. J. Wright, S. Lavorel, S. Dray et al. 2015. The global spectrum of plant forms and function. Nature 529:1–17.

Dormann, C. F., J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré et al. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36:27–46.

Dray, S., D. Bauman, G. Blanchet, D. Borcard, S. Clappe, G. Guenard et al. 2018. adespatial: Multivariate multiscale spatial analysis. R package version 0.2-0.

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

Dray, S., R. Pélissier, P. Couteron, M.-J. Fortin, P. Legendre, P. R. Peres-Neto et al. 2012. Community ecology in the age of multivariate multiscale spatial analysis. Ecological Monographs 82:257–275.

Fargione, J., C. S. Brown, and D. Tilman. 2003. Community assembly and invasion: An experimental test of neutral versus niche processes. Proc:eedings of the National Academy of Sciences of the United States of America 100:8916–8920.

Fox, J., S. Weisberg, B. Price, D. Adler, D. Bates, G. Baud-Bovy. 2018. Package 'car'. Vienna: R Foundation for Statistical Computing.

Gilbert, B., and M. J. Lechowicz. 2004. Neutrality, niches, and dispersal in a temperate forest understory. Proceedings of the National Academy of Sciences 101:7651–7656.

Griffith, D. A., and P. R. Peres-Neto. 2006. Spatial modeling in Ecology: the flexibility of eigenfunction spatial analyses. Ecology 87:2603–2613.

HilleRisLambers, J., P. B. Adler, W. S. Harpole, J. M. Levine, and M. M. Mayfield. 2012. Rethinking community assembly through the lens of coexistence theory. Annual Review of Ecology, Evolution, and Systematics 43:227–248.

Hooper, D. U., and J. S. Dukes. 2010. Functional composition controls invasion success in a California serpentine grassland. Journal of Ecology 98:764–777.

Kraft, N. J. B., P. B. Adler, O. Godoy, E. C. James, S. Fuller, and J. M. Levine. 2015. Community assembly, coexistence and the environmental filtering metaphor. Functional Ecology 29:592–599.

Kraft, N. J. B., G. M. Crutsinger, E. J. Forrestel, and N. C. Emery. 2014. Functional trait differences and the outcome of community assembly: An experimental test with vernal pool annual plants. Oikos 123:1391–1399.

Laliberté, E., and P. Legendre. 2010. A distance-based framework for measuring functional diversity from multiple traits. Ecology 91:299–305.

Laliberté, E., P. Legendre, and B. Shipley. 2014. FD: measuring functional diversity from multiple traits, and other tools for functional ecology. R package version 1.0-12.

Laughlin, D. C., and J. Messier. 2015. Fitness of multidimensional phenotypes in dynamic adaptive landscapes. Trends in Ecology and Evolution 30:487–496.

Lavorel, S., and E. Garnier. 2002. Predicting changes in community composition and ecosystem functioning from plant traits: Revisiting the Holy Grail. Functional Ecology 16:545–556.

Legendre, P., and E. D. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia 129:271–280.

Legendre, P., and L. Legendre. 2012. Numerical Ecology. Elsevier, Amsterdam.

Lin, Y.-C., L.-W. Chang, K.-C. Yang, H.-H. Wang, and I.-F. Sun. 2011. Point patterns of tree distribution determined by habitat heterogeneity and dispersal limitation. Oecologia 165:175–184.

Malhi, Y., C. A. J. Girardin, G. R. Goldsmith, C. E. Doughty, N. Salinas, D. B. Metcalfe et al. 2017. The variation of productivity and its allocation along a tropical elevation gradient: a whole carbon budget perspective. New Phytologist 214:1019–1032.

McIntire, E. J. B., and A. Fajardo. 2009. Beyond description: The active and effective way to infer processes from spatial patterns. Ecology 90:46–56.

Meynard, C. N., S. Lavergne, I. Boulangeat, L. Garraud, J. Van Es, N. Mouquet et al. 2013. Disentangling the drivers of metacommunity structure across spatial scales. Journal of Biogeography 40:1560–1571.

Miriti, M. N. 2006. Ontogenetic shift from facilitation to competition in a desert shrub. Journal of Ecology 94:973–979.

Myllymäki, M., T. Mrkvička, P. Grabarnik, H. Seijo, and U. Hahn. 2017. Global envelope tests for spatial processes. Journal of the Royal Statistical Society. Series B: Statistical Methodology 79:381–404.

Oksanen, J., F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn et al. 2017. Package 'vegan': Community ecology package (version 2.4-3).

Pansu, M., and J. Gautheyrou. 2007. Handbook of soil analysis: mineralogical, organic and inorganic methods. Springer Science & Business Media.

Pavoine, S., E. Vela, S. Gachet, G. De Bélair, and M. B. Bonsall. 2011. Linking patterns in phylogeny, traits, abiotic variables and space: A novel approach to linking environmental filtering and plant community assembly. Journal of Ecology 99:165–175.

Pebesma, E. J. 2004. Multivariable geostatistics in S: The gstat package. Computers and Geosciences 30:683–691.

Peres-Neto, P. R., P. Legendre, S. Dray, and D. Borcard. 2006. Variation partitioning of species data matrices: estimation and comparison of fractions. Ecology 87:2614–2625.

Pérez-Harguindeguy, N., S. Diaz, E. Garnier, S. Lavorel, H. Poorter, P. Jaureguiberry et al. 2013. New Handbook for standardized measurment of plant functional traits worldwide. Australian Journal of Botany 61:167–234.

Platt, T., and K. L. Denman. 1975. Spectral analysis in ecology. Annual Review of Ecology and Systematics 6:189–210.

Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research. Sankhya A26:329–358.

Ripley, B. D. 1981. Spatial statistics. John Wiley and Sons.

Schöb, C., C. Armas, M. Guler, I. Prieto, and F. I. Pugnaire. 2013. Variability in functional traits mediates plant interactions along stress gradients. Journal of Ecology 101:753–762.

Šidák, Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 62:626–633.

Suding, K. N., S. Lavorel, F. S. Chapin, J. H. C. Cornelissen, S. Díaz, E. Garnier et al. 2008. Scaling environmental change through the community-level: A trait-based response-and-effect framework for plants. Global Change Biology 14:1125–1140.

Tichý, L. 2016. Field test of canopy cover estimation by hemispherical photographs taken with a smartphone. Journal of Vegetation Science 27:427–435.

Valiente-Banuet, A., and M. Verdú. 2008. Temporal shifts from facilitation to competition occur between closely related taxa. Journal of Ecology 96:489–494.

Velázquez, E., I. Martínez, S. Getzin, K. A. Moloney, and T. Wiegand. 2016. An evaluation of the state of spatial point pattern analysis in ecology. Ecography 39:1042–1055.

Velázquez, E., C. E. T. Paine, F. May, and T. Wiegand. 2015. Linking trait similarity to interspecific spatial associations in a moist tropical forest. Journal of Vegetation Science 26:1068–1079.

Vellend, M., D. S. Srivastava, K. M. Anderson, C. D. Brown, J. E. Jankowski, E. J. Kleynhans et al. 2014. Assessing the relative importance of neutral stochasticity in ecological communities. Oikos 123:1420–1430.

Vleminckx, J., J.-L. Doucet, J. Morin-Rivat, A. B. Biwolé, D. Bauman, O. J. Hardy. 2017. The influence of spatially structured soil properties on tree community assemblages at a landscape scale in the tropical forests of southern Cameroon. Journal of Ecology 105:354–366.

Wagner, H. H., and S. Dray. 2015. Generating spatially constrained null models for irregularly spaced data using Moran spectral randomization methods. Methods in Ecology and Evolution 6:1169–1178.

Wang, X., T. Wiegand, Z. Hao, B. Li, J. Ye, and F. Lin. 2010. Species associations in an old-growth temperate forest in north-eastern China. Journal of Ecology 98:674–686.

Weiss, A. D. 2001. Topographic Positions and Landforms Analysis (poster), ESRI International User Conference, July 2001. San Diego, CA: ESRI.

Wiegand, T., S. Gunatilleke, and N. Gunatilleke. 2007a. Species Associations in a Heterogeneous Sri Lankan Dipterocarp Forest. The American Naturalist 170:E77–E95.

Wiegand, T., S. Gunatilleke, N. Gunatilleke, and T. Okuda. 2007b. Analyzing the spatial structure of a Sri Lankan tree species with multiple scales of clustering. Ecology 88:3088–3102.

Wiegand, T., A. Huth, S. Getzin, X. Wang, Z. Hao, C. V. S. Gunatilleke et al. 2012. Testing the independent species' arrangement assertion made by theories of stochastic geometry of biodiversity. Proceedings of the Royal Society B: Biological Sciences 279:3312–3320.

Wiegand, T., and K. A. Moloney. 2014. Handbook of spatial point pattern analysis in ecology. Chapman & Hall/CRC, Boca Raton.

Wilson, M. F. J., B. O'Connell, C. Brown, J. C. Guinan, and A. J. Grehan. 2007. Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope. Page Marine Geodesy.

Wright, I. J., P. B. Reich, M. Westoby, D. D. Ackerly, Z. Baruch, F. Bongers et al. 2004. The worldwide leaf economics spectrum. Nature 428:821–827.

Discussion générale

I. Assemblage des communautés et partition de la variation

Dans la majeure partie de cette thèse, le cadre d'analyse et d'interprétation écologique de l'assemblage des communautés a été celui de la partition de variation. Comme nous l'avons vu au fil des chapitres, cette méthode offre plusieurs éléments importants et complémentaires d'information quant aux processus structurant les communautés. Néanmoins, dans les Chapitres III, IV et V, différentes sources de biais – potentiellement très élevés – de plusieurs de ces éléments d'informations écologiques (fractions explicatives) ont été mises en évidence et corrigées. Nous allons maintenant passer en revue les fractions de la partition de variation, l'information qu'elles portent en théorie, l'interprétation parfois multiple que l'on peut en faire, les biais méthodologiques soulignés dans le travail et les solutions apportées.

Dans une partition de variation « classique » à deux composantes (environnementales, **E** et spatiale, **S**) d'une communauté **Y**, l'analyse génère une fraction environnementale pure – [a] –, une fraction environnementale partageant des structures spatiales de **Y** – [b], ou *shared space-environement fraction* (SSEF) –, une fraction spatiale pure – [c] – et la fraction correspondant aux résidus du modèle, c'est-à-dire la proportion de variation inexpliquée – [d] (voir Fig. 5 de l'*Introduction*).

I.1. Théorie de la niche : l'effet de l'environnement

L'effet de l'environnement est donc repris par les fractions [a], [b] et [ab]. Il s'agit là de l'effet de la **sélection**, dans le cadre conceptuel synthétique de Vellend (2010). L'avantage potentiel de considérer [a] et [b] séparément est que, si [ab] donne une information quantitative globale importante de l'ampleur de l'influence environnementale sur la composition de la communauté, la fraction [b] permet de visualiser spatialement la structure et la dimension des patterns de **Y** reflétant des patterns similaires présents dans **E**. Tel que suggéré dans le Chapitre VI, l'échelle précise de ces patterns peut, en outre, être estimée en ajustant un modèle théorique de semivariogramme au semivarigramme empirique du pattern (voir Fig. 4 et 5 du Chapitre VI).
La fraction [b] est ainsi, en théorie, la signature d'une dépendance spatiale de **Y** vis-à-vis de **E** (***induced spatial dependence***, ISD ; Legendre et Legendre 2012). Toutefois, un apport majeur de cette thèse a été de mettre à disposition un test non biaisé de cette fraction (Chapitre V), la fraction [b] ne pouvant jusqu'à présent pas être testée. Alors qu'une source fondamentale de structuration des communautés était minimisée au profit de la fraction [a] – d'interprétation plus sûre – ou était inférée avec le risque d'interpréter une fraction due au hasard (Peres-Neto et Legendre 2010), une ISD peut à présent être testée de manière non biaisée. L'application de cette procédure de test requiert néanmoins le respect de trois conditions préliminaire au test (voir Chapitre V) et l'utilisation d'une procédure optimisée de sélection d'un ensemble de prédicteurs spatiaux à partir d'un ensemble de matrices de pondération spatiale candidates (Chapitre III et IV). Les conditions font référence à la présence de structures environnementale et spatiale globales significative dans **Y**, ainsi qu'à la présence d'une

structure spatiale significative dans **E**. Alors que ces conditions visent à limiter le plus possible le risque de détecter un signal fortuit, l'optimisation de la sélection d'un ensemble de prédicteurs spatiaux vise à maximiser la puissance du test ainsi que la précision de détection des patterns spatiaux.

La procédure de test permet également de différencier une fraction [b] (ajustée) négative étant le fruit d'un artefact lié à la procédure ajustement (voir Peres-Neto et al. 2006) d'une fraction [b] négative correspondant à une réelle ISD, qu'elle teste dans ce dernier cas avec une puissance élevée (voir Table S4, Chapitre V). Finalement, il a aussi été montré que les problèmes de fractions [b] négatives (ne portant pas d'information écologique réelle) causées par un effet supresseur entre la fraction [b] et la fraction [c] pouvaient être évités en s'assurant d'utiliser une procédure adéquate de sélection de prédicteurs spatiaux (Chapitre III) pour qu'il ne reste plus de variables MEM modélisant des structures de **E** sans être utiles à la modélisation de structures de **Y** (voir Appendix S4 du Chapitre V). Ainsi, tant que les précautions mentionnées ci-dessus sont respectées, la procédure de test de la fraction [b] est fiable quant à la détection d'ISD associées à une fraction [b] négative.

Comme cela a été discuté dans le Chapitre V, la fraction [a] – la portion d'explication pourvue par **E** et qui n'est pas spatialement structurée – a généralement été considérée seule, faute de test fiable de la fraction [b]. Néanmoins, un effet de structuration majeur des communautés lié à **E** est attendu dans la fraction [b] plutôt que dans la fraction [a], dès lors que les variables physiques et chimiques naturelles sont de nature intrinsèquement spatiale (Legendre et Legendre 2012). En outre, il est très probable que l'environnement dans la fraction [a] soit lui aussi spatialement structuré – de par la nature spatiale de toute variable écologique – à une échelle trop fine ou trop large que pour être détectée par l'étendue de l'étude et/ou le design d'échantillonnage. La distinction entre [a] et [b] relèverait donc plus de considérations de limites de détection spatiale que de leur nature non spatialement structurée *versus* spatialement struturée. Ainsi, on peut imaginer qu'une taille de quadrat d'échantillonnage légèrement diminuée ou une étendue d'étude plus grande pourrait permettre de détecter la structure spatiale auparavant indétectable de [a]. Une partie ou toute la fraction [a] pourrait dans ce cas passer dans la fraction [b] et ne serait alors plus testée ni interprétée, selon la manière classique d'effectuer et interpréter les résultats d'une partition de variation. Ainsi, quantitativement, considérer la fraction globale de **E** – [ab] – fournit une information importante, dès lors qu'elle donne une notion de l'ampleur des processus déterministes (sélection ; aux échelles englobées par le design d'échantillonnage, voir Chase 2014, Garzon-Lopez et al. 2014). La fraction [b] permet, quant à elle, une visualisation et une quantification de l'échelle des patterns d'ISD dans **Y**. Quant à la fraction [a], elle pourrait indiquer que restent à découvrir des structures spatiale plus fines ou plus larges (selon la nature des variables environnementales correspondantes) de l'environnement et potentiellement des structures induites par celles-ci dans **Y**.

I.2. Théorie neutre : Dispersion limitée et stochasticité

En théorie, si toutes les variables environnementales pertinentes ont été échantillonnées, les structures spatiales de **Y** non-expliquées par **E** seraient causées soit par une capacité de **dispersion limitée** (Legendre et Legendre 2012, Borcard et al. 2018), soit par un **évènement historique** (perturbations, par ex. chablis, tornade, feu etc.). Ceci souligne la plus haute importance conférée à un échantillonnage de toutes les variables environnementales pertinentes et les plus précises possibles (variables directement liés à la physiologie des espèces à préférer aux variables indirectes, constituant des *proxy* de variables directes, voir Dormann et al. 2012), dès lors que toute variable environnementale importante mais non mesurée sera responsable d'une part des fraction [c] et [d], selon qu'elle est spatialement structurée ou non (Borcard et Legendre 1994, Legendre et Legendre 2012).

De plus, il a été montré que la quantité et la qualité des variables environnementales incluses dans la partition de variation peut influencer les fractions relatives à **E** à un point tel que les perceptions des processus de niche et neutres peuvent s'en retrouver inversées (Jones et al. 2008, Chang et al. 2013). Ceci explique la quantité de variables édaphiques utilisées dans les Chapitre I, II, III et VI. Ce n'est que lorsque les variables écologiques clés ont été considérées explicitement dans l'analyse que l'on peut interpréter de manière fiable les fractions [c] et [d] comme résultant de processus neutres (respectivement de dispersion limitée et de dérive, Vellend 2010), tout en restant prudent quant à la possibilité d'une variable omise dans l'analyse ou de micro-habitats ponctuels et non capturés par le design environnemental (Borcard et al. 2018). De même, ce n'est qu'au prix de nombreuses mesures de variables environnementales que l'on peut donner une interprétation fiable de l'amplitude de la fraction [d] comme étant la signature de procesuss stochastiques non spatialement structurés (**dérive écologique**) par rapport à la possibilité qu'il s'agisse de l'effet de variables environnementales non mesurées.

La fraction [c] peut néanmoins également refléter un évènement historique ayant structuré la communauté dans le passé sans qu'une trace mesurable ne soit restée (Legendre et Legendre 2012). La visualisation sur carte des patterns spatiaux inexpliqués de **Y** (Borcard et al. 2004, McIntire et Fajardo 2009) de même que la définition précise de leur échelle (voir Chapitre VI, mais aussi Sharma et al. 2012) peuvent permettre d'orienter l'inférence des processus du côté des processus neutres ou des évènements historiques. Ainsi, par exemple, alors que certains patterns spatiaux résiduels du Chapitre VI auraient potentiellement pu être causés par une dispersion limitée (Fig. 5c, d, voire b), une signature d'évènement passé ou de variable omise semblent plus probables pour le pattern de la Fig. 5a.

I.3. Limitations du cadre d'interprétation de la partition de variation

Au vu des points prémentionnés, la partition de variation s'avère être un cadre d'analyse et d'interprétation puissant et riche d'informations précieuses quant aux processus sous-tendant l'assemblage de la communauté et à leur importance relative. Toutefois, cet outil – aussi utile soit-il – a ses limites.

Tout d'abord, il est important de rappeller que dans le cadre global des approches corrélatives – auquel appartient la partition de variation – une **corrélation** ne suppose pas automatiquement de lien de **causalité** (Borcard et al. 2004, Dormann et al. 2012, Brown et al. 2016). Ainsi, une fraction [b] significative, par exemple, peut correspondre à une dépendance spatiale induite (ISD), mais peut aussi provenir d'une variable environnementale non mesurée influençant en parralèle et de façon similaire une variable environnementale mesurée et **Y** (Borcard et al. 2004). Dans les approches corrélatives telles que les analyses basées sur des régressions, c'est dans les hypothèses de départ que réside la causalité, d'où l'importance accordée à la formulation d'**hypothèses a priori**, c'est-à-dire avant d'effectuer l'analyse (McIntire et Fajardo 2009). Toutefois, on peut s'attendre à ce que la présence d'un tel facteur confondant soit d'autant moins probable que le pattern spatial partagé par **E** et **Y** est complexe. Le cas particulier des gradients environnementaux grossiers selon les coordonnées X et/ou Y (tendance linéaire) est discuté plus loin.

Deuxièmement, si la partition de variation offre un cadre très utile pour séparer des effets déterministes liés à la niche des espèces (effet de la sélection, Vellend 2010) d'effets neutres ou historiques, la distinction de mécanismes plus précis au sein de ces deux catégories est plus difficile, voire impossible sans approche complémentaire.

Ainsi, une ISD indique qu'une variable environnementale a imposé sa structure à celle de la communauté. Le mécanisme responsable de l'ISD pourrait être causé par l'exclusion de certaines espèces par **incompatibilité physiologique** liée à certaines valeurs de variables environnementales (effet sur la niche fondamentale des espèces). Le succès (par ex. taux de croissance) des espèces est alors directement l'effet de l'environnement (véritable filtre abiotique). D'autre part, le pattern d'ISD pourrait être issu d'interactions biotiques médiées par la variable abiotique sous-tendant le pattern. Les espèces ont dans ce cas une différence relative de *fitness* pouvant s'exprimer comme une fonction de l'environnement, les rendant plus ou moins compétitives selon la valeur de la variable abiotique concernée. Des espèces moins compétitives dans certaines conditions environnementales pourraient ainsi être exclues localement et ce bien que leur optimum écologique puisse potentiellement se trouver dans les zones dont elles ont été exclues. Ceci mène à une **spécialisation de l'habitat** de certaines espèces pouvant générer le pattern observé. Enfin, les interactions biotiques peuvent encore générer des patterns spatiaux de composition spécifique d'autres manières, notamment par des patterns d'associations d'espèces résultant d'une **limitation de la similarité** permise par un partitionnement des ressources entre espèces fonctionnellement complémentaires (Wright et Westoby 2002, Díaz et al. 2016, HilleRisLambers et al. 2012). Ce dernier mécanisme génère des patterns d'associations spatiales interspécifiques, comme celui détecté sur base de la surface foliaire à l'échelle du voisinage direct des individus dans le Chapitre VI, dès lors que les espèces exploitant la ressource concernée par l'ISD de façon suffisamment complémentaires peuvent se trouver à une certaine distance l'une de l'autre.

La partition de variation ne permet pas de différencier ces différents mécanismes d'assemblage sous-tendant potentiellement une ISD. Néanmoins, son utilisation est une première étape permettant de mettre en évidence cet effet de sélection spatialement structuré et les variables environnementales qui lui sont associées, ce qui permet alors de formuler de nouvelles hypothèses quant aux mécanismes précis possiblement concernés. Ces mécanismes peuvent alors être confirmés par l'**utilisation combinée de méthodes et approches complémentaires** (voir sections suivantes).

Une mise en garde relative à la perception des processus relatifs à la niche ou à des effets stochastiques est également nécessaire (Brown et al. 2016). En effet, la résolution spatiale d'une étude (étendue, grain et interval d'échantillonnage) doit impérativement être définie de sorte à ce que les abondances des espèces et des variables environnementales varient à une échelle similaire. En outre, l'étendue d'échantillonnage doit être telle que toutes les espèces de la communauté soient présentes, sans pourtant capturer un renouvellement d'espèces résultant de la superposition de plusieurs communautés. Sans cette précaution, l'amplitude des effets de l'habitat peut être biaisée par le fait qu'à une échelle supérieure, les effets déterministes sont justement si fort que certaines espèces ont localement été complètement ou presque complètement exclues de la zone de communauté étudiée, ce qui diminuerait artifiellement la perception de l'influence de l'habitat (Grime 2006, Brown et al. 2016). Ainsi, les processus de niche « cryptiques » peuvent ne pas être détectés nécessairement comme étant déterministes. Cette problématique d'échelle est bien illustrée dans la Fig. 3 de l'*Introduction* (voir Chase 2014, Garzon-Lopez et al. 2014 pour une discussion détaillée des liens entre échelle et perception des classes de processus).

De même, Brown et al. (2016) soulignent que la dispersion limitée n'est un processus neutre qu'à condition qu'elle soit la même pour toutes les espèces, dès lors que la théorie neutre implique que les espèces sont fonctionnellement équivalentes (Hubbell 2001) et ont donc une capacité équivalente de dispersion. Or, il est bien connu que la dispersion peut varier fortement entre espèces (par ex. masse de propagules ou modes de dispersion différents) (Seidler et Plotkin 2006, Low et McPeek 2014).

Finalement, il est une suite de mécanismes précis qu'une partition de variation seule ne peut détecter. En effet, si les espèces doivent s'accomoder de conditions abiotiques et biotiques, elles les affectent également, de sorte que la niche d'une espèce inclut tant sa réponse que son impact sur les environnements abiotique et biotique (Hutchinson 1957, Tilman 1982, Lavorel et Garnier 2002, Chase et Leibold 2003, Suding et al. 2008). De même, la réponse d'une espèce à son environnement abiotique influence ses interactions biotiques et vice versa (Tilman 1982), que ce soit par l'épuisement d'une ressource, par une restitution de litière améliorante ou acidifiante, ou par l'accumulation d'ennemis naturels (herbivores, pathogènes ; Levine et HilleRisLambers 2009, Yamazaki et al. 2009).

Il a été montré dans le Chapitre VI que la théorie contemporaine de la coexistence offre un cadre conceptuel fin pour la compréhension des processus mentionnés ci-dessus au travers des concepts de **mécanismes stabilisateurs** (*stabilising niche differences*) et de **différences relatives de valeur adaptative** (*fitness*) (Chesson 2000, HilleRisLambers et al. 2012). Notons qu'ici, la *fitness* utilise l'espèce comme point de comparaison (dominance compétitive) – et non pas l'individu, comme en biologie évolutive. Ainsi les mécanismes stabilisateurs sont ceux qui rendent le succès d'une espèce négativement dépendant de sa fréquence (*negative frequency-dependent growth rate*) et ont pour conséquence de favoriser la coexistance, dès lors qu'une espèce devenant rare est favorisée (Chesson 2000). L'hétérogénéité des conditions environnementales (McCarthy-Neumann et Kobe 2010), les ennemis naturels, le microbiote du sol (Petermann et al. 2008, Yamazaki et al. 2009, Swamy et Terborgh 2010) ou encore l'allélopathie (McCarthy-Neumann et Kobe 2010) sont autant d'exemples de processus stabilisateurs favorisant la coexistence des espèces. En l'absence de ces processus, les différences relatives de *fitness* entre espèces mènent à l'exclusion de celles dont les traits ne sont pas suffisament proches de la ou des valeur(s) optimale(s) dans les conditions environnementales locales. Des différences de tolérance à l'herbivorie (HilleRisLambers et al. 2010), de capacité à prélever une ressource limitante (Suding et al. 2005) ou encore de croissance influencée par les conditions environnementales (Cuma et al. 2018) sont des exemples pour lesquels un avantage compétitif peut mener à l'exclusion de certaines espèces, en l'absence de compromis fonctionnels stabilisateurs.

Ce dernier paragraphe permet de prendre la mesure du chemin qu'il reste encore à parcourir entre l'analyse de données par la partition de variation associée à des vecteurs propres spatiaux et la théorie fine des mécanismes précis d'assemblage. Toutefois, cette limitation n'est pas propre qu'à la partition de variation et, comme illustré dans le Chapitre VI, une combinaison d'approches analytiques complémentaires peut permettre de diminuer la distance séparant le cadre théorique des limitations d'interprétations des analyses réalisées. L'utilisation d'approches complémentaires à la partition de variation est, comme nous le verrons par la suite, une voie prometteuse qui a encore besoin d'être développée.

II. Approche fonctionnelle de l'assemblage des communautés

Dans les Chapitres II et VI, des traits fonctionnels d'arbres ont été utilisés dans des applications différentes. Dans le Chapitre II, ils ont servi à expliquer les variations de composition spécifique de la communauté des champignons ectomycorhiziens sur base des stratégies fonctionnelles des arbres-hôtes. Dans le Chapitre VI, les traits foliaires des arbres ont été utilisés pour expliquer un pattern d'association spatiale interspécifique et ont permis de mettre en évidence un mécanisme stabilisateur favorisant la coexistence d'espèces complémentaires d'un point de vue fonctionnel à une échelle de voisinage direct. Ces exemples illustrent l'utilité des approches fonctionnelles pour aborder les mécanismes d'assemblage des communautés.

L'utilisation de traits permet en effet d'établir un lien entre processus et fonction, dès lors qu'un pattern d'agrégation ou de répulsion peut directement être lié à une ou plusieurs fonctions des plantes. Ceci permet d'émettre des hypothèses plus ciblées sur les mécanismes précis qui sous-tendent l'effet stabilisateur (divergence des valeurs de traits) ou d'exclusion (convergence ou divergence des valeurs de traits) détecté. Dans le Chapitre VI par exemple, la corrélation négative des paires d'espèces présentant un pattern d'association bivarié à une échelle de 3 m avec la similarité des valeurs de surface foliaire a permis de détecter que ce trait est probablement lié à un mécanisme stabilisateur lié à l'exploitation de l'eau ou de la lumière.

Alors que l'approche utilisée dans la thèse visait à expliquer des abondances ou des positions d'individus par des variables environnementales ou des traits fonctionnels, une approche complémentaire consiste à chercher directement à expliquer des **patterns de traits fonctionnels** sur base de variation de conditions environnementales (Kraft et al. 2008, Cornwell et Ackerly 2009, Adler et al. 2013, Delhaye et al. 2016). Les traits pouvant être directement liés à certaines fonctions vitales des plantes, un pattern de convergence ou de divergence des valeurs de certains traits permet non seulement de détecter la présence d'un mécanisme de coexistence, mais aussi de directement diriger la réflexion de la nature du mécanisme sur les fonctions auxquelles les traits sont liés (HilleRisLambers et al. 2012). Bien entendu, ceci n'est valable qu'à condition d'étudier des traits pertinents, c'est-à-dire dont on est sûre que la valeur influence la valeur adaptative de l'espèce dans l'environnement abiotique et biotique étudié (Laughlin et al. 2018). Ceci peut être facilité par l'utilisation de traits physiologiques, directement liés à la fonction de la plante qu'on suppose être à la base d'un mécanisme de coexistence (par ex. Malhi et al. 2015, 2017, Blonder et al. 2017).

Une approche complémentaire aux approches centrées sur la composition spécifique de la communauté ou sur l'utilisation de traits consiste à considérer explicitement les relations phylogénétiques entre espèces comme *proxy* de traits affectés par la sélection mais non mesurés ou difficilement mesurables (Webb et al. 2002, Guénard et al. 2013, de Bello et al. 2017). L'interprétation écologique de patterns de sous ou surdispersion phylogénétiques dépend néanmoins de l'hypothèse, rarement vérifiée, que la niche est phylogénétiquement conservée (Cavender-Bares et al. 2009). Cette approche, potentiellement complémentaire à l'utilisation de traits fonctionnels, reste relativement controversée (Kraft et al. 2007, Cavender-Bares et al. 2009, Mayfield et Levine 2010, Pavoine et Ricotta 2013) et ne sera pas développée ici.

III. Complémentarité des approches d'étude de l'assemblage des communautés

Que ce soit sur base des outils statistiques présentés et développés dans les Chapitres III, IV et V ou sur bases des approches fonctionnelles et/ou phylogénétiques abordées ci-dessus, l'approche corrélative, utilisée seule, est limitée de manière ultime en termes d'inférences de processus d'assemblage de communauté (Dormann et al. 2012, HilleRisLambers et al. 2012, Laughlin et al. 2018). Ceci s'explique par le fait que, tel que discuté pour une dépendance spatiale induite (ISD) dans le cadre de la partition de variation, un pattern donné peut souvent autant être la signature d'un mécanisme stabilisateur que d'une exclusion compétitive résultant de différences de *fitness*.

Les patterns suggérant une dépendance spatiale induite obtenus sur base de la partition de variation du Chapitre VI (Fig. 4) illustrent bien cette problématique. Soit le pattern de la communauté résulte de différences de niche fondamentale (viabilité physiologique) parmi les espèces, soit ce sont des interactions biotiques qui sont à la base du pattern (voir spécialisation de l'habitat et limitation de la similarité), sans compter que des mêmes variables environnementales ont souvent une dynamique complexe et liée à la fois aux mécanismes de stabilisation et d'exclusion (HilleRisLambers et al. 2012). De même, un resserrement de la gamme des traits (convergence) peut indiquer l'action sous-jacente d'un filtre abiotique ne permettant qu'aux espèces présentant les traits adaptés de s'installer et de se maintenir, comme il peut s'agir de l'effet d'un filtre biotique de compétition intense ne laissant s'installer que les espèces suffisamment compétitives (par ex. présentant un syndrome fonctionnel de dominance ; égalisation de la valeur sélective, Chesson 2000, Mayfield et Levine 2010).

De plus, à moins d'utiliser un trait physiologique intimement lié à une fonction précise de la plante, la mise en évidence d'un processus de coexistence générant un pattern pour des traits morphologiques – les plus souvent utilisés en raison de leur facilité de mesure (de Bello et al. 2017) – ne permet pas toujours la détection claire de la fonction clé affectée chez l'espèce (Adler et al. 2013). Ceci est notamment dû au fait que les traits morphologiques sont souvent des *proxy* de différentes fonctions (la SLA, par exemple, est liée à l'utilisation de l'eau et à la compétitivité pour l'azote, Suding et al. 2005).

L'utilisation combinée d'études de terrain et d'expériences de manipulation de l'assemblage de la communauté ou de manipulation de la source des effets stabilisateurs ou différences de *fitness* relatives offre une voie complémentaire pour démêler effets stabilisateurs et d'exclusion (voir HilleRisLambers et al. 2012 pour une synthèse et des exemples). Néanmoins, la capacité de telles études de manipulations à inférer des processus et leurs conséquences dans le monde réel est régulièrement remise en cause (par ex. HilleRisLambers et al. 2012, Vellend et al. 2013).

Une dernière approche de terrain très complémentaire consiste à étudier les **taux démographiques** des espèces de la communauté (germination, recrutement, croissance, mortalité) et leur variation en fonction de variables environnementales et des traits fonctionnels. L'avantage de cette approche est que (1) les taux démographiques sont le résultat net de la somme de tous les processus antagonistes d'assemblage des communautés, et (2) tout processus stabilisateur influence la coexistence par le biais d'un taux de croissance supérieur des individus d'une espèce lorsque celle-ci est rare (Chesson 2000). Ceci peut être mis en évidence sur base des taux de croissance des individus par exemple. Des taux de croissance dépendant négativement de la fréquence chez diverses espèces à par exemple permis de détecter une signature claire d'effets stabilisateurs et d'en comprendre le mécanisme (Petermann et al. 2008, Yamazaki et al. 2009, Clark 2010, McCarthy-Neumann et Kobe 2010, Blonder et al. 2018).

Tel que détaillé dans l'Introduction, une porte d'entrée privilégiée vers la somme de ces questions et considérations écologiques sur la nature et le fonctionnement des mécanismes de l'assemblage des communautés d'arbres est l'étude détaillée des patterns spatiaux (et temporels) d'organisation de la composition taxonomique, fonctionnelle et démographique des communautés. C'est donc à présent sur ces considérations plus méthodologiques que je vais revenir.

IV. Inférence de processus à partir de patterns spatiaux

IV.1. Evolution des méthodes d'analyses spatiales

Dans le Chapitre III, une synthèse représentative des publications de domaines variés utilisant des vecteurs propres spatiaux a été réalisée. Celle-ci a mis en évidence qu'une proportion élevée (~40 %) des travaux publiés depuis plus de 10 ans utilisent ces méthodes de façon sous-optimale à hautement biaisée, compromettant ainsi l'inférence de processus à partir des patterns ou simplement la détection de patterns spatiaux corrects. Dans le Chapitre II, les *smoothed* MEM ont été utilisées pour la partie centrale des résultats, évitant ainsi la question, encore épineuse à ce moment, de la sélection d'un sous-ensemble de vecteurs propres. Le principe de cette méthode consiste donc à garder la totalité des variables MEM générées et à les regrouper en un certain nombre d'ensembles (voir Munoz 2009, Dray et al. 2012). Néanmoins, pour la partition de variation de ce même chapitre, une sélection de variables

MEM a été réalisée avec la méthode d'optimisation incorrecte basée sur le AIC (Dray et al. 2006). Les résultats relatifs à cette analyse, de même que ceux de près de 40% des études publiées depuis 2006, sont donc potentiellement biaisés et pourraient être réévalués. Le Chapitre III a également permis de définir les précautions évitant une inflation du taux d'erreur de type I et a mené à une définition claire des méthodes de sélection de variables MEM les plus adaptées en fonction de la nature des données et des objectifs visés.

Dans le Chapitre IV, une série de **matrices de pondérations spatiales** (SWMs) a été mise à l'épreuve en termes de puissance et de précision d'estimation du R^2 spatial. Il a ainsi été montré que les *distance-based* MEM – utilisées dans la majorité des travaux – étaient peu puissantes et imprécises en comparaison avec les *graph-based* MEM lorsque le design d'échantillonnage était irrégulier. Il convient ici de mentionner que, dans le Chapitre II, si une *graph-based* SWM (*Gabriel graph*) a bien été utilisée dans le cadre de la détection de structures spatiales associées ou non à l'environnement (Fig. 1 et 2), il aurait convenu d'utiliser des *graph-based* MEM pour la partition de variation aussi, et ce sans même considérer de *distance-based* MEM dans la procédure d'optimisation.

En outre, les conclusions du Chapitre III concernant le contrôle de l'erreur de type I ainsi que la sélection d'une méthode adaptée de sélection de vecteurs propres ont pu être mises à profit dans l'élaboration d'une approche d'**optimisation de la sélection de la SWM** pouvant s'adapter à l'objectif visé par l'utilisation de vecteurs propres spatiaux. Il a été montré que cette optimisation augmente la puissance de détection ainsi que la précision de la description des patterns spatiaux de variables réponses présentant des structures d'échelles et de degré d'autocorrélation variés.

Dans le Chapitre V, ces améliorations méthodologiques ont à leur tour permis le développement d'un test non biaisé de la variation de composition d'une communauté expliquée à la fois par un ensemble de variables environnementales et par des prédicteurs spatiaux (*shared space-environment fraction*, SSEF), dans le cadre du partitionnement de la variation. Ce test de la SSEF est un élément crucial qui s'ajoute à la boîte à outil de l'écologie des communautés, dès lors qu'il permet de tester la présence d'une dépendance spatiale induite par l'environnement au sein de la communauté (*induced spatial dependence*, ISD). Toutefois, tel que présenté dans le Chapitre V et discuté précédemment, la validité statistique ainsi que la puissance du test dépendent conjointement de la bonne utilisation des MEM (voir chapitre III et IV) ainsi que de trois tests préliminaires limitant le risque de faux positifs (voir tests globaux et présence d'un signal spatial global dans le jeu de données environnemental).

Les Chapitre III à V soulignent ainsi l'importance d'une compréhension fine et d'une utilisation adéquate des outils statistiques d'étude des structures spatiales afin de mettre en évidence les processus sous-jacents. Dans le cas contraire, une illustration a été donnée dans ces trois chapitres des risques de perte de puissance, de précision ou encore les détections injustifiées de patterns spatiaux. Ces problèmes mènent alors à des interprétations écologiques biaisées (voir surestimation de la fraction spatiale pure et de la SSEF dans l'exemple de la Fig. 4, Chapitre III). Ces biais divers dans l'utilisation de vecteurs propres spatiaux et donc de la partition de variation empêchent une comparaison fiable entre études. Ils compromettent également la réalisation de méta-analyses visant à expliquer des patterns globaux d'influence de certains processus (par ex. Soininen 2016). Cette problématique devrait être résolue par une adoption généralisée des « bonnes pratiques » proposées dans ces trois chapitres.

Il semble également important de souligner que, lorsqu'elles sont utilisées correctement (voir chapitre III et IV), les MEM ne détectent pas de structuration spatiale lorsque la communauté n'est pas spatialement structurée (voir Fig. 3a et Fig. 2a, b des Chapitres III et IV respectivement). Une SSEF significative n'est pas non plus détectée lorsque la communauté et l'environnement sont tous deux structurés mais présentent des patterns indépendants (voir scénario additionnel 1, Table S4 du Chapitre V). Ces résultats confortent le fait que les MEM, seules ou combinées à un cadre de partition de variation, sont tout à fait fiables en termes de risques de faux positifs, contrairement à ce qui était avancé par Gilbert et Bennett (2010) (voir détails dans la discussion du Chapitre V).

Ces trois chapitres de développements méthodologiques illustrent bien que si une approche empirique de terrain ne permet pas toujours de dêmeler certains mécanismes d'assemblages de communautés (par exemple, Mayfield et al. 2005, ou pattern d'ISD de la Fig. 4 du Chapitre VI), l'évolution des méthodes statistiques améliore continuellement la capacité d'inférer des processus écologiques à partir de patterns spatiaux de composition de communauté, de traits ou de relations phylogénétiques (par ex. Pavoine et al. 2011, Dray et al. 2012, Guénard et al. 2013, de Bello et al. 2017, Clappe et al. 2018).

IV.2. Perspectives de développements méthodologiques

IV.2.1. Matrice de pondération spatiale et réalité de terrain

Dans le Chapitre IV, il a été montré que l'optimisation de la sélection d'une matrice de pondération spatiale (SWM) à partir de différents candidats réalistes augmentait à la fois la puissance et la précision des MEM pour détecter des patterns complexes. A l'avenir, une voie prometteuse pour encore affiner la détection de ces patterns pourrait consister à intégrer des considérations d'écologie du paysage (par ex. barrières de dispersion naturelles ou anthropiques) ainsi que de biologie des organismes étudiés (par ex. évitement de certains éléments paysagers, directionnalité de la dispersion) afin de définir plus finement les chemins de connexions entre zones utilisées pour construire la SWM. Il s'agit donc d'intégrer plus d'information biologique et écologique de terrain afin de rendre une méthode purement mathématique plus proche de la réalité de la connectivité des communités étudiées. Les *spatial graphs* (voir *Minimum Planar Graph* discuté dans le Chapitre IV ; Fall et al. 2007), surfaces de résistances (Spear et al. 2010), chemins de moindre coût (par exemple, Rayfield et al. 2010, Mui et al. 2017) et l'utilisation de la théorie des circuits (Hanks et Hooten 2013) se sont montrés particulièrement utiles dans ce but et sont autant d'options intéressantes à explorer. Ceci permettrait en outre de restreindre le nombre de SWMs candidates à tester et donc de limiter la diminution de puissance liée à la correction du seuil de siginificativité dans la procédure d'optimisation (voir chapitre IV). C'est néanmoins principalement aux échelles régionales et globales que le gain de précision résultant de ces apports seront les plus importants.

IV.2.2. Puissance, échelle et signe de l'autocorrélation spatiale

Un deuxième élément qui mériterait d'être approfondi est celui du lien entre puissance de la méthode MEM, échelle spatiale d'un pattern à détecter et valeur propre des variables MEM nécessaires à sa modélisation. En effet, il a été montré dans le Chapitre III que la puissance des MEM décroissait fortement lorsque les variables MEM nécessaires à la modélisation d'un pattern avaient une valeur propre s'approchant de zéro, en valeur absolue. Pour mieux comprendre ce phénomène, il est important de clairement définir le lien entre la valeur propre et l'échelle d'un pattern et ce, tant pour les patterns d'autocorrélation spatiale positive que négative (valeurs propres positives et négatives respectivement ; Bauman et al. 2018a) (Fig. 1).

En effet, une certaine confusion existe entre l'échelle d'un pattern et le signe de son autocorrélation spatiale. Ainsi, il arrive que les patterns de large et fine échelles (encore appelés « globaux » ou « locaux ») soient respectivement associés à une autocorrélation spatiale positive et négative (Thioulouse et al. 1995, Jombart et al. 2008). S'il est clair que les patterns négativement autocorrélés sont, de façon générale, plus fins que les patterns positivement autocorrélés, les patterns spatiaux d'agrégation les plus fins (donc positivement autocorrélés) méritent tout de même le terme de pattern de fine échelle (Fig. 1). Ainsi, il conviendrait de définir une terminologie explicite du point de vue du signe de l'autocorrélation considérée afin d'éviter la confusion selon laquelle seules les patterns spatiaux négativement autocorrélés sont appelés patterns d'échelle fine.

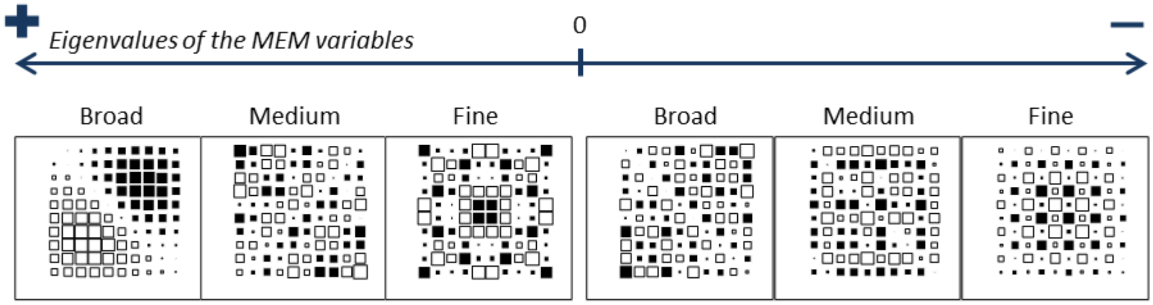


Figure 1 : Illustration du lien entre échelle spatiale, valeur absolue et signe des valeurs propres (*eigenvalues*) associées aux variables MEM. *Broad*, *Medium* et *Fine* correspondent respectivement à des exemples de patterns spatiaux de large, moyenne et fine échelle.

IV.2.3. Aliasing effects

Dans les Chapitre II et VI, la problématique des « ***aliasing effects*** » a été soulevée. Pour rappel, ces derniers correspondent à des biais de détection d'échelle de patterns spatiaux résultant d'un sous-échantillonnage des zones d'observations de la variables réponse (c'est-à-dire, les quadrats) (Platt et Denman 1975). On parle de sous-échantillonnage car le biais résulte justement du fait que la dimension de la somme des quadrats devant être considérés pour détecter le signal du pattern spatial est plus élevée que celle du pattern réel. Ceci a pour conséquence de donner l'impression à tort que l'échelle réelle du pattern est celle qui correspond à la l'espace occupé par des quadrats adjacents de même nature (voir illustration de la Fig. 2). Ainsi, par exemple, un pattern se répétant tous les 3 m pourrait n'apparaître qu'à une échelle de près de 12 m si la distance entre les points d'échantillonnage est de 4 mètres (Fig. 2). De potentiels *aliasing effects* sont toujours présents aux échelles les plus fines qui peuvent être détectées sur base d'un design d'échantillonnage donné. Toute étude d'écologie des communauté cherchant à inférer des processus à partir des patterns détectés à l'aide de vecteurs propres spatiaux ou d'autres méthodes spatiales a donc de bonnes chances de se heurter à une détection de patterns d'échelle relativement fine biaisée par un *aliasing effect*. Cependant, cette problématique ne semble avoir été abordée jusqu'à présent que par le biais des *smoothed* MEM (Munoz 2009) (adoptée dans le Chapitre II), rarement utilisées (11 citations de Munoz 2009 d'après Google Scholar en juin 2018).

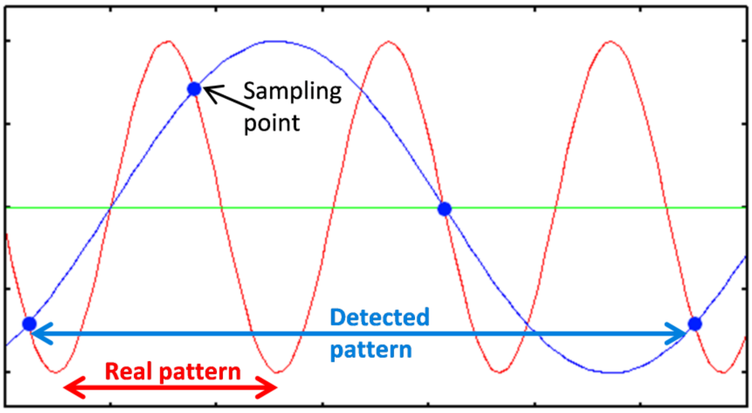


Figure 2 : Illustration schématique d'un *aliasing effect*. Le sous-échantillonnage (points bleux) ne rend la détection du pattern réel (sinusoïde rouge) possible qu'à une échelle trop large (sinusoïde bleue). Chaque point bleu peut être assimilé à un quadrat dans lequel la composition spécifique est étudiée.

Une étude de simulations permettrait de mieux comprendre quantitativement les situations où et la façon dont les *aliasing effects* influencent les patterns détectés par les MEM ou d'autres méthodes spatialement explicites. Ces simulations pourraient consister en la génération d'agrégats d'individus de taille variable sur une grille (voir méthodologie du Chapitre V). Ces agrégats pourraient être secondairement influencés, ou non, par une variable environnementale spatialement structurée à large, moyenne ou fine échelle. La dimension (échelle) de référence des patterns complexes de distribution des individus pourrait être caractérisée par ajustement d'un processus de points (*point process*, voir chapitre VI) considérant à la fois la variable environnementale filtrante et une dispersion limitée suivant la même fonction que celle utilisée pour générer les agrégats initiaux (par exemple, processus de points inhomogène de Cox, Baddeley et al. 2015). Ensuite, un échantillonnage pourrait être effectué selon différentes modalités (contigus, régulier, irrégulier, regroupé, voir Gilbert et Bennett 2010, Bauman et al. 2018b) et il s'agirait alors de tester la précision de récupération du pattern initial à l'aide de la méthode MEM optimisée (voir chapitre III et IV). L'échelle des patterns détectés par les MEM serait caractérisée à l'aide de l'ajustement d'un modèle de semivariogramme théorique au semivariogramme observé du pattern (voir caractérisation de l'échelle des patterns détectés dans le Chapitre VI). Une telle étude théorique pourrait être un premier pas vers une considération explicite – ne fut-ce que dans la discussion des résultats – de cette source de biais d'interprétations écologiques.

IV.2.4. Gradients spatiaux linéaires

Dans le Chapitre V, il a été montré que la procédure de test de la *shared space-environment fraction* (SSEF, ou fraction [b]) détecte une fraction partagée entre l'environnement et les variables spatiales uniquement quand une dépendance spatiale induite (ISD) a été utilisée pour générer les données (Fig. 3a, b). Il existe néanmoins une exception : lorsque l'abondance d'une ou plusieurs espèces et une variable environnementale augmentent ou diminuent graduellement ensemble selon l'axe X et/ou Y tout le long de l'étendue du site d'étude. L'inférence d'un processus écologique pose problème pour un tel gradient spatial, encore appelé « **tendance linéaire** » (*spatial trend*). En effet, si une tendance linéaire peut réellement indiquer un lien direct entre une variable environnementale et la composition de la communauté (c'est-à-dire, une ISD), elle peut également être le signe d'un processus prenant place à une échelle dépassant l'étendue du site d'étude et influençant à la fois les variables environnementales et les abondances des espèces sans qu'il n'y ait pour autant de lien direct entre celles- ci (Borcard et al. 2004). Dans ce dernier cas, l'aspect grossier et non répliqué du pattern peut facilement mener à une corrélation fortuite élevée et donc à une SSEF significative. Inférer une ISD à partir d'une tendance linéaire est donc très délicat.

Borcard et al. (2004) ont suggéré, à l'époque des *principal coordinates of neighbour matrices* (PCNM, Borcard et Legendre 2002) précurseurs des MEM (Dray et al. 2006), qu'en présence d'une tendance linéaire significative, les variables réponses et explicatives concernées devaient être « détendancées » (c'est-à-dire prendre les résidus d'un modèle des variables en fonction des coordonnées X-Y concernées par la tendance linéaire ; *detrending*) avant l'utilisation de la méthode PCNM. La raison avancée pour ce traitement préalable est que non seulement une tendance linéaire est délicate à interpréter et devrait être considérée séparément, mais qu'en plus le *detrending* éviterait l'utilisation superflue de vecteurs propres pour modéliser un pattern grossier que les coordonnées X-Y pourraient capturer. Ceci permettrait *in fine* de préserver les vecteurs propres PCNM pour la détection de patterns plus complexes. Ainsi, détendancer augmenterait la puissance des PCNM pour détecter des patterns fins éventuels tout en évitant les problèmes d'interprétation.

Néanmoins, tel que mentionné dans le Chapitre V, la pratique d'élimination d'une tendance linéaire avant d'effectuer une analyse spatiale basée sur des vecteurs propres n'a encore jamais été réellement évaluée, que ce soit avec les PCNM ou avec les MEM. Or, on peut se demander si une telle pratique augmente en effet la puissance et la précision des MEM pour détecter des patterns plus complexes, ou si au contraire cela n'a pas plutôt l'effet inverse. En effet, puisque des structures spatiales ne peuvent être modélisées sans inflation du taux d'erreur de type I qu'à condition que le modèle MEM global soit significatif (voir chapitre III et IV), et sachant que la puissance des MEM diminue pour les structures spatiales fines, on pourrait s'attendre à ce qu'éliminer une tendance linéaire diminue la puissance statistique du modèle MEM global sur les données détendancées (dès lors que le modèle global pourrait avoir une moindre puissance).

J'ai testé cette hypothèse au travers de simulations dont le principe, le code R et les résultats sont présentés en annexe (Appendix S1). Il s'agissait de simuler l'abondance d'une espèce ($\boldsymbol{y}$) de façon à ce que celle-ci soit la superposition d'une ou deux tendances linéaires générées à partir des coordonnées

X et/ou Y, d'un pattern spatial supplémentaire d'échelle large, intermédiaire ou fine, généré à partir de variables MEM, et d'une variation aléatoire représentant la dérive écologique. L'amplitude relative de la dérive écologique et des structures spatiales fluctuaient de sorte à ce que le R^2 spatial de référence se trouve dans une gamme de valeurs réalistes (entre 0.17 et 0.65, voir Appendix S1). La puissance et la précision de détection de la structure spatiale totale était alors testée pour deux procédures consistant soit à directement utiliser la procédure d'optimisation des MEM (voir chapitre IV), soit à tester la présence d'une tendance, détendancer les données et ensuite utiliser la procédure d'optimisation des MEM. La puissance et précision ont également été évaluées de façon similaire en utilisant les PCNM originelles. La procédure de simulation a été appliquée pour un $\boldsymbol{y}$ échantillonné selon deux types de design d'échantillonnage : régulier et aléatoire.

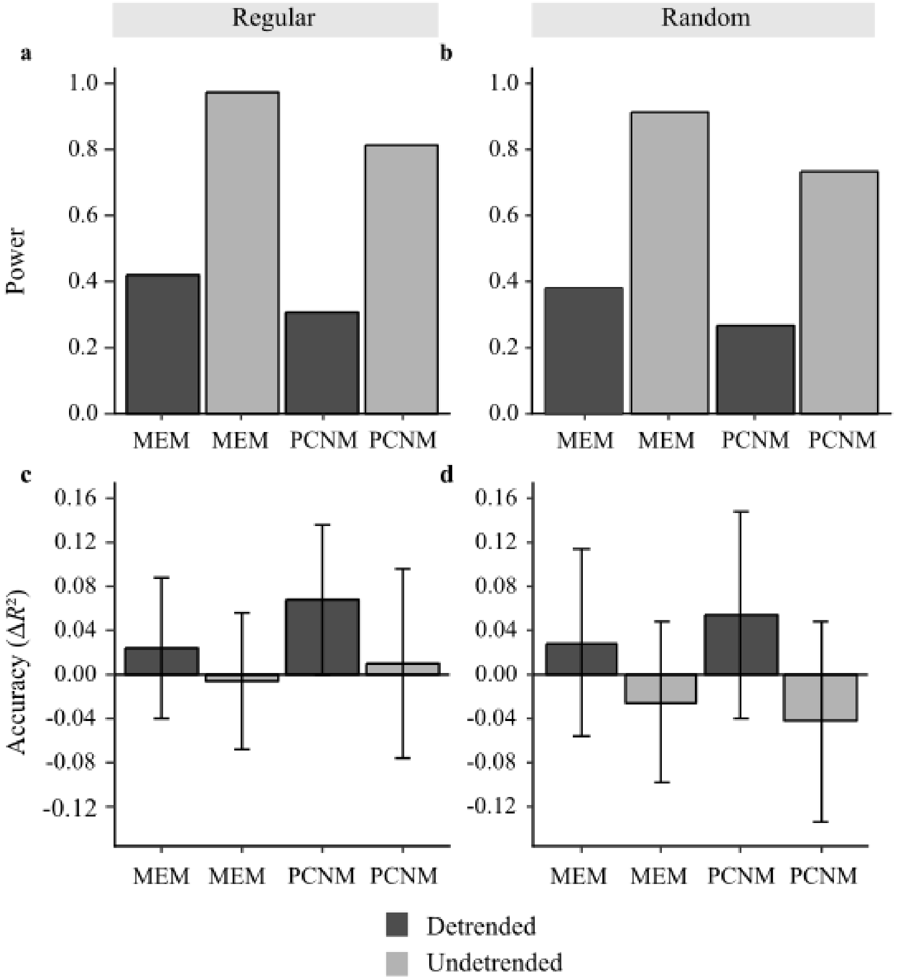


Figure 3 : Puissance statistique et précision d'estimation du R^2 spatial des simulations visant à comparer l'utilisation d'un *detrending* précédant l'utilisation de la procédure optimisée des MEM (Chapitre IV) ou l'utilisation des PCNM (*Detrended*) avec une utilisation directe des méthodes (*Undetrended*). Les barres représentent la puissance (a, b) et la précision (c, d ; $\Delta R^2 = R^2_{\text{simulé}} - R^2_{\text{référence}}$) moyennes de 1000 simulations d'une variable réponse résultant de la superposition de une ou deux tendances linéaires, une structure de large, moyenne ou fine échelle (MEM) et un bruit aléatoire d'amplitude variable (voir Appendix S1). a, b : La puissance est définie comme la proportion des 1000 simulations pour laquelle le test global des MEM ou PCNM était significatif (*p*-value < 0.05). c, d : Les valeurs négatives et positives indiquent respectivement une sous-estimation et une surestimation du R^2 de référence. Les barres d'erreur sont les écart-types.

Les résultats obtenus à partir de 1000 simulations indiquent de façon frappante qu'indépendament du type de design d'échantillonnage, la puissance statistique, tant des MEM optimisées que des PCNM, était fortement diminuée lorsqu'un *detrending* précédait l'utilisation des vecteurs propres (puissance des MEM avec et sans *detrending* respectivement ~0.4 et > 0.9) (Fig. 3a, b). De plus, le *detrending* n'améliorait pas la précision de détection des patterns spatiaux complexes, mais menait même à une légère surestimation de l'ampleur réelle du signal spatial (Fig. 3c, d ; voir détails dans l'Appendix S1).

Il semblerait donc que si une ISD apparente correspondant à une tendance linéaire ne devrait pas être interprétée sur un même pied d'égalité qu'une SSEF associée à un pattern plus complexe, la tendance linéaire ne devrait pas pour autant être éliminée, faute de quoi le risque de laisser passer des structures plus fines augmenterait fortement. Je suggèrerais, sur base de ces résultats préliminaires, qu'une solution serait de tester la présence d'une tendance linéaire et, le cas échéant, éviter d'interpréter les variables environnementales qui y sont associées, sans pour autant retirer la tendance linéaire des données analysées.

Ces quatre points de perspectives de développements méthodologiques fournissent un bon aperçu des exemples d'avancées futures qui permettront une inférence toujours plus fine et fiable des processus écologiques à partir de patterns spatiaux, mais aussi une prise de conscience des limites de ces méthodes afin d'éviter des inférences abusives. Ceci souligne également l'importance des études de simulations telles que celles utilisées dans les Chapitres III, IV et V.

V. Du processus vers le pattern

En parallèle des approches corrélatives utilisées dans la thèse et discutées jusqu'à présent, des approches dites mécanistes, orientées directement vers les processus (***process-based models***), permettent une compréhension complémentaire des patterns de distributions des espèces (Adler et al. 2006, Clark 2010, Schurr et al. 2012, Cabral et al. 2017, Munoz et al. 2018). Ces approches considèrent les processus ou mécanismes écologiques de façon explicite, contrairement aux approches corrélatives qui considèrent les processus implicitement par l'intermédiaire des patterns et sont donc souvent limitées pour identifier des liens de causalité (Dormann et al. 2012, Cabral et al. 2017).

Alors que les approches corrélatives lient directement les variables environnementales à l'occurrence ou à l'abondance des espèces, les modèles basés sur les processus formulent l'écologie d'une espèce à partir de fonctions mathématiques, de façon réductionniste, en contrôlant donc la causalité. L'abondance ou l'occurrence de l'espèce émerge ainsi indirectement du modèle (voir Dormann et al. 2012 pour une synthèse). Cette dichotomie entre les approches corrélatives et mécanistes est toutefois plus conceptuelle que réelle, dès lors qu'il existe un continuum de méthodes entre les extrèmes de ces deux approches. Ce continuum peut être caractérisé par (1) à quel point le modèle est exclusivement construit à partir de processus mécanistes et (2) le degré de calibration du modèle, c'est-à-dire la quantité de paramètres du modèles devant être ajustés à partir de données réelles. Différents types de modèles ont été développés et se placent le long de ce continuum, chacune des méthodes présentant individuellement des avantages et des limitations, de sorte qu'il est généralement utile d'utiliser une approche analytique combinée (Dormann et al. 2012). Un exemple de cette combination a notamment été présenté dans le Chapitre VI, par l'ajustement d'un paramètre de taille de pattern de dispersion à partir des données réelles de la distribution des individus de chaque espcèce (au moyen des outils de l'analyse de semis de points), lequel a alors été utilisé comme aide à l'interprétation de l'approche corrélative qu'est la partition de variation.

Là où la causalité n'est jamais sûre dans l'approche corrélative, c'est par contre au travers de cette dernière que de nouveaux processus ou interactions entre processus peuvent être découverts, dès lors que les données explorées contiennent toute la complexité présente dans la nature. Cette approche est donc généralement préférée pour *comprendre* un système et générer des hypothèses. Comme discuté précédemment, ces dernières peuvent alors être testées expérimentalement (Suding et al. 2005, HilleRisLambers et al. 2012), ou être testées dans des approches mécanistes utilisant une paramétrisation provenant de l'approche corrélative (par exemple, Adler et al. 2006, Clark 2010). En conséquence, si l'approche corrélative – utilisée avec des variables environnementales pertinentes, en quantité et de qualité suffisante (Jones et al. 2008, Chang et al. 2013) – permet de générer des hypotheses sur le rôle de divers processus, l'inférence de la causalité entre les patterns détectés et le processus doit être réalisée avec prudence. En outre, les approches corrélatives visent avant tout à expliquer un système donné, mais possède une capacité de généralisation ou de transfert vers d'autres espèces ou sites relativement faible (voir Peterson et al. 1999, Hein et al. 2007, Pearman et al. 2008).

Dans l'approche mécaniste, par contre, les processus et leurs interactions doivent être définis a priori, de sorte que seuls les processus intégrés au modèle pourront être détectés (Dormann et al. 2012). De plus, les modèles mécanistiques nécessitant généralement d'être au moins un peu paramétrisés, le modélisateur doit tout de même choisir les variables de terrain mesurées qu'il considère pour modéliser les processus et paramètres inconnus. En outre, au plus les données réelles sont utilisées pour paramétriser le modèle, au ces mêmes données pourront être utilisées pour vérifier que le modèle est correcte. Le modèle mécaniste servira alors souvent plutôt un but plus *prédictif* qu'explicatif. Les modèles mécanistiques permettent par exemple de prédire des propriétés dynamiques de la distribution de espèces et de l'assemblage de la communauté, tel que l'influence de perturbations (Schumacher et Bugmann 2006), ou la confirmation d'une différence de niche stabilisatrice ou de la signature d'une différence relative de fitness dans un type de pattern observé (Adler et al. 2006, 2012, Clark 2010).

Les approches corrélatives et mécanistes sont donc complémentaires et les perspectives les plus prometteuses d'avancées tant dans la compréhension des mécanismes d'assemblage des communautés que dans la prédiction de leur composition le long de gradients spatiaux et/ou temporels naturels de contraintes environnementales ou face aux perturbations anthropiques directes (par ex. fragmentation de l'habitat) et indirectes (par ex. changements climatiques) prendront très probablement naissance de l'utilisation combinée de ces approches complémentaires (Dormann et al. 2012, HilleRisLambers et al. 2012, Aiello-Lammens et al. 2017, Cabral et al. 2017, Munoz et al. 2018).

VI. Conséquences des développements méthodologiques sur le terrain : Choix d'un design d'échantillonnage

Les améliorations et apports méthodologiques des Chapitres III et IV ont mis en évidence que le succès dans la détection de patterns spatiaux complexes variait fortement selon l'échelle spatiale du pattern, selon le degré d'autocorrélation dans les variables réponses, mais également selon le type de design d'échantillonnage adopté. Si dans un premier temps ces éléments ont permis d'évaluer les performances des étapes clés des MEM, la méthode à présent optimisée et non biaisée pourrait en second lieu aider à définir les types de design d'échantillonnage qui optimisent la détection de patterns spatiaux dans les communautés.

La question du choix d'un design d'échantillonnage est une question cruciale qui se pose à toute équipe de chercheurs lors de la mise en place d'une stratégie d'échantillonnage, que ce soit en écologie des (méta-)communautés (Legendre et Legendre 2012, Bauman et al. 2018a, b, Brind'Amour et al. 2018), en épidémiologie (par ex. Artois et al. 2016, Thanapongtharm et al. 2016) ou dans d'autres domaines confrontés à des phénomènes spatialisés. Ainsi, un même effort d'échantillonnage peut être atteint avec différentes combinaisons 1) de grain d'étude, 2) de nombre de sites (par ex. quadrats ou dispositifs de forêts) échantillonnés, 3) d'étendue couverte par l'échantillonnage et 4) de répartition spatiale des sites les uns par rapport aux autres. Néanmoins, on peut fortement s'attendre à ce que le choix de ces propriétés de l'échantillonnage influence la détection de patterns spatiaux dans la communauté (Legendre et al. 2009, Legendre et Legendre 2012, Lin et al. 2013, Bauman et al. 2018a, b, Brind'Amour et al. 2018).

L'utilisation jusqu'à présent très fortement majoritaire de *distance-based* MEM (*db*-MEM) a imposé une contrainte sévère sur le design d'échantillonnage, dès lors que la résolution spatiale minimale des patterns détectables est pour cette famille de MEM liée à la plus petite distance qui permet de maintenir tous les points d'échantillonnage connectés (le plus long lien d'un *minimum spanning tree*) (Borcard et Legendre 2002, Legendre et Legendre 2012). Cette thèse a cependant montré que cette contrainte d'un design d'échantillonnage régulier pouvait à présent être relaxée.

En effet, le Chapitre IV a mis l'emphase sur les *graph-based* MEM (*gb*-MEM), beaucoup plus adaptées aux designs d'échantillonnage irréguliers ou stratifiés. En outre, une méthode d'optimisation puissante, non-biaisée et flexible a été mise sur pied afin de sélectionner un ensemble de variables MEM parmi un groupe de matrices de pondération spatiale pré-sélectionnées sur mesure par rapport au type de design d'échantillonnage (Bauman et al. 2018b). Cette étude a montré que des puissances statistiques et précisions très élevées pouvaient ainsi être obtenues à partir de différents types de designs d'échantillonnage et pour des patterns d'échelles contrastées.

En parallèle, Brind'Amour et al. (2018) ont récemment montré que les *db*-MEM pouvaient être utilisées avec une précision et puissance statistique relativement supérieure à celles associées à leur utilisation classique, pour des échantillonnages irréguliers. Pour ce faire, les *db*-MEM sont générées à partir des coordonnées spatiales des points échantillonnés auxquelles sont ajoutées les coordonnées spatiales de points artificiels, non échantillonnés, qui rendent le design d'échantillonnage initialement irrégulier plus proche d'un design régulier. Les MEM sont générées à partir de cet ensemble augmenté de points, suite à quoi les éléments des vecteurs propres spatiaux correspondant aux points ajoutés sont extraits. Cette pratique a comme désavantage que les variables MEM deviennent colinéaires (perte de leur propriété d'orthogonalité) et qu'elles ne maximisent plus l'indice *I* de Moran. Cependant, elle augmente la puissance statistique et la précision des *db*-MEM par rapport à leur utilisation classique, diminuant donc l'importance de la régularité du design d'échantillonnage.

Ces avancées parallèles dans le domaine des méthodes basées sur des vecteurs propres spatiaux permettent ainsi de relaxer la contrainte d'un design d'échantillonnage régulier. Elles mènent néanmoins également à la question du choix du design optimal.

Parmi les grands types de designs d'échantillonnage utilisés en écologie des communautés se retrouvent 1) le design régulier à sites (ou quadrats) adjacents (voir Chapitre I), 2) le design à sites espacés de façon régulière, 3) le design aléatoire (voir Chapitres III et IV), le design « en grappes » ou stratifié (plusieurs groupes de sites plus proches entre eux que des sites des autres groupes), ainsi que des intermédiaires entre certaines de ces catégories (par ex. design en grappes à sites distribués aléatoirement ou régulièrement et de manière contiguë ou pas au sein des groupes).

Dans les cas où une hétérogénéité environnementale est apparente (par ex. topographie, proximité d'un cours d'eau) ou connue car étudiée au préalable, on pourrait s'attendre à ce que – pour un nombre de sites, une étendue et un grain d'étude fixes – un design d'échantillonnage stratifié consistant à positionner les sites en groupes sur des conditions environnementales contrastées soit le meilleur choix. En effet, non seulement un signal spatial d'échelle large pourrait apparaître sur base des différentes grappes de sites, probablement plus homogènes en leur sein qu'elles ne le sont entre elles (si la variable environnemental considérée pour le choix du placement des grappes est en effet importante pour les communautés étudiées), mais en plus, des patterns spatiaux beaucoup plus fins pourraient être détectés au sein des grappes grâce à la puissance fournie par le nombre élevés de sites s'y trouvant et donc relativement beaucoup plus proches entre eux que de tout autre site au sein d'un autre groupe.

Dans le cas où aucun élément paysager ni connaissance environnementale préalable ne sont disponibles, un design d'échantillonnage aléatoire pourrait avoir plus de chance de détecter des structures d'échelle spatiale fine qu'un design régulier, dès lors que les distances entre sites varient et des sites proches pourraient favoriser localement la détection de structures fines. Il est néanmoins également possible qu'un design régulier soit plus puissant pour détecter une structure répétée à intervalles réguliers, pour laquelle la puissance statistique pourrait être insuffisante sur base d'un design irrégulier ou aléatoire qui manquerait éventuellement des zones clés de l'étendue étudiée.

Les études de simulations générant des motifs spatiaux d'échelles et intensités variables sur des grilles complètes ensuite échantillonnées selon différentes modalités (types de design d'échantillonnage, étendues spatiales) seraient nécessaires pour vérifier ces différentes hypothèses et tester l'effet de différentes stratégies d'échantillonnages sur les performances statistiques des MEM (Brind'Amour et al. 2018, Bauman et al. 2018a, b). Ce cadre de travail devrait ainsi prochainement permettre de définir les designs d'échantillonnage qui maximisent la puissance et la précision de détection de patterns spatiaux de différentes échelles. Ceci fournira une suite de recommandations spatialement explicites pour la mise en place d'une stratégie d'échantillonnage selon différents éléments, tels que l'hétérogénéité de l'environnement, l'échelle supposée d'un processus (par ex. dispersion limitée, dépendance spatiale induite) ou encore des contraintes de terrain empêchant l'échantillonnage de certaines zones.

Conclusion

Cette thèse a permis de mettre en évidence le rôle fondamental que jouent les approches méthodologiques dans l'avancée de la compréhension des mécanismes de structuration des communautés. Il est apparu clairement au fil des pages qu'il existe une kyrielle de méthodes puissantes permettant d'aborder des aspects complémentaires de l'écologie des espèces composant la communauté. Ces méthodes permettent une approche complémentaire et une compréhension plus profonde de l'assemblage des communautés quand elles sont utilisées en parallèle ou de façon intégrée. De même, l'utilisation combinée d'approches corrélatives spatialement explicites, d'approches expérimentales sur terrain, de simulations et de modèles mécanistiques basés sur les processus semble être la prochaine étape prioritaire dans l'approfondissement de la compréhension des mécanismes précis de coexistence et d'assemblage des communautés d'arbres.

Au cours du travail, il a aussi été montré à plusieurs reprises que tant un choix inadapté de méthode, un non-respect des conditions d'application de tests, ou une utilisation biaisée ou sous-optimale de méthodes pouvait mener à des interprétations écologiques biaisées, voir opposées à la réalité. Ceci souligne le rôle essentiel du développement continu de méthodes spatialement explicites ainsi que de la compréhension précise de la puissance et des limites de ces méthodes pour répondre aux questions fondamentales de l'écologie des communautés.

Références

Adler, P. B., H. J. Dalgleish, and S. P. Ellner. 2012. Forecasting plant community impacts of climate variability and change: When do competitive interactions matter? Journal of Ecology 100:478–487.

Adler, P. B., A. Fajardo, A. R. Kleinhesselink, and N. J. B. Kraft. 2013. Trait-based tests of coexistence mechanisms. Ecology Letters 16:1294–1306.

Adler, P. B., J. HilleRisLambers, P. C. Kyriakidis, Q. Guan, and J. M. Levine. 2006. Climate variability has a stabilizing effect on the coexistence of prairie grasses. Proceedings of the National Academy of Sciences 103:12793–12798.

Aiello-Lammens, M. E., J. A. Slingsby, C. Merow, H. K. Mollmann, D. Euston-Brown, C. S. Jones et al. 2017. Processes of community assembly in an environmentally heterogeneous, high biodiversity region. Ecography 40:561–576.

Artois, J., Newman, S., Dhingra, M. S., Chaiban, C., Linard, C., Cattoli, G. et al. 2016. Clade-level Spatial Modelling of HPAI H5N1 Dynamics in the Mekong Region Reveals New Patterns and Associations with Agro-Ecological Factors. Scientific Reports 6: 30316.

Baddeley, A., E. Rubak, and R. Turner. 2015. Spatial point patterns: methodology and applications with R. Chapman & Hall/CRC.

Bauman, D., T. Drouet, S. Dray, and J. Vleminckx. 2018a. Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors. Ecography 41(10): 1638–1649.

Bauman, D., T. Drouet, M.-J. Fortin, and S. Dray. 2018b. Optimizing the choice of a spatial weighting matrix in eigenvector-based methods. Ecology: doi: 10.1002/ecy.2469.

de Bello, F., P. Šmilauer, J. A. F. Diniz-Filho, C. P. Carmona, Z. Lososová, T. Herben et al. 2017. Decoupling phylogenetic and functional diversity to reveal hidden signals in community assembly. Methods in Ecology and Evolution 8:1200–1211.

Blonder, B., R. E. Kapas, R. M. Dalton, B. J. Graae, J. M. Heiling, and Ø. H. Opedal. 2018. Microenvironment and functional-trait context dependence predict alpine plant community dynamics. Journal of Ecology 106:1323–1337.

Blonder, B., N. Salinas, L. Patrick Bentley, A. Shenkin, P. O. Chambi Porroa, Y. Valdez Tejeira et al. 2017. Predicting trait-environment relationships for venation networks along an Andes-Amazon elevation gradient. Ecology 98:1239–1255.

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

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

Borcard, D., P. Legendre, C. Avois-Jacquet, and H. Tuomisto. 2004. Dissecting the spatial structure of ecological data at multiple scales. Ecology 85:1826–1832.

Borcard, D., P. Legendre, and F. Gillet. 2018. Numerical ecology with R, second edition. Use R! Series. Springer International Publishing, Cham, Switzerland.

Brind'Amour, A., Mahévas, S., Legendre, P., Bellanger, L. 2018. Application of Moran Eigenvector Maps (MEM) to irregular sampling designs. Spatial Statistics 26: 56–68.

Brown, B. L., E. R. Sokol, J. Skelton, and B. Tornwall. 2016. Making sense of metacommunities: dispelling the mythology of a metacommunity typology. Oecologia 183:643–652.

Cabral, J. S., L. Valente, and F. Hartig. 2017. Mechanistic simulation models in macroecology and biogeography: state-of-art and prospects. Ecography 40:267–280.

Cavender-Bares, J., K. H. Kozak, P. V. A. Fine, and S. W. Kembel. 2009. The merging of community ecology and phylogenetic biology. Ecology Letters 12:693–715.

Chang, L. W., D. Zelený, C. F. Li, S. T. Chiu, and C. F. Hsieh. 2013. Better environmental data may reverse conclusions about niche- and dispersal-based processes in community assembly. Ecology 94:2145–2151.

Chase, J. M. 2014. Spatial scale resolves the niche versus neutral theory debate. Journal of Vegetation Science 25:319–322.

Chase, J. M., and M. A. Leibold. 2003. Ecological niches: linking classical and contemporary approaches. University of Chicago Press.

Chesson, P. 2000. Mechanisms of maintenance of species diversity. Annual Review of Ecology and Systematics 31:343–366.

Clappe, S., S. Dray, and P. R. Peres-Neto. 2018. Beyond neutrality: disentangling the effects of species sorting and spurious correlations in community analysis. Ecology:doi: 10.1002/ecy.2376.

Clark, J. S. 2010. Individuals and the variation needed for high species diversity in forest trees. Science 327:1129–1132.

Cornwell, W. K., and D. D. Ackerly. 2009. Community assembly and shifts in plant trait distributions across an environmental gradient in coastal California. Ecological Monographs 79:109–126.

Cuma, F. M., D. Bauman, B. M. Bazirake, Y. Mleci, M. Kalenga, M. N. Shutcha et al. 2018. Edaphic specialisation in relation to termite mounds in Katanga (DR. Congo): a reciprocal transplant experiment with congeneric tree species. Journal of Vegetation Science: doi: 10.1111/jvs.12675.

Delhaye, G., C. Violle, M. Séleck, E. Ilunga wa Ilunga, I. Daubie, G. Mahy et al. 2016. Community variation in plant traits along copper and cobalt gradients. Journal of Vegetation Science:Doi:10.1111/jvs.12394.

Dormann, C. F., S. J. Schymanski, J. Cabral, I. Chuine, C. Graham, F. Hartig et al. 2012. Correlation and process in species distribution models: Bridging a dichotomy. Journal of Biogeography 39:2119–2131.

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

Dray, S., R. Pélissier, P. Couteron, M.-J. Fortin, P. Legendre, P. R. Peres-Neto et al. 2012. Community ecology in the age of multivariate multiscale spatial analysis. Ecological Monographs 82:257–275.

Fall, A., M. J. Fortin, M. Manseau, and D. O'Brien. 2007. Spatial graphs: Principles and applications for habitat connectivity. Ecosystems 10:448–461.

Garzon-Lopez, C. X., P. A. Jansen, S. A. Bohlman, A. Ordoñez, and H. Olff. 2014. Effects of sampling scale on patterns of habitat association in tropical trees. Journal of Vegetation Science 25:349–362.

Gilbert, B., and J. R. Bennett. 2010. Partitioning variation in ecological communities: do the numbers add up? Journal of Applied Ecology 44:1071–1082.

Grime, J. P. 2006. Trait convergence and trait divergence in herbaceous plant communities: Mechanisms and consequences. Journal of Vegetation Science 17:255–260.

Guénard, G., P. Legendre, and P. Peres-Neto. 2013. Phylogenetic eigenvector maps: A framework to model and predict species traits. Methods in Ecology and Evolution 4:1120–1131.

Hanks, E. M., and M. B. Hooten. 2013. Circuit theory and model-based inference for landscape connectivity. Journal of the American Statistical Association 108:22–33.

Hein, S., B. Binzenhöfer, H. J. Poethke, R. Biedermann, J. Settele, and B. Schröder. 2007. The generality of habitat suitability models: A practical test with two insect groups. Basic and Applied Ecology 8:310–320.

HilleRisLambers, J., P. B. Adler, W. S. Harpole, J. M. Levine, and M. M. Mayfield. 2012. Rethinking community assembly through the lens of coexistence theory. Annual Review of Ecology, Evolution, and Systematics 43:227–248.

HilleRisLambers, J., S. G. Yelenik, B. P. Colman, and J. M. Levine. 2010. California annual grass invaders: The drivers or passengers of change? Journal of Ecology 98:1147–1156.

Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography. Princeton University Press, Princeton, New Jersey.

Hutchinson, G. E. 1957. Population studies: animal ecology and demography: concluding remarks. Page 22:415-27 Cold Spring Harb. Symp. Quant. Biol.

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

Jones, M. M., H. Tuomisto, D. Borcard, P. Legendre, D. B. Clark, and P. C. Olivas. 2008. Explaining variation in tropical plant community composition: Influence of environmental and spatial data quality. Oecologia 155:593–604.

Kraft, N. J. B., W. K. Cornwell, C. O. Webb, and D. D. Ackerly. 2007. Trait evolution, community assembly, and the phylogenetic structure of ecological communities. The American Naturalist 170:271–283.

Kraft, N. J. B., R. Valencia, and D. D. Ackerly. 2008. Functional traits and niche-based tree community assembly in an amazonian forest. Science 322:580–582.

Laughlin, D. C., R. T. Strahan, P. B. Adler, and M. M. Moore. 2018. Survival rates indicate that correlations between community-weighted mean traits and environments can be unreliable estimates of the adaptive value of traits. Ecology Letters 21:411–421.

Lavorel, S., and E. Garnier. 2002. Predicting changes in community composition and ecosystem functioning from plant traits: Revisiting the Holy Grail. Functional Ecology 16:545–556.

Legendre, P., Mi, X., Ren, H., Ma, K., Yu, M., Sun, I.-F. et al. 2009. Partitioning beta diversity in a subtropical broad-leaved forest of China. Ecology 90(3): 663–674.

Legendre, P., and L. Legendre. 2012. Numerical Ecology. Elsevier, Amsterdam.

Levine, J. M., and J. HilleRisLambers. 2009. The importance of niches for the maintenance of species diversity. Nature 461:254–257.

Lin, G., Stralberg, G., Gong, G., Huang, Z., Ye, W., Wu, L. 2013. Separating the effects of environment and space on tree species distribution: from population to community. PLoS ONE 8(2): e56171.

Lowe, W. H., and M. A. McPeek. 2014. Is dispersal neutral? Trends in Ecology and Evolution 29:444–450.

Malhi, Y., C. E. Doughty, G. R. Goldsmith, D. B. Metcalfe, C. A. J. Girardin, T. R. Marthews et al. 2015. The linkages between photosynthesis, productivity, growth and biomass in lowland Amazonian forests. Global Change Biology 21:2283–2295.

Malhi, Y., C. A. J. Girardin, G. R. Goldsmith, C. E. Doughty, N. Salinas, D. B. Metcalfe et al. 2017. The variation of productivity and its allocation along a tropical elevation gradient: a whole carbon budget perspective. New Phytologist 214:1019–1032.

Mayfield, M. M., M. F. Boni, G. C. Daily, and D. Ackerly. 2005. Species and functional diversity of native and human-dominated plant communities. Ecology 86:2365–2372.

Mayfield, M. M., and J. M. Levine. 2010. Opposing effects of competitive exclusion on the phylogenetic structure of communities. Ecology Letters 13:1085–1093.

McCarthy-Neumann, S., and R. K. Kobe. 2010. Conspecific plant-soil feedbacks reduce survivorship and growth of tropical tree seedlings. Journal of Ecology 98:396–407.

McIntire, E. J. B., and A. Fajardo. 2009. Beyond description: the active and effective way to infer processes from spatial patterns. Ecology 90:46–56.

Mui, A. B., B. Caverhill, B. Johnson, M. J. Fortin, and Y. He. 2017. Using multiple metrics to estimate seasonal landscape connectivity for Blanding's turtles (Emydoidea blandingii) in a fragmented landscape. Landscape Ecology 32:531–546.

Munoz, F. 2009. Distance-based eigenvector maps (DBEM) to analyse metapopulation structure with irregular sampling. Ecological Modelling 220:2683–2689.

Munoz, F., M. Grenié, P. Denelle, A. Taudière, F. Laroche, C. Tucker et al. 2018. ecolottery: Simulating and assessing community assembly with environmental filtering and neutral dynamics in R. Methods in Ecology and Evolution 9:693–703.

Pavoine, S., and C. Ricotta. 2013. Testing for phylogenetic signal in biological traits: The ubiquity of cross-product statistics. Evolution 67:828–840.

Pavoine, S., E. Vela, S. Gachet, G. De Bélair, and M. B. Bonsall. 2011. Linking patterns in phylogeny, traits, abiotic variables and space: A novel approach to linking environmental filtering and plant community assembly. Journal of Ecology 99:165–175.

Pearman, P. B., A. Guisan, O. Broennimann, and C. F. Randin. 2008. Niche dynamics in space and time. Trends in Ecology and Evolution 23:149–158.

Peres-Neto, P. R., and P. Legendre. 2010. Estimating and controlling for spatial structure in the study of ecological communities. Global Ecology and Biogeography 19:174–184.

Peres-Neto, P. R., P. Legendre, S. Dray, and D. Borcard. 2006. Variation partitioning of species data matrices: estimation and comparison of fractions. Ecology 87:2614–2625.

Petermann, J. S., A. J. F. Fergus, L. A. Turnbull, and B. Schmid. 2008. Janzen-Connell effects are widespread and strong enough to maintain diversity in grasslands. Ecology 89:2399–2406.

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

Platt, T., and K. L. Denman. 1975. Spectral analysis in ecology. Annual Review of Ecology and Systematics 6:189–210.

Rayfield, B., M. J. Fortin, and A. Fall. 2010. The sensitivity of least-cost habitat graphs to relative cost surface values. Landscape Ecology 25:519–532.

Schumacher, S., and H. Bugmann. 2006. The relative importance of climatic effects, wildfires and management for future forest landscape dynamics in the Swiss Alps. Global Change Biology 12:1435–1450.

Schurr, F. M., J. Pagel, J. S. Cabral, J. Groeneveld, O. Bykova, R. B. O'Hara et al. 2012. How to understand species' niches and range dynamics: A demographic research agenda for biogeography. Journal of Biogeography 39:2146–2162.

Seidler, T. G., and J. B. Plotkin. 2006. Seed dispersal and spatial pattern in tropical trees. PLoS Biology 4:2132–2137.

Sharma, S., P. Legendre, D. Boisclair, and S. Gauthier. 2012. Effects of spatial scale and choice of statistical model (linear versus tree-based) on determining species–habitat relationships. Canadian Journal of Fisheries and Aquatic Sciences 69:2095–2111.

Soininen, J. 2016. Spatial structure in ecological communities - a quantitative analysis. Oikos 125:160–166.

Spear, S. F., N. Balkenhol, M. J. Fortin, B. H. McRae, and K. Scribner. 2010. Use of resistance surfaces for landscape genetic studies: Considerations for parameterization and analysis. Molecular Ecology 19:3576–3591.

Suding, K. N., S. L. Collins, L. Gough, C. Clark, E. E. Cleland, K. L. Gross et al. 2005. Functional- and abundance-based mechanisms explain diversity loss due to N fertilization. Proceedings of the National Academy of Sciences 102:4387–4392.

Suding, K. N., S. Lavorel, F. S. Chapin, J. H. C. Cornelissen, S. Díaz, E. Garnier et al. 2008. Scaling environmental change through the community-level: A trait-based response-and-effect framework for plants. Global Change Biology 14:1125–1140.

Swamy, V., and J. W. Terborgh. 2010. Distance-responsive natural enemies strongly influence seedling establishment patterns of multiple species in an Amazonian rain forest. Journal of Ecology 98:1096–1107.

Thanapongtharm, W., Linard, C., Chinson, P., Kasemsuwan, S., Visser, M., Gaughan, A. E. et al. 2016. Spatial analysis and characteristics of pig farming in Thailand. BMC Veterinary Research 12(1): doi: 10.1186/s12917-016-0849-7.

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

Tilman, D. 1982. Resource competition and community structure. Princeton university press.

Vellend, M. 2010. Conceptual synthesis in community ecology. The Quarterly Review of Biology 85:183–206.

Vellend, M., L. Baeten, I. H. Myers-Smith, S. C. Elmendorf, R. Beausejour, C. D. Brown et al. 2013. Global meta-analysis reveals no net change in local-scale plant biodiversity over time. Proceedings of the National Academy of Sciences 110:19456–19459.

Webb, C. O., D. D. Ackerly, M. A. McPeek, and M. J. Donoghue. 2002. Phylogenies and community ecology. Annual Review of Ecology and Systematics 33:475–505.

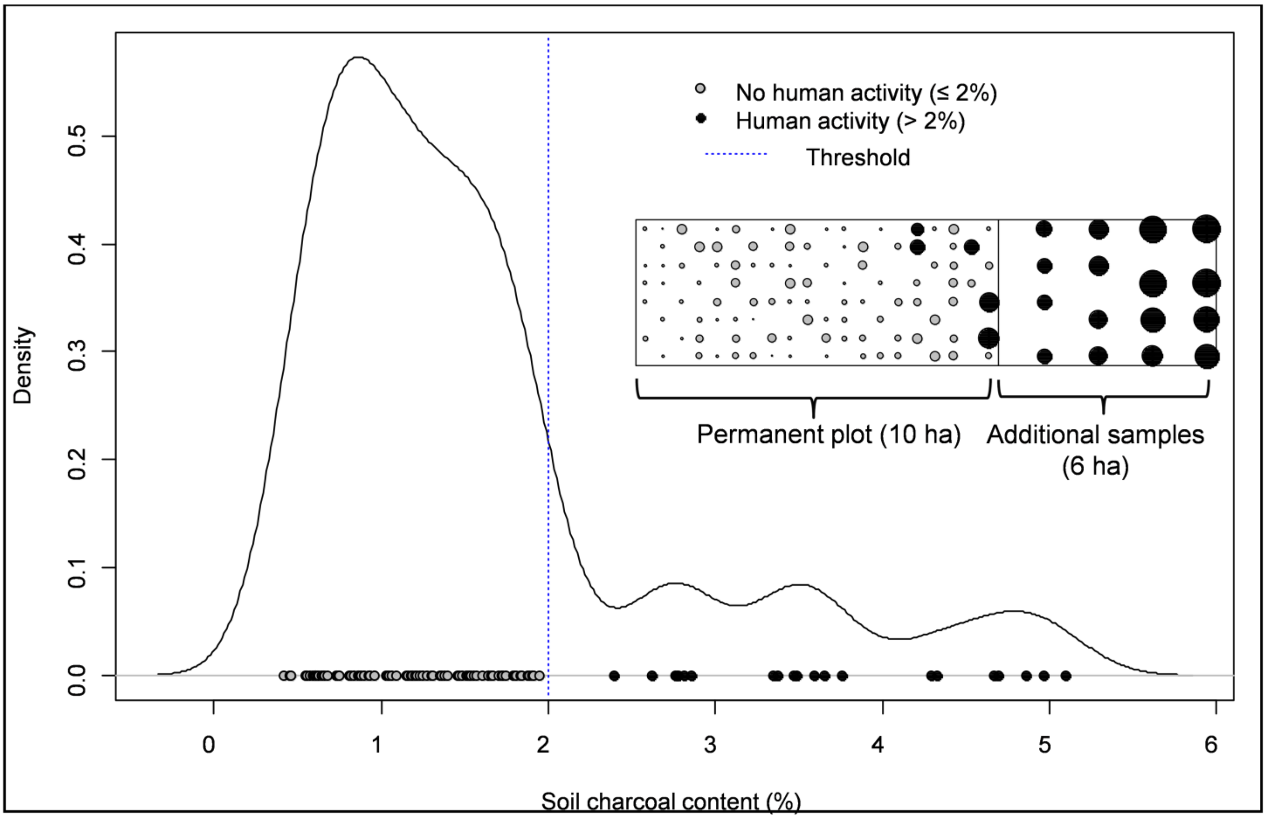
Yamazaki, M., S. Iwamoto, and K. Seivva. 2009. Distance- and density-dependent seedling mortality caused by several diseases in eight tree species co-occurring in a temperate forest. Plant Ecology 201:181–196.

Annexes

I. Chapitre I : Supporting information

The supplementary material presented below is also available at *Journal of Plant Ecology* online.

Appendix S1: Frequency distribution of the soil charcoal content within the dynamic plot coulped with an additional sampled area presenting attested agricultural pracitices. We set a threshold of charcoal content on the basis of the departure from the Gaussian distribution. This limit separates natural burning to anthropogenic slash and burn activities. Filled circles above the absciss correspond to the measured charcoal content. Gray circles: ≤ 2% values; black circles: > 2% values. The left-handed and right-handed rectangles below the legend represent the forest dynamic plot and the additional sampled area, respectively. The filled circles within the rectangles represent the sampled quadrats and have a size equal to their value of charcoal content (min = 0.4%, max = 5.1%). The colour code of the circles is the same as described above.



Appendix S2: Table of soil factors in the five habitats defined by MRT (*n:* number of quadrats). *P*-values of general Kruskal-Wallis tests between habitats: * < 0.05, ** < 0.01, *** < 0.001. Mean values followed by the same letter are not significantly different (Kruskal-Wallis test).

		Descriptive statistics of the whole dynamic plot			***MRT-habitat comparison (mean + sd)***				
Soil variables	**Unit**	**Mean + sd**	**Min**	**Max**	***H1* (*n* = 37)**	***H2* (*n* = 35)**	***H3* (*n* = 29)**	***H4* (*n* = 28)**	***H5* (*n* = 31)**
Bulk density	-	1.36 ± 0.10 ***	1.07	1.55	1.35 ± 0.05 a	1.35 ± 0.03 a	1.36 ± 0.02 a	1.32 ± 0.02 b	1.32 ± 0.02 b
Stoniness index	-	0.63 ± 0.59 ***	0.00	2.00	1.39 ± 0.28 a	0.95 ± 0.43 b	0.57 ± 0.46 c	0.17 ± 0.35 d	0.10 ± 0.13 d
Clay	%	26 ± 9 ***	13	44	19 ± 6 c	19 ± 2 d	21 ± 3 b	35 ± 9 a	34 ± 8 a
CEC	cmol$_c$/kg	1.97 ± 0.72 ***	0.86	3.44	1.51 ± 0.76 c	1.67 ± 0.08 b	1.62 ± 0.10 b	1.89 ± 0.23 a	1.99 ± 0.15 a
Al-Sat	%	54.7 ± 5.9 ***	40.1	66.3	59.2 ± 9.8 a	54.6 ± 3.1 b	60.5 ± 3.1 a	51.4 ± 5.2 c	47.2 ± 4.7 d
EC	µS	26.4 ± 9.3 ***	11.0	43.7	28.5 ± 8.1 a	24.8 ± 5.0 c	23.7 ± 5.1 d	23.3 ± 7.0 bc	24.2 ± 9.6 b
C/N	-	12.92 ± 0.92 ***	11.17	15.31	13.01 ± 0.32 a	12.96 ± 0.34 a	12.49 ± 0.23 b	12.64 ± 0.31 b	12.89 ± 0.19 a
OM	%	3.8 ± 1.1 ***	2.0	8.8	3.5 ± 0.6 b	3.1 ± 0.3 d	3.3 ± 0.3 c	4.1 ± 0.4 a	3.6 ± 0.3 b
E4/E6	-	5.72 ± 0.71	3.14	7.30	5.79 ± 0.2	5.79 ± 0.03	5.8 ± 0.02	5.8 ± 0.03	5.8 ± 0.02
pH$_{H2O}$	-	5.05 ± 0.25 ***	4.50	5.53	4.88 ± 0.19 c	5.05 ± 0.16 b	4.91 ± 0.18 c	5.07 ± 0.20 b	5.14 ± 0.23 a
pH$_{KCl}$	-	4.44 ± 0.20 ***	4.13	5.30	4.41 ± 0.15 b	4.47 ± 0.12 ab	4.35 ± 0.12 c	4.33 ± 0.08 c	4.49 ± 0.25 a
Al$_{exch.}$	cmol$_c$/kg	0.77 ± 0.26 ***	0.00	1.67	0.73 ± 0.25 c	0.79 ± 0.23 c	0.84 ± 0.28 a	0.83 ± 0.36 b	0.70 ± 0.20 d
DpH	-	-0.60 ± 0.22 ***	-1.30	-0.05	-0.46 ± 0.21 a	-0.58 ± 0.19 b	-0.55 ± 0.22 ab	-0.74 ± 0.22 c	-0.65 ± 0.20 c
P$_{EDTA}$	µg/g	2.77 ± 1.64 ***	0.31	4.87	3.1 ± 0.2 a	2.82 ± 0.21 b	2.58 ± 0.09 c	2.33 ± 0.32 d	2.19 ± 0.32 e
P$_{Olsen}$	µg/g	2.2 ± 1.5 ***	0.0	10.7	2.3 ± 0.4 a	2.2 ± 0.1 b	2.2 ± 0.1 ab	1.7 ± 0.3 c	1.8 ± 0.3 c
Ca$_{avail}$	µg/g	10 ± 7 ***	15	214	8 ± 4 b	13 ± 8 a	7 ± 2 c	10 ± 8 b	12 ± 9 b
Mg$_{avail}$	µg/g	22 ± 7 ***	13	100	12 ± 4 c	19 ± 8 b	13 ± 6 c	19 ± 6 b	24 ± 6 a
K$_{avail}$	µg/g	170 ± 70 ***	67	317	115 ± 55 d	120 ± 10 c	135 ± 19 b	200 ± 45 a	210 ± 45 a
Al$_{avail}$	µg/g	340 ± 110 ***	151	529	270 ± 50 d	280 ± 10 c	290 ± 20 b	370 ± 60 a	370 ± 50 a
Fe$_{avail}$	µg/g	117 ± 43 ***	36	249	129 ± 11 a	122 ± 7 b	114 ± 4 c	102 ± 12 d	96 ± 13 e
Mn$_{avail}$	µg/g	37 ± 38 ***	4	93	12 ± 7 d	24 ± 8 b	18 ± 8 c	35 ± 22 a	36 ± 18 a
Zn$_{avail}$	µg/g	1.06 ± 0.44 ***	0.45	1.64	0.96 ± 0.14 abc	1.02 ± 0.13 a	0.99 ± 0.10 ab	0.89 ± 0.14 bc	0.88 ± 0.17 c

B_{avail}	µg/g	0.12 ± 0.12 ***	0.05	0.22	0.09 ± 0.03 b	0.11 ± 0.02 a	0.1 ± 0.01 ab	0.09 ± 0.01 b	0.10 ± 0.02 a
Ca_{tot}	µg/g	33 ± 13 ***	4	127	58 ± 12 ab	19 ± 14 c	55 ± 32 b	77 ± 22 a	17 ± 9 c
Mg_{tot}	µg/g	26.8 ± 4.3 ***	15.2	36.4	24.9 ± 4.2 b	26.2 ± 3.6 b	23.4 ± 4.9 b	30.4 ± 3.4 a	29.6 ± 2.6 a
K_{tot}	µg/g	2280 ± 645 ***	1112	4783	2630 ± 210 a	2380 ± 150 b	2240 ± 70 c	2100 ± 160 d	2010 ± 160 e
Al_{tot}	µg/g	910 ± 220 ***	514	1722	860 ± 120 b	800 ± 110 bc	750 ± 200 c	1130 ± 300 a	1060 ± 170 a
Fe_{tot}	µg/g	32420 ± 9780 ***	16353	76757	36940 ± 3490 b	30590 ± 3440 d	33560 ± 4970 c	40530 ± 4060 a	37280 ± 4760 b
Mn_{tot}	µg/g	75 ± 34 ***	18	416	49 ± 10 c	60 ± 8 b	59 ± 10 b	106 ± 41 a	110 ± 42 a
Mo_{tot}	µg/g	0.94 ± 0.61 ***	0.72	1.10	0.89 ± 0.06 b	0.84 ± 0.08 c	0.98 ± 0.08 a	0.97 ± 0.07 a	0.91 ± 0.05 b
P_{tot}	µg/g	190 ± 50 ***	87	378	187 ± 20 b	158 ± 12 c	161 ± 23 c	220 ± 31 a	210 ± 33 a
Zn_{tot}	µg/g	23 ± 12 ***	6	50	16 ± 2 c	16 ± 1 bc	17 ± 2 b	25 ± 6 a	26 ± 6 a
B_{tot}	µg/g	60.2 ± 9.7 ***	33.3	96.7	58.5 ± 0.7 b	58.1 ± 3.4 b	54.2 ± 2.2 c	60.2 ± 2.5 a	60.5 ± 2.5 a
charcoal	%	1.24 ± 0.6 ***	0.42	3.65	1.05 ± 0.46 a	1.04 ± 0.45 a	1.09 ± 0.48 a	1.51 ± 0.64 ab	1.73 ± 0.75 b

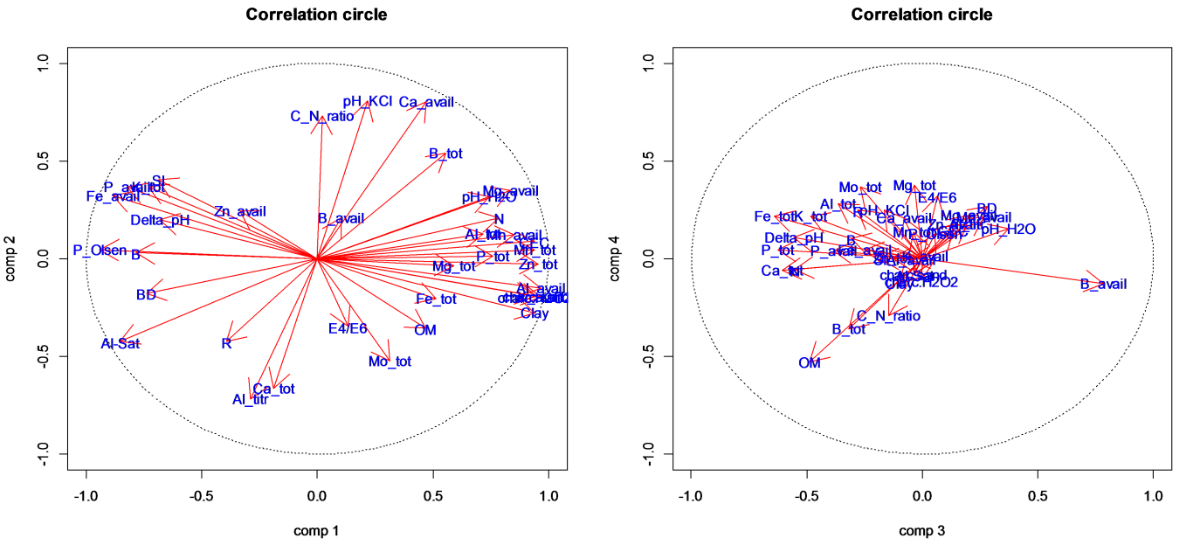
Appendix S3: Table of the Pearson's correlation coefficients between soil variables. The coefficients were tested for significance using toroidal randomizations (9999 iterations). Since the '*Man*' variable is a binary variable, the coefficient of intra-class correlation was used to compute correlation between this variable and the quantitative variables. Codes of significance level: *** for P<=0.001 and >=4999 replicates; ** for P<=0.01 and >=499 replicates; * for P<=0.05 and >=99 replicates (bilateral tests).

	SI	**BD**	**pH$_{H2O}$**	**OM**	**C/N**	**P$_{Olsen}$**	**E4/E6**	**CEC**	**Al-Sat**	**Al$_{avail}$**	**B$_{avail}$**	**Ca$_{avail}$**	**Fe$_{avail}$**	**K$_{avail}$**	**Mg$_{avail}$**	**Mn$_{avail}$**	**P$_{EDTA}$**
SI																	
BD	0.46																
pH$_{H2O}$	-0.39	-0.45															
OM	-0.37	-0.63	-0.01														
C/N	0.33	-0.33	0.11	-0.02													
P$_{Olsen}$	0.60 *	0.66 *	-0.61	-0.51 *	-0.12												
E4/E6	-0.18	0.13	0.01	0.05	-0.15	-0.17											
CEC	-0.67 *	-0.71 **	0.76 **	0.28	0.09	-0.86 **	0.10										
Al-Sat	0.46	0.64 *	-0.78 **	-0.19	-0.21	0.76 *	0.02	-0.83 **									
Al$_{avail}$	-0.71	-0.68 *	0.62 *	0.49	0.00	-0.95 **	0.18	0.89 **	-0.74 *								
B$_{avail}$	0.04	0.12	0.47	-0.30	0.15	-0.13	0.06	0.19	-0.15	0.01							
Ca$_{avail}$	0.01	-0.44	0.56 *	-0.16	0.57	-0.44	-0.15	0.56	-0.75 **	0.36	0.13						
Fe$_{avail}$	0.82 **	0.58	-0.63 **	-0.44	0.19	0.80 *	-0.21	-0.85 *	0.61	-0.90 *	-0.17	-0.10					
K$_{avail}$	-0.78 *	-0.68	0.63 **	0.49	-0.09	-0.90 **	0.19	0.90 **	-0.74 **	0.97 **	0.03	0.29	-0.95 *				
Mg$_{avail}$	-0.52	-0.60 *	0.79 **	0.08	0.16	-0.74 **	0.02	0.92 **	-0.90 **	0.74 **	0.28	0.74 **	-0.67 *	0.75 **			
Mn$_{avail}$	-0.63 **	-0.57 *	0.78 **	0.17	0.01	-0.75 **	0.10	0.91 **	-0.84 **	0.77 **	0.25	0.54	-0.76 *	0.78 **	0.88 **		
P$_{EDTA}$	0.81 **	0.48	-0.62 **	-0.36	0.25	0.76 *	-0.21	-0.81 *	0.54	-0.84 *	-0.27	-0.05	0.98 **	-0.89 *	-0.64 *	-0.75 **	
Zn$_{avail}$	0.24	0.16	-0.16	-0.30	0.14	0.35 *	-0.05	-0.20	0.13	-0.38 *	0.10	0.04	0.34	-0.36	-0.11	-0.01	0.31
Clay	-0.75 *	-0.70	0.55	0.63 *	-0.10	-0.91 **	0.19	0.82 *	-0.66	0.96 **	-0.03	0.18	-0.91 *	0.96 **	0.62	0.69 *	-0.84 *
Al$_{tot}$	-0.27	-0.58	0.48	0.27	0.12	-0.65 *	0.07	0.65	-0.61 *	0.69 **	-0.16	0.48	-0.47	0.65	0.57	0.54	-0.37
B$_{tot}$	-0.10	-0.64	0.41	0.40	0.50	-0.58 *	-0.19	0.45	-0.69 **	0.44	0.02	0.66 *	-0.19	0.37	0.49 *	0.41	-0.10
Ca$_{tot}$	0.05	0.07	-0.58 *	0.49 *	-0.42 *	0.06	0.11	-0.32 *	0.44 **	-0.06	-0.52 *	-0.55 **	0.15	-0.08	-0.48 *	-0.37	0.17
Fe$_{tot}$	-0.26	-0.33	0.19	0.52 *	-0.22	-0.49	0.17	0.35	-0.37	0.51	-0.40 *	0.12	-0.31	0.50	0.26	0.28	-0.21
K$_{tot}$	0.77 **	0.44	-0.55	-0.36	0.23	0.68	-0.18	-0.70 *	0.47	-0.72	-0.35	0.06	0.92 **	-0.80	-0.53	-0.65 *	0.95 **

Mg$_{tot}$	-0.25	-0.34	0.51 *	0.13	-0.06	-0.49 -*	0.23	0.53 *	-0.48 *	0.53 **	0.00	0.25	-0.49	0.55 **	0.46	0.50 *	-0.44
Mn$_{tot}$	-0.64 *	-0.65 **	0.67 **	0.35	0.00	-0.88 -**	0.08	0.92 **	-0.84 **	0.90 **	0.05	0.56	-0.77 *	0.88 **	0.86 **	0.86 **	-0.73 *
Mo$_{tot}$	-0.34	-0.15	0.08	0.30	-0.43	-0.25	0.17	0.25	-0.06	0.36	-0.20	-0.18	-0.34	0.38	0.10	0.20	-0.31
P$_{tot}$	-0.36	-0.68	0.37	0.62 **	0.11	-0.75 -**	0.09	0.62	-0.63	0.76 *	-0.36	0.42	-0.49	0.71	0.51	0.48	-0.37
Zn$_{tot}$	-0.67	-0.72 **	0.60 *	0.49	0.04	-0.93 -**	0.14	0.89 **	-0.80 **	0.94 **	-0.03	0.48	-0.81 *	0.92 **	0.79 **	0.78 **	-0.74 *
Man	0.34	0.25	0.35 *	-0.09	-0.03	0.47	0.31 *	0.48 *	0.33	0.56 *	0.02	0.21	0.53 **	0.51 *	0.45 *	0.32	0.48 *

	Zn$_{avail}$	**Clay**	**Al$_{tot}$**	**B$_{tot}$**	**Ca$_{tot}$**	**Fe$_{tot}$**	**K$_{tot}$**	**Mg$_{tot}$**	**Mn$_{tot}$**	**Mo$_{tot}$**	**P$_{tot}$**
Clay	-0.42 **										
Al$_{tot}$	-0.30	0.64									
B$_{tot}$	-0.09	0.44	0.48								
Ca$_{tot}$	-0.10	0.05	0.00	-0.23							
Fe$_{tot}$	-0.23	0.56	0.60 *	0.34	0.45						
K$_{tot}$	0.28	-0.78 *	-0.23	-0.12	0.21	-0.05					
Mg$_{tot}$	-0.37 *	0.52 *	0.70 **	0.18	-0.07	0.35	-0.36				
Mn$_{tot}$	-0.24	0.84 **	0.68 *	0.55	-0.17	0.53 *	-0.60	0.48 *			
Mo$_{tot}$	-0.07	0.41	0.30	-0.10	0.43 *	0.55 *	-0.20	0.25	0.37		
P$_{tot}$	-0.32	0.76	0.77 *	0.61	0.24	0.85 **	-0.21	0.46	0.74 *	0.42	
Zn$_{tot}$	-0.32	0.92 **	0.69 *	0.56	-0.09	0.56 *	-0.64	0.47 *	0.95 **	0.35	0.79 **
Man	-0.05	0.44	0.21	0.04	0.27 *	-0.08	0.35	0.09	0.35	-0.04	0.23

Appendix S4: Correlation circles of the first four principal components of the PCA of soil parameters.



II. Chapitre II : Supporting information

The supporting information is available at FEMSEC online:
https://academic.oup.com/femsec/article/92/10/fiw151/2197817#82439523

Appendix S1: For each soil sample, a stoniness index (SI) was estimated on the field after sieving (2-mm mesh) by a discreet quantitative index taking the values of 0 (null SP), 1 (half of the sieve area or less covered by gravels) or 2 (all the sieve area covered). Soil texture was determined by wet sieving and the pipette method after OM destruction with H_2O_2 and clay dispersion by Na citrate. The pH-H_2O and the electrical conductivity (EC) were respectively measured with glass electrodes (Mettler-Toledo) and a conductimeter (VWR EC300) in a ratio 1:5 between soil and water. The pH-KCl and exchangeable Al_{exch} were determined in a soil extract (1:5 ratio) in 1 M KCl and measuring the derivative of the titration curves for Al_{exch} (Radiometer Copenhagen TIM900). The ΔpH was obtained by calculating the difference between pH-KCl and pH-H_2O. The plant-available elements ((Al, B, Ca, Fe, K, Mg, Mn, P-EDTA and $Zn)_{avail}$) were extracted with 0.5 M ammonium acetate 0.03 M EDTA at pH 4.65 and measured by inductively coupled plasma optical emission spectroscopy (ICP-OES) with CCD detector (Varian, Vista MPX). Bioavailable phosphorus (P-Olsen) was extracted with Na bicarbonate and determined by colorimetry. Total forms of elements (Al, B, Ca, Fe, K, Mg, Mn, Mo, P and $Zn)_{tot}$ were dissolved in Teflon vials by tri-acid (HCl-HNO_3-HF) attack of finely ground soil samples on a hot plate. The dry residue was re-dissolved in HNO_3 and total element concentrations were determined by ICP-OES. The effective cation exchange capacity (ECEC) was calculated as the sum of exchangeable Ca, K, Mg concentrations and titrated Al (Al_{exch}), expressed in $cmol_c$ kg^{-1}. The Al^{3+} saturation rate (Al-Sat) of the exchange complex corresponds to the proportion of Al^{3+} on the total ECEC. Organic matter content (OM) was calculated by mass loss of a sample after dry ashing at 550 °C. The extinction coefficient in visible light (E4/E6) allows to know the relative importance of humic and fulvic acids in the soil OM and is related to the humification stage. This coefficient was obtained measuring the absorbance at 465 and 665 nm of a soil extract with 0.5 M NaOH during 16 h after centrifugation (10000 rnd/min). The soil nitrogen content (N) and carbon-to-nitrogen ratio (C/N) were computed after measuring soil nitrogen and carbon contents by flash combustion at 1350 °C in a CN elemental analyser (Dumas method, ISO 10694).

Appendix S2: Before running MEM analysis, community data were detrended as advised by Legendre and Legendre (2012). The spatial gradients, if any, were analysed separately. Numerous spatial weighting matrices *W* were tested as recommended by Dray et al. (2006). Two types of connectivity matrices were systematically tested: 1) a connectivity matrix based on the Delaunay triangulation scheme between the sampled quadrats and 2) a connectivity matrix based on distances (radius around a quadrat). For the latter, 20 radius values were tested, ranging from the highest distance of the minimum spanning tree to the distance corresponding to the maximum value of spatial correlation in the community. For each one of these connectivity matrices, two weighting matrices were tested: 1) $f_1 = 1 - (d/dmax)^y$ and 2) $f_2 = 1/d^y$ where d = the Euclidean distance between both considered quadrats, $dmax$ = the maximum Euclidean distance of the connectivity matrix, and y = a parameter tested for values ranging from 1 to 10. Connectivity matrices were also tested without weighting matrix. The selected final spatial weighting matrix was the one presenting the lowest AIC_c value, following Dray *et al.*, 2006.

Figure S1: The permanent plot consists of 160 quadrats of 25 × 25 m, represented on the right-hand scheme by the small squares. Grey squares represent the 34 quadrats in which the fungal EM community was sampled. The dots in the right-hand scheme illustrate the 510 soil samples taken from 102 quadrats.

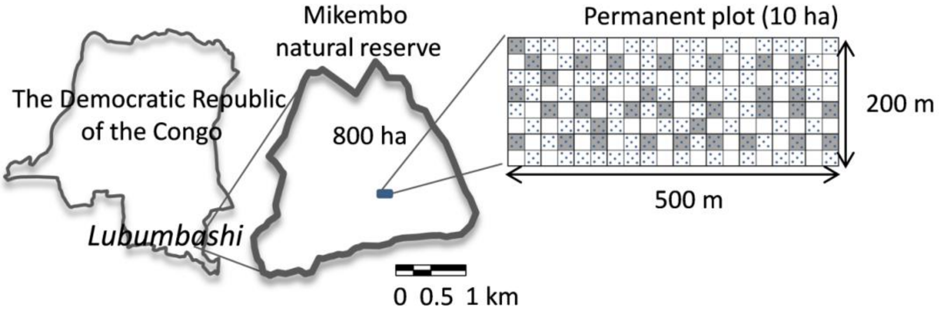


Figure S2: Fitted site score triplots of the redundancy analysis of the Hellinger-transformed EM fungal community constrained by the selected soil variables (Fig. S2a and S2b) and host FT variables (Fig. S2c) - scaling 2. The bottom and left-hand scales are for the quadrats and the species, the top and right-hand scales are for the soil variables. Soil abbreviations are presented in Appendix S1.

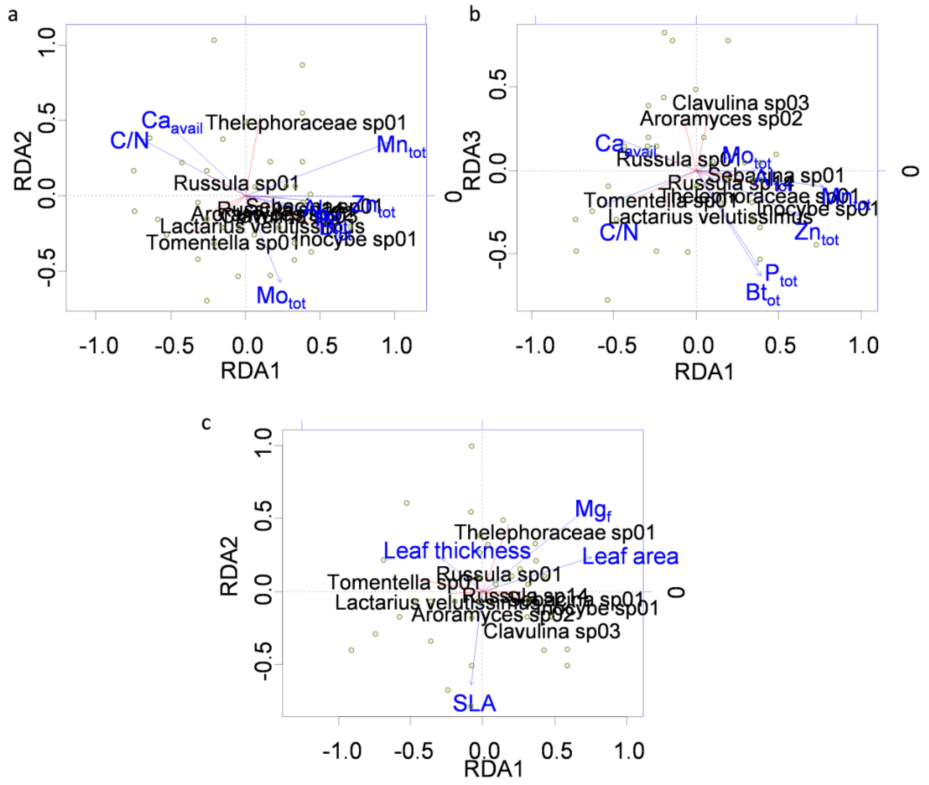


Figure S3: Phylogenetic tree of host trees constructed with Phylomatic (Webb et al., 2005) and Phylocom (Webb et al., 2008) and representation of the values of leaf area, specific leaf area, and foliar Mg content. The circle sizes next to each species are proportional to the value of the corresponding functional trait. Functional traits were not measured for species displaying no symbol (least abundant species). Similar trait values are observed within different genera and families, indicating no trait conservatism. For instance, *Brachystegia spiciformis* and *B. wangermeeana* (Fabaceae) present very different leaf areas (93 and 34 cm², respectively) while the most similar species to *B. wangermeeana* for this trait is *Marquesia macroura* (25 cm2; Dipterocarpaceae). Similarly, *Julbernardia paniculata* has a much greater leaf area than *J. globiflora* (166 and 71 cm2, respectively; Fabaceae), which itself has a leaf area much more similar to that of *Uapaca nitida* (74 cm2; Phyllanthaceae).The same trend can be observed for the SLA, with for instance *B. wangermeeana* having a SLA closer to that of *Uapaca nitida* than *B. spiciformis* (113, 110 and 103 cm^{2}g^{-1}, respectively); *n* indicates the abundance of the species in the plot.

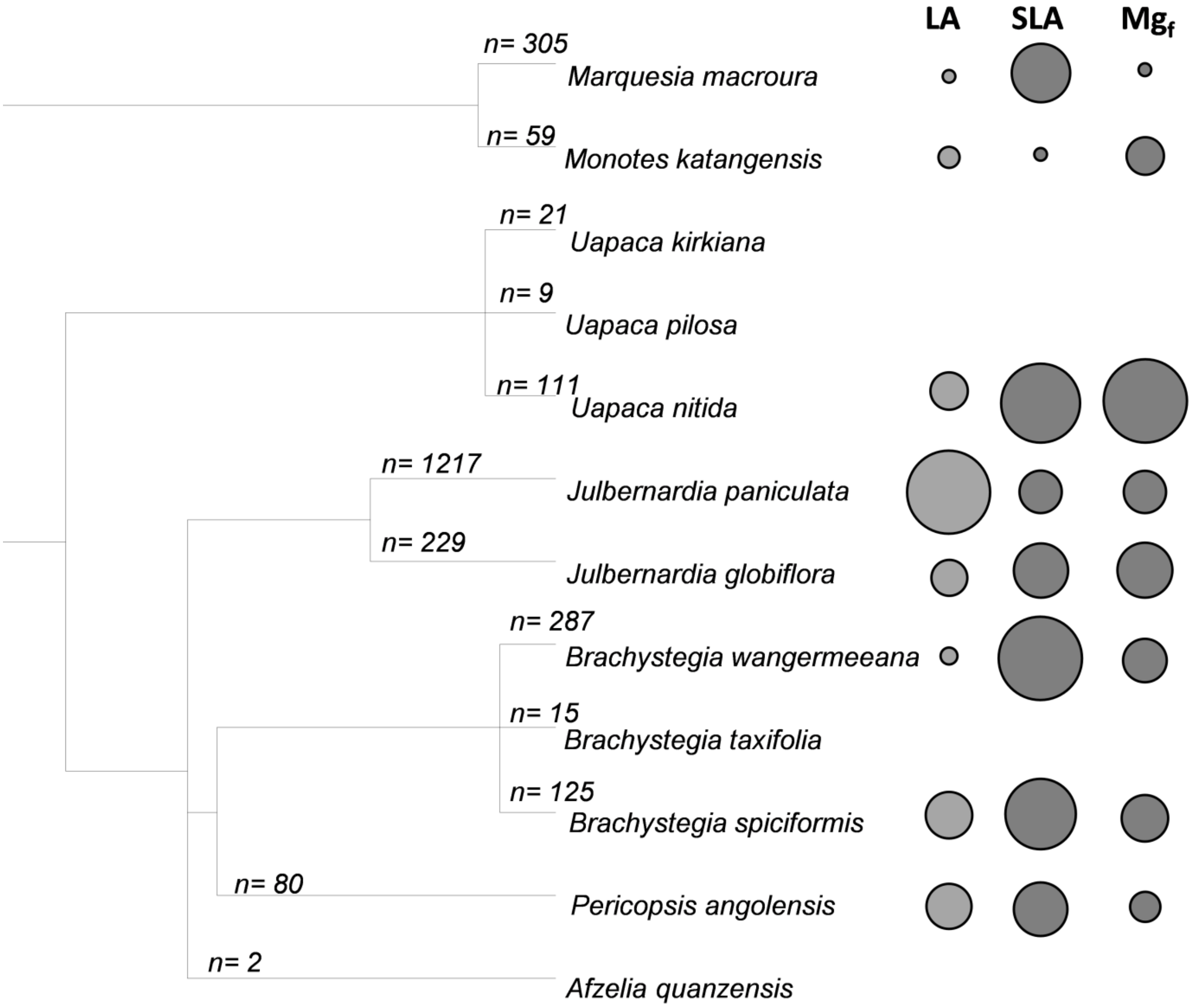


Table S1: New accession numbers for OTUs of the belowground ectomycorrhizal community.

Table S1 is too big to be displayed here. It is available at FEMSEC online.

Table S2: Table of soil variables, displaying general descriptive statistics as well as spatial parameter values after fitting semivariogram models to the empirical semivariograms (*gstat* package). The semivariograms models used for the fitting were: Gaussian (Gau), Spherical (Sph), Wave (Wav), Matern (Mat), Linear (Lin). Soil variable abbreviations are presented in supplementary Appendix S1. The λ parameter of each variable characterisises its box-cox transformation.

			Descriptive statistics				**Fitting models to empirical semivariograms**			
Soil variables	**Unit**	**λ**	**Mean**	**sd**	**Min**	**Max**	**Model**	**Range (m)**	**Nugget**	**psill**
BD		4.20	1.4	0.1	1.1	1.5	Sph	108	0.02	0.03
SI		-5.23	0.6	0.6	0.0	2.0	Sph	54	0.26	0.16
Clay	%	-0.39	26	9	13	44	Gau	379	0.00	0.03
EC	µs	-1.96	26	9	11	44	Wav	200	0.00	0.00
pH_{H2O}		-2.60	5.1	0.3	4.5	5.5	Sph	100	0.00	0.00
pH_{KCl}		-9.36	4.4	0.2	4.1	5.3	Wav	200	0.00	0.00
ΔpH		-3.12	-0.6	0.2	-1.3	0.0	Gau	150	0.00	0.00
$Al_{exch.}$	$cmol_c/kg$	-4.24	0.8	0.3	0.0	1.7				
Al_{avail}	µg/g	-0.09	343	110	151	529	Gau	306	0.01	0.03
B_{avail}	µg/g	-0.40	0.1	0.1	0.1	0.2	Gau	94	0.43	0.29
Ca_{avail}	µg/g	-0.49	97.0	358.3	15.3	213.5	Gau	113	0.01	0.00
Fe_{avail}	µg/g	1.29	117	43	36	249				
K_{avail}	µg/g	1.29	172	70	67	317	Wav	400	0.01	0.04
Mg_{avail}	µg/g	1.29	56	46	13	100	Gau	176	0.03	0.03
Mn_{avail}	µg/g	1.29	37.3	38.2	3.9	93.0	Gau	174	0.06	0.09
P_{avail}	µg/g	1.29	2.8	1.6	0.3	4.9	Wav	300	1.50	1.00
Zn_{avail}	µg/g	-0.04	1.1	0.4	0.5	1.6	Gau	190	0.04	0.02
P_{Olsen}	µg/g	0.36	2.2	1.5	0.0	10.7				
Al_{tot}	µg/g	-0.16	2277	7923	168	60561	Gau	171	0.03	0.01
B_{tot}	µg/g	0.41	60	10	33	97	Gau	68	0.43	0.16
Ca_{tot}	µg/g	0.52	33	13	-2	96				
Fe_{tot}	µg/g	0.16	32417	9778	16353	76757	Sph	240	1.14	1.14
K_{tot}	µg/g	-0.33	2284	646	1112	4783	Gau	339	0.00	0.00
Mg_{tot}	µg/g	-0.16	65	237	-23	1612	Wav	250	0.36	0.05
Mn_{tot}	µg/g	-0.25	103	86	18	416	Gau	527	0.03	0.14
Mo_{tot}	µg/g	-0.53	0.9	0.6	0.4	4.2	Wav	110	0.24	0.04
P_{tot}	µg/g	0.24	190	50	87	378	Gau	177	0.50	0.84
Zn_{tot}	µg/g	0.15	23.4	12.1	5.6	50.5	Gau	789	0.44	2.75
ECEC	$cmol_c/kg$	-0.40	2.0	0.7	0.9	3.4	Gau	1542	0.03	0.47
Al-Sat	%	1.34	42	13	0	72	Sph	325	1158	995
N	µg/g	-1.28	0.08	0.01	0.05	0.12	Gau	197	6	23

C/N		-1.94	12.9	0.9	11.2	15.3	Wav	220	0.00	0.00
OM	%	-0.39	3.8	1.1	2.0	8.8	Wav	200	0.02	0.01
E4/E6		2.74	5.7	0.7	3.1	7.3	Wav	400	207	46

Table S3: Table of the host species functional traits, displaying general descriptive statistics as well as parameter values after fitting semivariogram models to the empirical semivariograms (gstat package). The semivariograms models used for the fitting were: Gaussian (Gau), Spherical (Sph), Wave (Wav), Matern (Mat), Linear (Lin). Range1 correspond to the range estimated by the model, while range2 was estimated visually when a second plateau was reached further than the first one. The statistics and parameter estimates were computed based on the mean values per quadrat of each trait weighted by the abundances of host species.

		Descriptive statistics				**Fitting models to empirical semivariograms**				
FT variables	**Unit**	**Mean**	**sd**	**Min**	**Max**	**Model**	**Range 1 (m)**	**Range 2 (m)**	**Nugget**	**psill**
SLA	cm²/g	92.7	5.4	81.8	108.7	Sph	44	180	75.21	29.76
Leaflet area	cm²	23.2	8.5	2.8	44.2	Sph	53	300	109	36
Leaf area	cm²	97.9	33.1	39.1	165.8	Wav	100	-	1325	216
Max height	m	21.1	1.4	18.0	24.4	Wav	220	-	6.86	2.08
Mean annual increment	cm/year	0.5	0.1	0.2	0.6	Gau	76	-	0.01	0.00
Leaf decomposability	% g/day	0.61	0.01	0.58	0.63	Sph	26	-	0.00	0.00
Al_f	µg/g	24.0	8.4	13.1	40.6	Sph	35	-	75.06	34.79
B_f	µg/g	14.6	1.3	12.0	18.3	Sph	161	400	15.23	4.95
Ca_f	µg/g	4930	1310	2144	7020	Sph	40	370	2.E+06	6.E+05
Fe_f	µg/g	28.3	5.3	21.1	38.4	Sph	36	470	27.31	18.01
K_f	µg/g	8514	1105	6571	10696	Sph	31	470	1.E+06	8.E+05
Mg_f	µg/g	2512	263	1828	3168	Wav	80	-	1.E+05	3.E+04
Mn_f	µg/g	124	35	62	206	Wav	150	-	1454	341
P_f	µg/g	1553	91	1386	1713	Sph	34	450	7315	7552
Zn_f	µg/g	15.9	2.0	12.4	20.2	Sph	24	-	2.52	2.78
N	%	2.0	0.1	1.8	2.3	Sph	26	-	0.03	0.02
C/N ratio		25.0	1.4	22.3	27.9	Sph	48	400	5.08	1.48
Wood density	g/cm³	0.7	0.0	0.6	0.7	Sph	187	400	0.00	0.00
Leaf dry matter content	mg/g	481.1	36.6	391.2	528.8	Sph	56	250	1519	654
Leaf thickness	cm	0.026	0.002	0.019	0.030	Sph	52	400	0.00	0.00
Foliar lignin content	% g	34.8	1.3	32.6	38.2	Sph	30	-	2.11	3.36
Bark thickness	cm	0.9	0.1	0.8	1.1	Sph	34	470	0.00	0.00
Seed mass	mg	342.2	90.4	117.5	507.1	Wav	199	-	1.E+05	2.E+04

Table S4: Table of pairwise Pearson correlations between selected soil variables (rows) and the remaining soil variables presenting at least one significant correlation. Significance test were carried out by 999 torus-translation permutations. Significance level is illustrated by * (p < 0.05) and ** (p < 0.01). The last column gives the total number of variables significantly correlated. Variable abbreviations are presented in the corresponding Material and methods sections. Soil variable abbreviations are presented in Appendix S1.

	BD	SI	Clay	EC	pH_H₂O	pH_KCl	DpH	Al_{avail}	B_{avail}	Ca_{avail}	Fe_{avail}	K_{avail}	Mg_{avail}	Mn_{avail}	P_{avail}	P_{Olsen}
C/N	-0.32	0.34	-0.12	0.27	0.10	0.40	0.20	-0.02	0.13	0.57	0.21	-0.12	0.15	-0.01	0.27	-0.10
Ca_{avail}	-0.44	0.01	0.19	0.28	0.56*	0.57**	-0.12	0.36	0.13		-0.10	0.29	0.74**	0.54	-0.05	-0.44
B_{tot}	-0.65	-0.10	0.45	0.45	0.42	0.25	-0.23	0.45	0.02	0.66*	-0.20	0.38	0.49**	0.42	-0.11	-0.58*
Mn_{tot}	-0.66**	-0.65	0.85**	0.00	0.68*	0.37	-0.61**	0.90**	0.05	0.56	-0.78*	0.88**	0.86**	0.86**	-0.73*	-0.89**
Mo_{tot}	-0.16	-0.34	0.41	-0.03	0.09	0.06	-0.14	0.36	-0.20	-0.18	-0.34	0.39	0.10	0.21	-0.32	-0.26
P_{tot}	-0.68	-0.37	0.76	0.43	0.37	0.40	-0.19	0.77*	-0.36	0.43	-0.50	0.72	0.52	0.48	-0.37	-0.76**
Zn_{tot}	-0.73**	-0.67	0.92**	0.07	0.61*	0.35	-0.58**	0.95**	-0.03	0.48	-0.81*	0.93**	0.79**	0.78**	-0.75*	-0.93**

	Al_{tot}	B_{tot}	Ca_{tot}	Fe_{tot}	K_{tot}	Mg_{tot}	Mn_{tot}	P_{tot}	N	ECEC	Al-Sat	OM	**N° corr. var.**
C/N	0.12	0.51	-0.43*	-0.24	0.26	-0.09	-0.02	0.10	0.29	0.08	-0.22	-0.02	**1**
Ca_{avail}	0.49	0.66*	-0.55**	0.13	0.06	0.26	0.56	0.43	0.57	0.39	-0.73**	-0.10	**6**
B_{tot}	0.49		-0.24	0.34	-0.12	0.18	0.55	0.61	0.77*	0.36	-0.65*	0.48	**4**
Mn_{tot}	0.68*	0.55	-0.18	0.53	-0.60	0.48*		0.74*	0.74*	0.67*	-0.87**	0.42	**16**
Mo_{tot}	0.31	-0.10	0.44*	0.55*	-0.20	0.26	0.37	0.42	0.26	0.28	-0.11	0.35	**2**
P_{tot}	0.77*	0.61	0.24	0.86**	-0.21	0.47	0.74**		0.96*	0.48	-0.70	0.75**	**6**
Zn_{tot}	0.69*	0.56	-0.10	0.56*	-0.64	0.48*	0.96**	0.80**	0.80*	0.61*	-0.86**	0.56	**18**

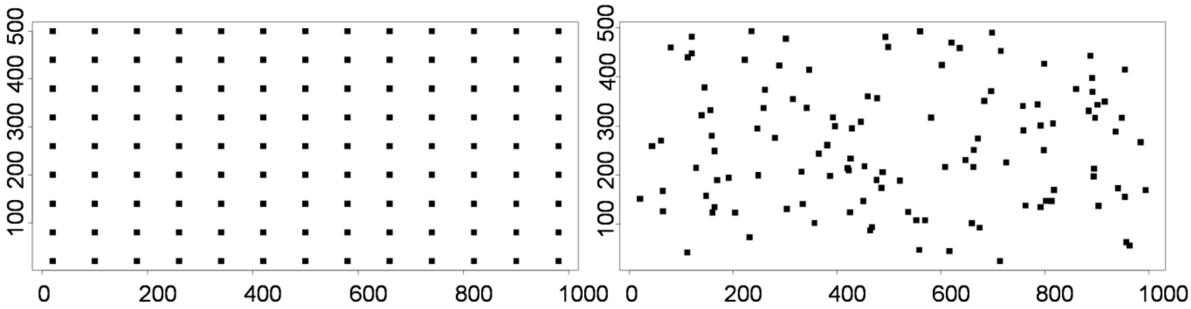
Table S5: Table of pairwise Pearson correlations between selected host tree FT variables (rows) and the remaining FT variables presenting at least one significant correlation. Significance test were carried out by 999 torus-translation permutations. Significance level is illustrated by * (p < 0.05) and ** (p < 0.01). The last column gives the total number of variables significantly correlated. Variable abbreviations are presented in the corresponding Material and methods sections.

	SLA	Leaf area	Leaflet area	Max height	Mean annual increment	Leaf decomposability	Al_f	B_f	Ca_f	Fe_f	K_f	Mn_f	P_f	Zn_f	N	C/N ratio	Leaf dry matter content	Leaf thickness	Foliar lignin	Bark thickness	Seed mass	**N° correlated variables**
SLA		-0.66**	-0.50**	0.02	-0.50**	-0.16	0.06	0.25	0.20	0.10	-0.51**	0.68**	-0.25	0.56**	0.58**	-0.62**	-0.03	-0.66**	0.45*	0.57**	0.26	**11**
Leaf area	-0.66**		0.30**	-0.04	0.22	0.09	-0.69**	-0.66**	0.37**	-0.67**	0.95**	-0.82**	0.51*	-0.78**	-0.34*	0.29	0.46*	0.47**	-0.70**	-0.86**	0.33**	**16**
Mg_f	0.01	0.19	0.02	-0.74**	-0.60**	-0.76*	-0.11	-0.25	0.44*	-0.02	0.14	-0.55**	0.15	0.15	0.12	-0.1	0.12	0.23	0.49	-0.40*	0.65**	**7**
Leaf thickness	-0.66**	0.47**	0.87**	-0.12	0.28	0.02	0.11	-0.33	-0.40*	0.06	0.41*	-0.57**	-0.08	-0.241635	-0.63**	0.75*	-0.41*		-0.33	-0.35	-0.08	**9**

III. Chapitre III : Supporting information

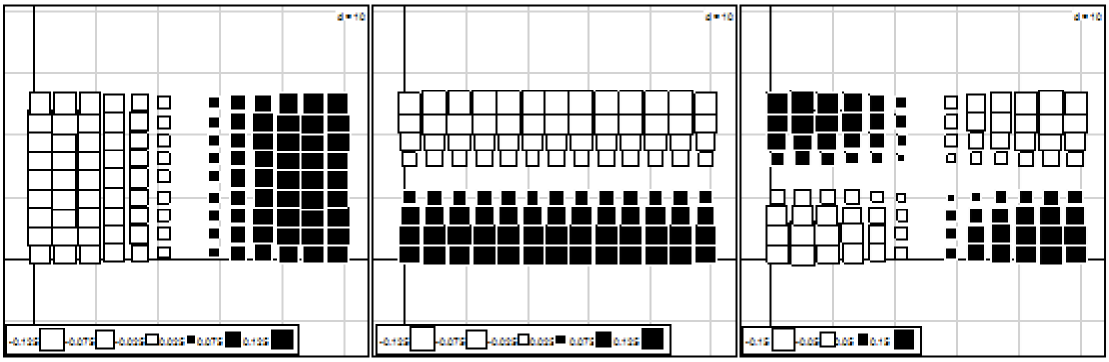
The supplementary material presented below is available at *Ecography* online: http://www.ecography.org/appendix/ecog-03380

Figure A1: Regular (left-hand) and random (right-hand) sampling designs used for the simulations. In both cases, 117 quadrats were used (black squares).

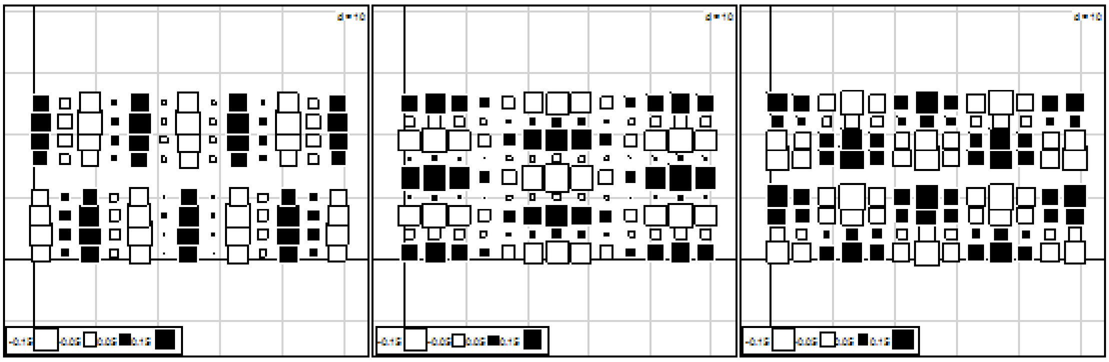


Appendix A1: Illustrations and methodological details of the generation of simulated structured species at broad, medium and fine scales.

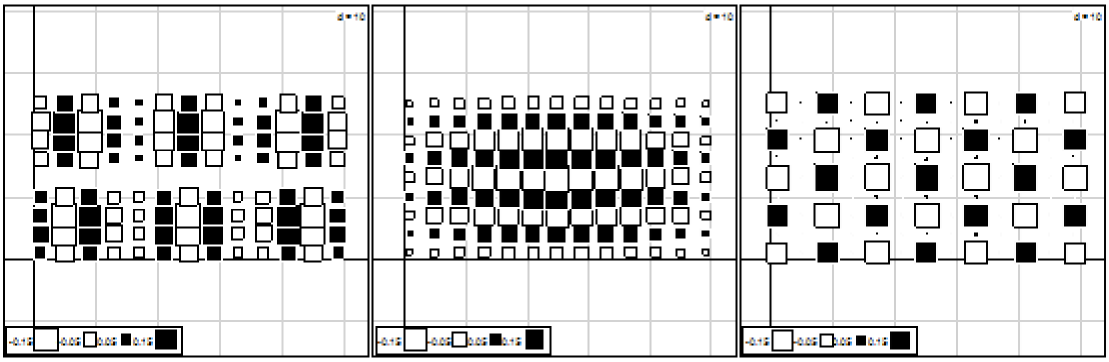
In order to test the power and a possible (adjusted R^2) bias of the spatial EV selection methods, positively autocorrelated response variables were constructed by linear combination of three MEM variables to which a random noise (mean of 0, standard deviation of 1) was added, following Jombart *et al.* (2009). The coefficient of each MEM is given in the legends of the figures below. Structured variables were built at broad, medium, and fine scales, using MEM 1 to 3, 25 to 27, and 56 to 58, respectively. The MEM used at the three scales and for the regular and random designs are illustrated below. An example of the resulting spatial patterns is also presented (for the regular design only, for brevity).



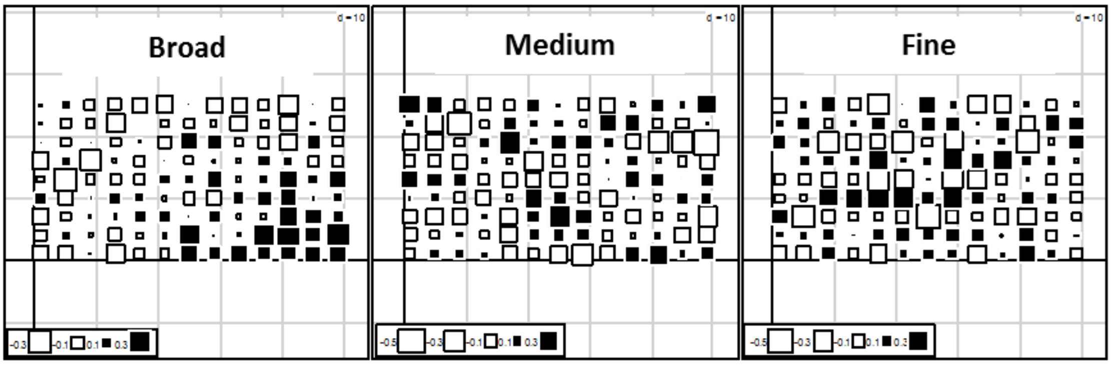
Appendix A1 Figure 6: MEM variables (1, 2 and 3, from the left to the right) used for constructing simulated species structured at a broad scale in a regular design. The linear combination coefficient of the MEM for the linear combination was 0.5 for the three variables. The bottom legend of each box relates each colour and size of the quadrats (white and black squares) to their coordinate on the corresponding eigenvector (see Introduction). Quadrats of the same colour and similar size are very alike (species abundance).



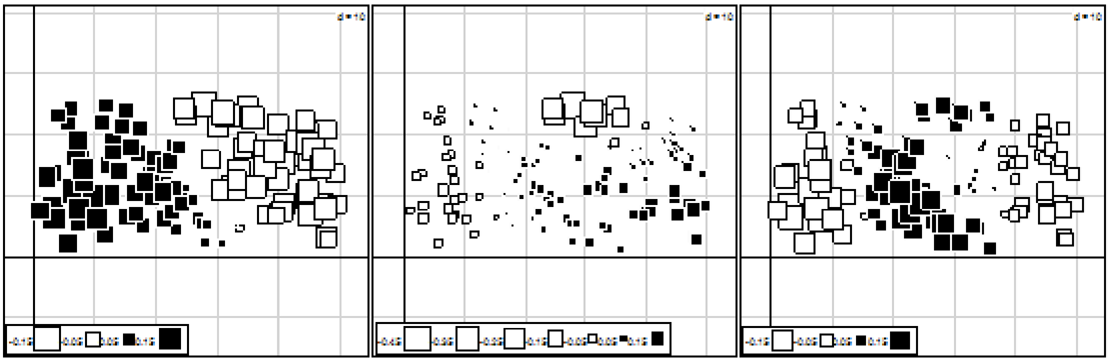
Appendix A1 Figure 7: MEM variables (25, 26 and 27, from the left to the right) used for constructing simulated species structured at a medium scale in a regular design. The linear combination coefficients of the MEM were 0.6, 1 and 0.8 for the three variables. See the Fig. 1 caption of the present appendix for the legend of the boxes.



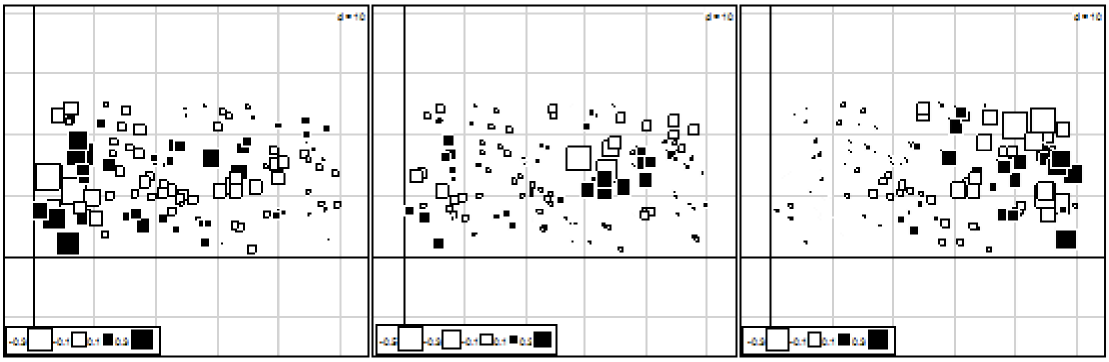
Appendix A1 Figure 8: MEM variables (56, 57 and 58, from the left to the right) used for constructing simulated species structured at a fine scale in a regular design. The linear combination coefficients of the MEM were 0.5, 1 and 1 for the three variables. See the Fig. 1 caption of the present appendix for the legend of the boxes.



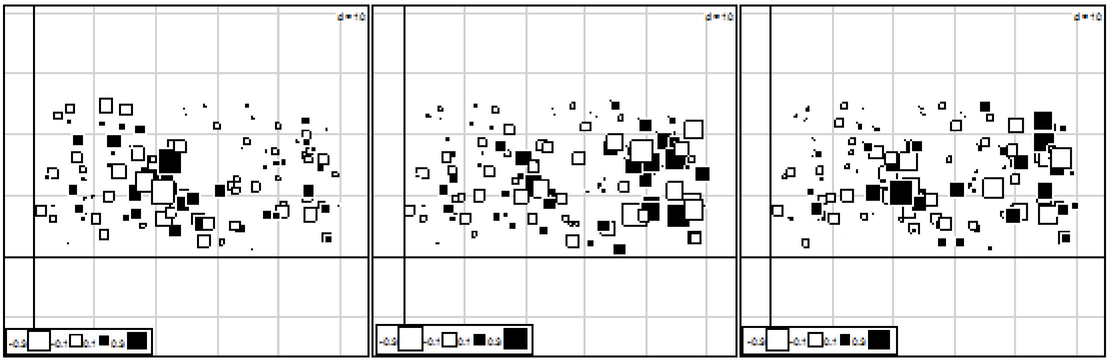
Appendix A1 Figure 9: Examples of a realisation of the linear combination of the MEM variables to which some random noise was added at broad, medium and fine scales, for the regular design. In this example, the three MEM variables explained 0.44, 0.61, and 0.70 (adjusted R^2) of the simulated distributions at broad, medium and fine scales, respectively.



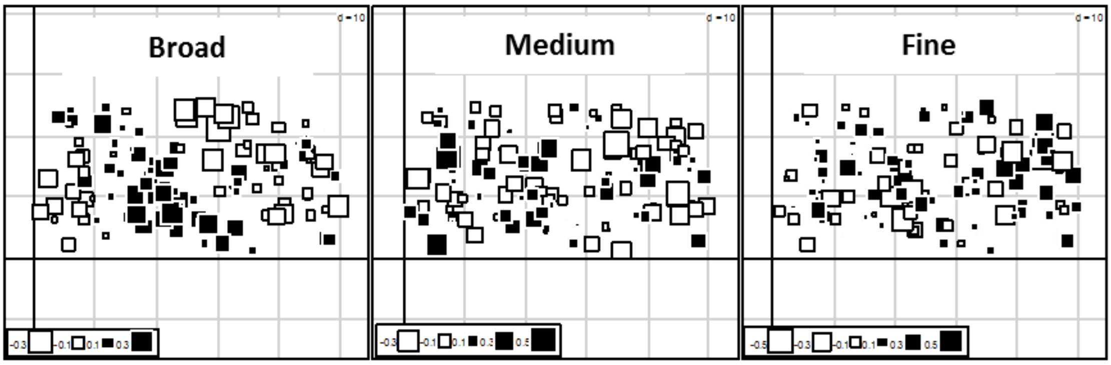
Appendix A1 Figure 10: MEM variables (1, 2 and 3, from the left to the right) used for constructing simulated species structured at a broad scale in a random design. The linear combination coefficient of the MEM for the linear combination was 0.5 for the three variables. See the Fig. 1 caption of the present appendix for the legend of the boxes.



Appendix A1 Figure 11: MEM variables (19, 20 and 21, from the left to the right) used for constructing simulated species structured at a medium scale in a random design. The linear combination coefficients of the MEM for the linear combination were 0.6, 1, and 0.8 for the three variables. See the Fig. 1 caption of the present appendix for the legend of the boxes.



Appendix A1 Figure 12: MEM variables (38, 39 and 40, from the left to the right) used for constructing simulated species structured at a fine scale in a random design. The linear combination coefficients of the MEM for the linear combination were 0.5, 1, and 1 for the three variables. See the Fig. 1 caption of the present appendix for the legend of the boxes.



Appendix A1 Figure 13: Examples of a realisation of the linear combination of the MEM variables to which some random noise was added at broad, medium and fine scales, for the random design. In this example, the three MEM variables explained 0.49, 0.58, and 0.67 (adjusted *R^2*) of the simulated distributions at broad, medium and fine scales, respectively.

Appendix S2: R code used for the simulation study.

The R code is too long to be presented here. It is available at http://www.ecography.org/appendix/ecog-03380

Appendix S3: R function *MEM.moransel* to perform a selection of spatial eigenvectors based on the minimization of the Moran's *I* of the model residuals (MIR selection).

Available online at http://www.ecography.org/appendix/ecog-03380
Note that the function is now integrated to the function *mem.select* of the *adespatial* package (argument *method* = *"MIR"*).

Appendix A4: List of the 301 references used for the review of methods using Moran's eigenvector maps (MEM) and spatial eigenvector mapping (SEVM).

The appendix is too long to be presented here. Is is online at http://www.ecography.org/appendix/ecog-03380

Appendix S5: Guidelines of spatial eigenvector selection in R.

Objective of this document

This document provides general guidelines for selecting an optimal subset of spatial eigenvectors (MEM variables) depending on the purpose of the study and on the univariate or multivariate nature of the response.

Useful packages

```
library(vegan)

## Warning: package 'vegan' was built under R version 3.4.4

library(adespatial)
library(spdep)

## Warning: package 'spdep' was built under R version 3.4.4
```

Data input

The oribatid mite dataset will be used to illustrate the eigenvector selection procedures (see Borcard et al. 1992, 1994 for details on the data).

```
data(mite)
data(mite.xy)

Y <- mite
C <- mite.xy
```

We transform the species data with the Hellinger transformation (more details in Legendre and Gallagher 2001):

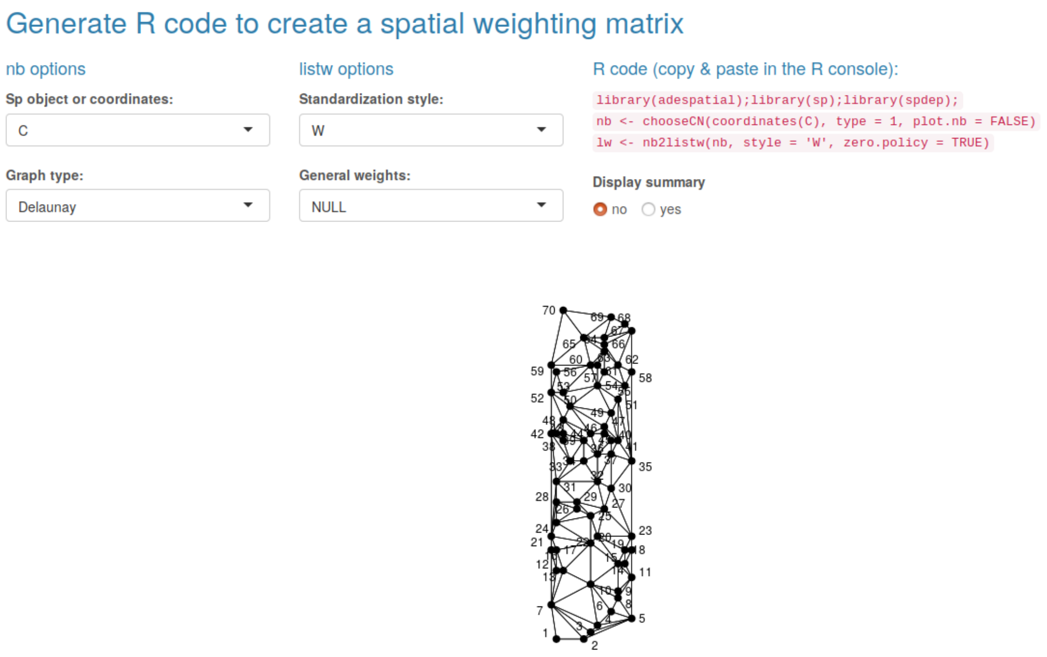
```
Y <- decostand(Y, method = "hellinger")
```

I. Moran's eigenvector maps (MEM): constructing the spatial variables

The MEM variables can be built from a huge variety of spatial weighting matrices (W). The W matrix is constructed by the Hadamard product of a connectivity matrix (B) defining which sites are connected and which are not, and a weighting matrix (A) either binary (no weighting) or continuous often causing connectivity to decrease with distance. The decrease can be linear, or follow a concave-down or concave-up curve (see Dray et al. 2006 for details), or any other function defined by the user. This R code does not provide the way to select an optimal W matrix, as this procedure still needs to be thoroughly adressed through an unbiased procedure.

The function `createlistw` of the `adespatial` package offers an interactive way to create and visualize different connectivity criteria. This can help deciding which connectivity scheme is more adapted for a given study case.

```
C <- as.matrix(C)
createlistw()
```



Generate R code to create a spatial weighting matrix

nb options

Sp object or coordinates: C

Graph type: Delaunay

listw options

Standardization style: W

General weights: NULL

R code (copy & paste in the R console):

```
library(adespatial);library(sp);library(spdep);
nb <- chooseCN(coordinates(C), type = 1, plot.nb = FALSE)
lw <- nb2listw(nb, style = 'W', zero.policy = TRUE)
```

Display summary

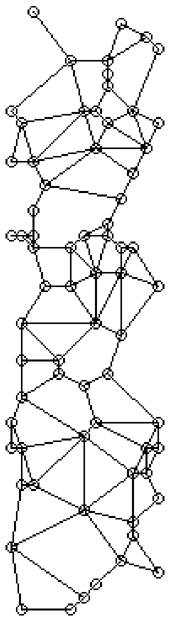
● no ○ yes

Screenshot of the interactive tool provided by the function `createlistw`.

Here, we use a Gabriel graph with no weighting to build the MEM variables, as an example to illustrate the unbiased eigenvector selection procedures. However, any W matrix can be used for the continuation of the procedure.

Gabriel graph of the mite dataset:

```r
nb <- chooseCN(C, type = 2, plot.nb = FALSE)
lw <- nb2listw(nb, style = 'B', zero.policy = TRUE)
par(mar = c(0, 0, 0, 0))
plot(nb, C)
```



We start by defining some parameters. First, we define whether we want the MEM variables modelling positively autocorrelated patterns (`"positive"`), negatively autocorrelated patterns (`"negative"`), or all n-1 MEM variables (`"all"`):

```
MEM_model <- "positive"
```

Then, we define the standardisation scheme of the `listw` object (see help of `nb2listw` function), and then build the set of spatial eigenvectors (stored in object `MEM`). Depending on the value of `MEM_model`, the object `MEM` contains all n-1 eigenvectors, or only the eigenvectors associated with positive or negative eigenvalues (corresponding to positively and negatively autocorrelated patterns, respectively).

```
style <- "B"

nb <- graph2nb(gabrielneigh(as.matrix(C), nnmult = 5), sym = TRUE)
listw <- nb2listw(nb, style = style)
MEM <- scores.listw(listw, MEM.autocor = MEM_model)
```

II. Eigenvector selection

II.1. Spatial filters - Controlling the spatial autocorrelation of OLS or GLM model residuals

When the response data is univariate, if the purpose of the spatial predictors is to control the spatial autocorrelation of an OLS or GLM model, then the most suited MEM variable selection is that of Griffith and Peres-Neto (2006). This procedure selects the smallest MEM subset minimising spatial autocorrelation (Moran's I) in the residuals.

As this selection procedure is restricted to univariate data, we only consider the second species of the community dataframe in our example.

```
Y <- mite
Y <- Y[, 2]
```

II.1.1. Selection of the spatial predictors based on the residuals of a model relating Y to a set of explanatory variables (X)

The Moran eigenvector filtering function ME allows removing spatial autocorrelation from the residuals of generalised linear models (see help of ME for details).

```
data(mite.env)
X <- mite.env[, 1:2]

select <- ME(Y ~., data = as.data.frame(X), listw = listw, family = gaussian,
             nsim = 99, alpha = 0.05)
MEM.select <- select$vectors
```

`MEM.select` can be used to control the spatial autocorrelation of our model by adding the MEM variables to the explanatory variables of the model.

II.1.2. Selection of the spatial predictors based on Y only (MIR approach in Bauman et al. 2017, Fig. 1, Step 3.3.)

The `MEM.moransel` function focuses on controling the spatial autocorrelation of Y, instead of the spatial autocorrelation of the residuals of the model relating Y to X. The function does the same as ME, but with a model that contains only an intercept term.

Call the MEM.moransel.R (Appendix A3 in Bauman et al. 2017), construct the spatial predictors and select a subset of them following the MIR procedure.

```
source("MEM.moransel.R")
moransel <- MEM.moransel(Y, listw, MEM.autocor = MEM_model, nperm = 999,
                         alpha = 0.05)
```

`MEM.moransel` returns two dataframes containing all the spatial predictors (MEM.all) and the subset of predictors selected by the procedure (MEM.select). If no significant spatial autocorrelation could be detected in Y, then an informing message is printed.

II.2. Selecting MEM variables to describe space as accurately as possible

If the response data is multivariate, and/or if the purpose of the spatial predictors is to capture as much spatial structure as possible in Y (i.e., maximise the spatial fit), whether it is related to a set of explanatory variables (X) or not, then the forward selection (FWD) of Blanchet et al. (2008) should be preferred.

A first mandatory step before performing the FWD is to check the significance of the global model, that is, the model of Y as a function of all the spatial predictors. The FWD can only be performed if this global test is significant at a predefined threshold of null hypothesis rejection (here, 0.05). This step was shown to control the Type I error rate that otherwise can be highly inflated (Blanchet et al. 2008).

If `MEM_model = "all"` (we are interested in both positively and negatively autocorrelated patterns), then two separate global tests are performed, on the MEM displaying positive and negative eigenvalues, respectively. A p-value correction for multiple testing is then applied (Sidak correction) and the FWD is performed only if at least one of the two tests is significant (see Blanchet et al. 2008).

The FWD with two stopping criteria consists in 1) searching the MEM variable that best explains Y (highest adjusted R^2 adjusted by the Ezekiel correction, 1929), then 2) to search for the next MEM best explaining the residuals of Y on the first selected MEM, etc. At each selection step, two stopping criteria are used to accept the next best MEM or stop the procedure: a) the p-value of the added MEM (as in the classical forward selection), and b) the adjusted R^2 of the global model (including all predictors). This second criterion was shown by Blanchet et al. (2008) to avoid model overfitting (one of the main issues of the classical forward selection).

Here, we consider the complete community dataframe (multivariate response):

```
Y <- mite
Y <- decostand(Y, "hellinger")
```

Eigenvector selection using the forward selection with double stopping criterion:

```
if (MEM_model != "all") {   # We consider only positively or negatively autocorrel
ated MEM

  if (anova(rda(Y, MEM), permutations = 9999)$Pr[1] <= 0.05) {
    # Global adjusted R-squared of the model
    R2adj <- RsquareAdj(rda(Y, MEM))$adj.r.squared
    # FWD with two stopping criteria
    fsel <- forward.sel(Y, MEM, adjR2thresh = R2adj, nperm = 999)
    # We order the selected MEM by decreasing eigenvalue
    sorted_sel <- sort(fsel$order)
    # Object containing the selected MEM
    MEM.select <- as.data.frame(MEM)[, c(sorted_sel)]
  } else print("No significant spatial autocorrelation was detected")

} else {   # We consider both positively and negatively autocorrelated predictors
  # List to save the positively and negatively autocorrelated MEM separately
  mem.sign <- vector("list", 2)
  signif <- c("FALSE", "FALSE")
  # We select the positive and negative MEM separately after testing the global
  # significance of both models at a corrected threshold value of null hypothesis
  # rejection (Sidak correction)
  for (i in 1:2) {
    if (i == 1) {   # Positive MEM
      mem <- MEM[, which(attributes(MEM)$values > 0)]
    } else {        # Negative MEM
      mem <- MEM[, which(attributes(MEM)$values < 0)]
    }
    # Global test of significance with the Sidak correction for multiple tests
    if (anova(rda(Y, mem), permutations = 9999)$Pr[1] <= (1-(1-0.05)^0.5)) {
      # Global adjusted R-squared of the model
      R2adj <- RsquareAdj(rda(Y, mem))$adj.r.squared
      # FWD with two stopping criteria
      fsel <- forward.sel(Y, mem, adjR2thresh = R2adj, nperm = 999)
      # We order the selected MEM by decreasing eigenvalue
      sorted_sel <- sort(fsel$order)
      # We save the selection of MEM
      mem.sign[[i]] <- as.data.frame(mem)[, c(sorted_sel)]
      signif[i] <- "TRUE"
    }
  }
  # MEM.select will contain both positive and negative MEM, only positive or only
  # negative MEM, depending on the significance of the global tests.
  if (length(which(signif == "FALSE")) != 2) {
    if (length(which(signif == "TRUE")) == 2) {
      MEM.select <- cbind(mem.sign[[1]], mem.sign[[2]])
    } else if (signif[1] == "TRUE") {
      MEM.select <- mem.sign[[1]]
    } else MEM.select <- mem.sign[[2]]
  } else print("No significant spatial autocorrelation was detected")
}

## Testing variable 1
## Testing variable 2
## Testing variable 3
## Testing variable 4
## Testing variable 5
## Testing variable 6
## Testing variable 7
## Testing variable 8
## Testing variable 9
## Testing variable 10
## Testing variable 11
## Testing variable 12
## Testing variable 13
## Testing variable 14
## Procedure stopped (alpha criteria): pvalue for variable 14 is 0.077000 (> 0.05)
```

The MEM variables of `MEM.select` can be used as (co)variables in OLS or GLM models (univariate reponse), or in an RDA or CCA (multivariate response).

References

Bauman, D. et al. 2017. Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors – Ecography 41(10): 1638–1649.

Blanchet, F. G. et al. 2008. Forward Selection of Explanatory Variables. - Ecology 89: 2623–2632.

Borcard, D. et al. 1992. Partialling out the spatial component of ecological variation. Ecology 73: 1045-1055.

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

Dray, S. et al. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). - Ecol. Modell. 196: 483–493.

Ezekiel, M. 1929. The application of the theory of error to multiple and curvilinear correlation. - J. Am. Stat. Assoc. 24: 99–104.

Griffith, D. A. and Peres-Neto, P. R. 2006. Spatial modeling in Ecology: the flexibility of eigenfunction spatial analyses. - Ecology 87: 2603–2613.

Legendre, P., and Gallagher, E. D. 2001. Ecologically meaningful transformations for ordination of species data. - Oecologia 129(2): 271-280.

Table A1: Review of 301 peer-reviewed papers published between 2006 and 2016 and using eigenvector-based methods (Moran's eigenvector maps and spatial eigenvector mapping).

The table is too large to be included here. It is available online at http://www.ecography.org/appendix/ecog-03380

Table A2: Type I error rates of the AIC, FWD and MIR approaches in a univariate frame, considering a regular and a random sampling design.

	Regular design						**Random design**					
	Type I error			***R^2_{adj} ± sd***			***Type I error***			***R^2_{adj} ± sd***		
Random response	**Fwd**	**AIC**	**MIR**	**Fwd**	**AIC**	**MIR**	**Fwd**	**AIC**	**MIR**	**Fwd**	**AIC**	**MIR**
Uniform	0.050	0.999	0.053	0.23 ± 0.06	0.24 ± 0.08	0.06 ± 0.03	0.051	0.982	0.049	0.16 ± 0.06	0.15 ± 0.06	0.05 ± 0.05
Normal	0.047	1.000	0.052	0.23 ± 0.05	0.24 ± 0.08	0.06 ± 0.04	0.052	0.981	0.054	0.16 ± 0.04	0.15 ± 0.06	0.05 ± 0.03
Exponential 1	0.045	1.000	0.050	0.23 ± 0.05	0.24 ± 0.07	0.06 ± 0.03	0.047	0.981	0.054	0.18 ± 0.05	0.15 ± 0.06	0.05 ± 0.03
Exponential cube	0.045	1.000	0.052	0.22 ± 0.07	0.22 ± 0.07	0.06 ± 0.05	0.051	0.976	0.051	0.22 ± 0.07	0.16 ± 0.08	0.07 ± 0.05

Table A3: Statistical power and mean R^2 estimation accuracy of a method of spatial predictor selection consisting in using the forward selection (FWD approach) with the AICc instead of the R^2 as criterion of selection. The significance of the global model was also used as a condition to enter the forward selection, and the R^2 of the global model was used as second stopping criterion, as in the original method (Blanchet et al. 2008). The ΔR2 is computed as the difference between the simulated and the real R^2. sd: standard deviation. R2adjReal: Mean adjusted R^2 calcultated by a linear regression of the response variable against the three MEM variables used to create the spatial component of the response variable.

	Regular sampling design									**Random sampling design**								
Scale	*Broad*			*Medium*			*Fine*			*Broad*			*Medium*			*Fine*		
	Power	**Mean**	sd	Power	**Mean**	sd	Power	**Mean**	sd	Power	**Mean**	sd	Power	**Mean**	sd	Power	**Mean**	sd
ΔR2	0.978	-0.005	0.090	1.000	-0.159	0.119	0.677	-0.593	0.056	0.999	-0.017	0.107	1.000	-0.022	0.048	1.000	-0.048	0.066
R2adjReal		0.430	0.061		0.667	0.042		0.694	0.039		0.432	0.063		0.671	0.041		0.693	0.039

Using the AICc or the R² as selection criterion in a forward selection (Blanchet et al. 2008) provides very similar results in terms of type I error rates, statistical power, and accuracy in a univariate context (see Fig. 3 of the paper). However, while the R^2 is a reliable statistics both for univariate and multivariate response variables, the multivariate AIC proposed by Godinez-Dominguez and Freire (2003) is a wrong analogy of the univariate AIC and is therefore not reliable (see discussion of the paper).

Table A4: Statistical power and mean R^2 estimation accuracy computed with the FWD and MIR approaches (i.e., methods having a correct type I error rate) on response variables displaying a fine-scaled and negatively autocorrelated spatial structure (10000 simulations). The AIC approach was not tested since it presents a highly inflated type I error rate and should therefore not be used. The simulated response variables were constructed as the sum of a linear combination of the last three negatively autocorrelated MEM variables and a random normal noise (N(0, 1)) . The ΔR2 is computed as the difference between the simulated and the real R^2. sd: standard deviation. R2adjReal: Mean adjusted R^2 calcultated by a linear regression of the response variable against the three MEM variables used to create the spatial component of the response variable.

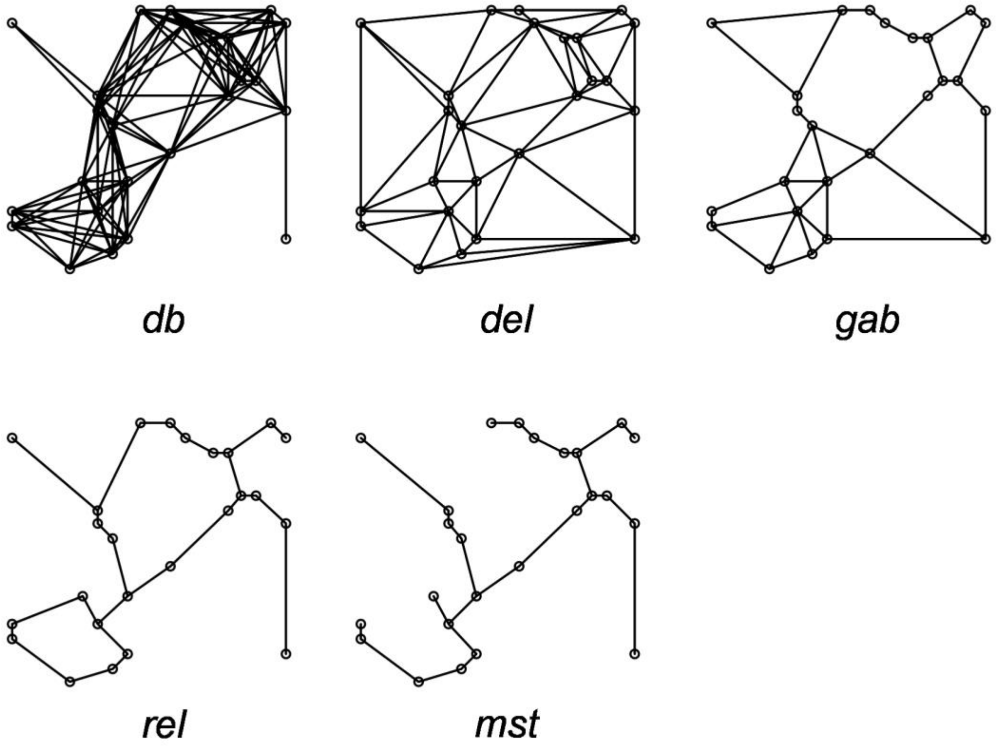
	Regular sampling design						**Random sampling design**					
Eigenvector selection	*FWD*			*MIR*			*FWD*			*MIR*		
	Power	**Mean**	sd	Power	**Mean**	sd	Power	**Mean**	sd	Power	**Mean**	sd
ΔR2	1.000	-0.011	0.041	1.000	0.000	0.025	1.000	-0.048	0.066	0.995	-0.079	0.117
R2adjReal		0.694	0.038		0.694	0.038		0.693	0.039		0.693	0.039

IV. Chapitre IV : Supporting information

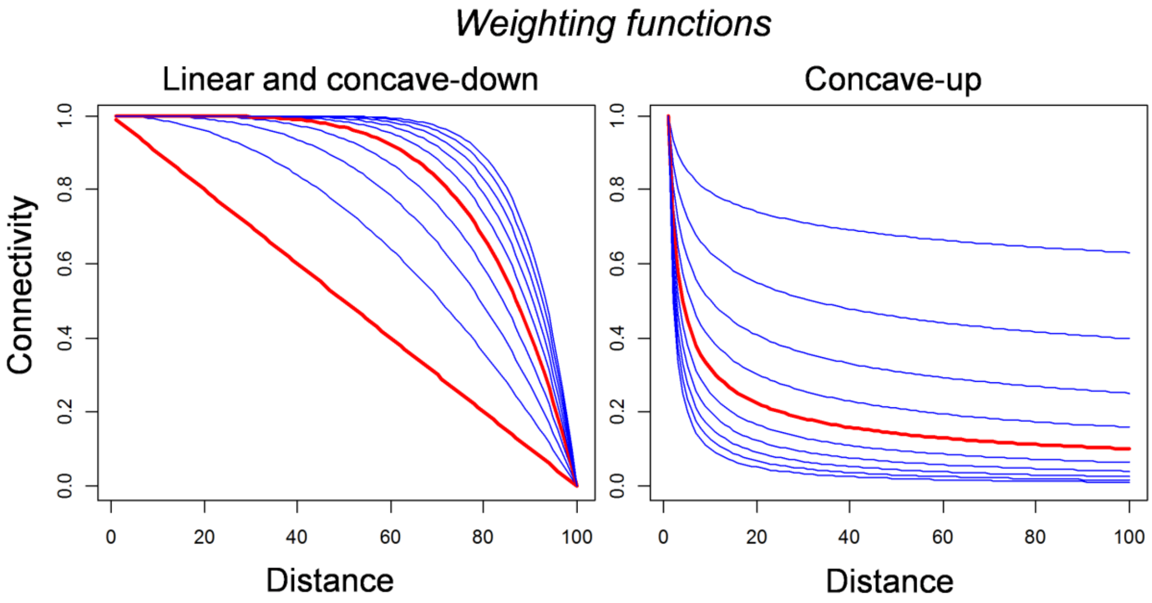
The additional supporting information may be found in the online version of the article at http://onlinelibrary.wiley.com/doi/10.1002/ecy.2469/suppinfo

Appendix S1: Supplementary figures.

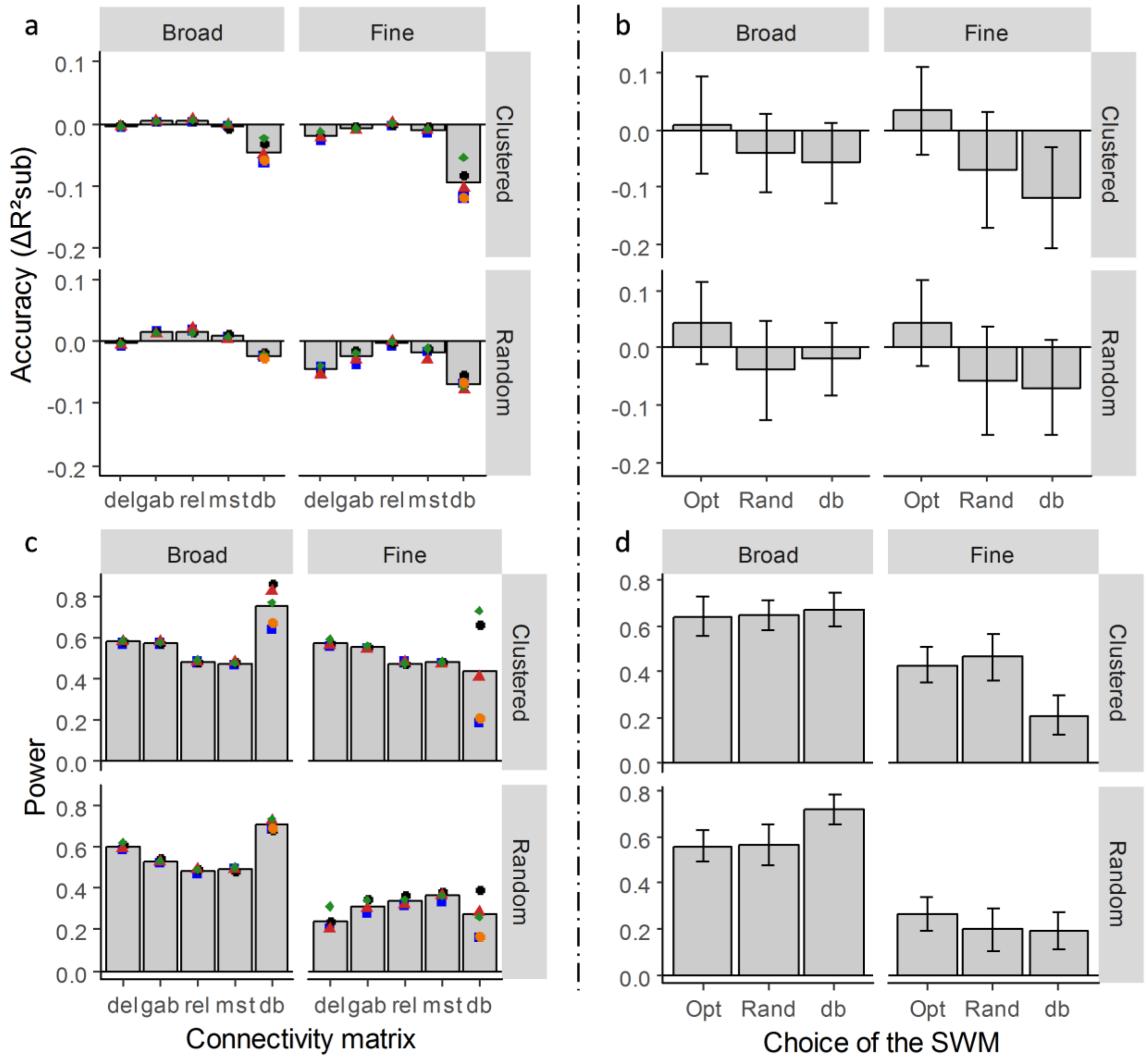
Appendix S1 – Figure S1: Illustration of distance-based and graph-based schemes on 25 randomly-distributed points. The distance-based connection scheme connects all points distant from less than a threshold value (here, the longest edge of the minimum spanning tree, like in PCNM). The graph-based connection schemes are inclusive, so that all the links of the minimum spanning tree (mst) are included in the relative neighborhood graph (rel), included in the Gabriel graph (gab), itself included in the Delaunay triangulation (del; see Legendre and Legendre 2012 for details). The number of connections is as follows: db ≥ del ≥ gab ≥ rel ≥ mst (see Fortin and Dale 2014).



Appendix S1 – Figure S2: Plot of connectivity against distance according to three weighting functions commonly used to build spatial weighting matrices. The left-hand plot represents a linear (straight red line) and concave-down (curves) decreases of connectivity between two points with respect to distance. The right-hand plot shows the concave-up decrease of connectivity with an increasing distance between two points. The linear, concave-down, and concave-up functions are defined by $f_{lin} = 1 - (d/d_{\max})$, $f_{down} = 1 - (d/d_{\max})^{\alpha}$, and $f_{up} = 1/d_{\max}^{\alpha}$, respectively, where d = the Euclidean distance between two sites, $d_{\max}$ = the maximum distance between two sites in a squared Euclidean distance matrix, and α is a parameter taking values of 2 to 10 (bottom curve to top curve in f_{down}), or 0.1 to 1 (top to bottom curve in f_{up}). The red curves correspond to the values of the parameters used in the simulations for the corresponding weighting functions (f_{down}: 5, f_{up}: 0.5).



Appendix S1 – Figure S3: R^2 estimation accuracy and statistical power of the different SWMs (a, c), the optimization method used with a forward selection criterion, the random choice of a SWM, and the arbitrary choice of *db*-MEM$_{\text{PCNM}}$ (b, d), at broad and fine spatial scales and for different types of sampling design ('Clustered' and 'Random'). These are the results for the low degree of SAC. a, c: The grey vertical bars correspond to the mean of the type I error rate (a), mean ΔR^2 (i.e., $R^2_{sim} - R^2_{ref}$) (c), and statistical power (e) of the different **A** matrices within each **B** matrix (x-axis). The symbols give the detailed values for the combinations of the matrices **B** and **A**. Squares: f_{bin}, black circles: f_{lin}, triangles: f_{down}, diamonds: f_{up}, orange circles: f_{PCNM}. b, d: R^2 estimation accuracy (b) and statistical power (d) of the optimization procedure with *p*-value correction ('Opt'), the random choice of a SWM among 57 candidates ('Rand'), and the *db*-MEM$_{\text{PCNM}}$ ('db'). a, b: Negative and positive values of ΔR^2 correspond to underestimations and overestimations of the actual R^2 (i.e. R^2_{ref}), respectively.



Appendix S2: Complement to the Methods section

Appendix S2 – Section 1: Generation of the clustered and random simulated sampling designs

We built a grid of 90 × 90 cells and 120 cells were drawn following a clustered or a random sampling design. The clustered design was constructed to produce three clusters of 40 cells. To do so, the grid was divided into nine contiguous squares of 30 × 30 cells. Three squares were randomly chosen and, for each of them, 40 cells were randomly drawn in their central 20 x 20 area. This ensured to simulate a clustered sampling (see lower right-hand figure of Fig. 1) as the distance between the sampled cells of two different clusters is at least equal to 10 cells. The random design consisted in a uniform sample of 120 cells drawn from the 90 x 90 grid (see upper right-hand figure of Fig. 1).

Appendix S2 – Section 2: Type I error rate of additional simulations

In addition to computing the type I error rate of the 21 types of SWM, we also evaluated whether the type I error rate was inflated by two common practices (Borcard et al. 2011) that consist in optimizing either the value of a parameter in a weighting function or the value of the connectivity threshold distance in *db*-MEM. To do so, we generated a SWM with *gab* as connectivity criterion, for which 10 α values were tested in f_{down} and f_{up} (values ranging between 1 and 10, and between 0.1 and 1, respectively). We also generated *db*-MEM using ten connectivity threshold distances, ranging from the largest edge of a minimum spanning tree to the maximum distance between two cells. For each simulation and in both cases, we then selected the parameter value or distance threshold of the SWM that provided the highest R^2, following Dray et al. (2006) and Borcard et al. (2011). We then tested the significance of the resulting SWM as described in the *Material and Methods* section of the paper. This simulation procedure was iterated 1000 times and the type I error rate was computed as the proportion of significant SWMs detected (significance threshold of 0.05).

Testing different α exponent values in the concave-down or concave-up functions and selecting the one leading to the highest R^2 inflated the type I error rate (0.12 when using five exponent values). The same problem appeared when testing different threshold distances for *db*-MEM. The type I error rate dropped below 0.04 when applying a *p*-value correction for multiple tests (Šidák 1967).

Appendix S2 – Section 3: Statistical power and R^2 estimation accuracy of the different SWMs

The SWMs were evaluated on the basis of their statistical power and R^2 estimation accuracy in a set of scenarios in which the type of sampling design, the degree of SAC, and the spatial scale of the structuring pattern varied. Each one of these parameters had two modalities (Fig. 1): the sampling design was either clustered or random (see *Type I error rate* section above), the degree of SAC was either high or low (see below), and the response variable was structured either at broad or at fine spatial scale.

The simulation procedure began by generating a set of positive MEM variables on the complete grid (90 × 90 cells) with a SWM considering connected the cells sharing either an edge or a vertex (i.e., 'queen contiguity' (Griffith 2017), yielding eight neighbors for each cell, except for the edge cases). This connectivity criterion is neither a distance-based nor a graph-based criterion, so that no family of MEM is favored. A response variable of a population, $\mathbf{y}_{pop}$, was then built at the level of the complete grid by a standardized linear combination (**u**) of three MEM variables ($\mathrm{MEM}_{\mathrm{STD}}$; MEM 6 to 8, and 40 to 42 for broad and fine scales, respectively) to which a standardized normal random noise (**e**) was added. To generate $\mathbf{y}_{pop}$, these two components were then multiplied by λ and 1 – λ, respectively, were λ is the structuring intensity or degree of SAC ($\lambda_{\mathrm{strong}} = 0.55$, $\lambda_{\mathrm{weak}} = 0.35$), as illustrated in the following equation: $\mathbf{y}_{pop} = \lambda\mathbf{u} + (1 - \lambda)\mathbf{e}$.

Then, 120 cells of the grid were subsampled following a clustered or a random sampling design, leading to $\mathbf{y}_{sub}$ and $\mathbf{u}_{sub}$ (i.e., the sampled subset of $\mathbf{y}_{pop}$ and corresponding subset of **u**). The reference value of the spatial R^2 for the subsampled grid was computed as $\mathrm{cor}^2(\mathbf{y}_{sub,},\mathbf{u}_{sub})$ and equal to ~0.6 ± 0.1 and ~0.2 ± 0.1 for a strong and weak degree of SAC, respectively. Then, a complete set of positive MEM variables was generated on the basis of the coordinates of the sampled cells of the grid using the 29 types of SWMs defined earlier.

For each of these SWMs, a test of the R^2 of the linear model considering the response $\mathbf{y}_{sub}$ against the complete set of positive MEM variables (global test) was performed using 999 permutations of the model residuals, following Anderson and Legendre (1999). If the test was significant (p-value ≤ 0.05), then a forward selection with two stopping criteria (Blanchet et al. 2008) was applied to select a subset of spatial predictors. We then computed the R^2, adjusted by Ezekiel's correction (Ezekiel 1929), of the linear regression of $\mathbf{y}_{sub}$ against this subset of spatial predictors (R^2_{sim}). The R^2_{sub} estimation accuracy (hereafter ΔR^2_{sub}) was finally computed as the difference between the reference R^2 value and the R^2 obtained in the simulation with one of the SWMs (i.e., $\Delta R^2_{sub} = R^2_{sim} - R^2_{sub}$), so that negative and positive values indicated underestimation or overestimation of the real spatial signal, respectively. The complete simulation procedure was iterated 1000 times, and the power was computed for each SWM as the proportion of significant simulations. The R^2 estimation accuracy was computed as the mean ΔR^2_{sub} of the significant simulations.

Appendix S2 – Section 4: Optimization criteria for the selection of the SWM

In the paper, we used the R^2_{adj} of the model explaining the response variable (**y**) by a subset of MEM variables selected by forward selection to optimize the selection of the SWM. However, according to the objective of the study, different criteria can be used in the optimization procedure. We propose three optimization criteria for different purposes, describe them below, and illustrate them in an R tutorial (Appendix S3).

In the great majority of cases, the MEM variables are used in combination with a set of actual (e.g. environmental) predictors (**X**). The objective is then either to describe as accurately as possible the spatial structures of a response dataset (**y**) and relate them to the ones possibly present in **X** (e.g. variation partitioning, Peres-Neto and Legendre 2010), or to remove the spatial autocorrelation from the residuals of a model of **y** against **X** to respect the condition of independence of the model residuals. In these cases, and for a given SWM, a selection of spatial predictors is necessary to avoid model overfitting and a loss of power to detect the contribution of **X** to the variability of **y** (Griffith 2003, Dray et al. 2006, Blanchet et al. 2008, Peres-Neto and Legendre 2010, Diniz-Filho et al. 2012).

If the analysis of interest is of this kind, then the selection of the SWM should be optimized on the basis of what will be used, that is, a subset of spatial predictors. Depending on the objective, either the forward selection with double stopping criterion (Blanchet et al. 2008) or the minimization of the Moran's index *I* in the residuals (MIR selection, see Bauman et al. 2018) should be used.

1) The selection of a subset of MEM variables on the basis of the forward selection with double stopping criterion has recently been shown to be the most powerful and accurate method of spatial eigenvector selection (Bauman et al. 2018). If the objective of the MEM variables is the accuracy of the spatial patterns captured in **y**, then this method is the way to go. Our optimization procedure then tests the significance of all the candidate SWMs by 9999 permutations (i.e. global tests), corrects the p-values according to the number of matrices tested, and performs a forward selection with double stopping criterion within all the significant SWMs. The optimized SWM selected is the one within which the highest forward-selected R-squared (R2) is reached. It is therefore the subset of MEM variables that guides the selection of a SWM.

Note that **y** can be a univariate or a multivariate response dataset containing species abundances, for instance, but it can also be the residuals of the model of **y** against **X**. In the latter case, using the forward selection optimization criterion is the most appropriate choice if the objective is to optimize the selection of the spatial structures of **y** unexplained by **X** (i.e. the residual spatial patterns).

2) If the objective is to estimate the coefficients of **X** in a model of type ordinary least squares, or generalized linear models of **y** against **X**, then the objective of the spatial predictors is only to remove any significant spatial autocorrelation from the model residuals, and the accuracy of the spatial patterns captured is secondary. In this case, one should add as few spatial predictors as possible to the model (to avoid large standard errors), and the best practice will be to optimize the selection of the SWM and its subset of spatial predictors on the basis of the residuals of the model. Bauman et al. (2018) showed that performing the eigenvector selection on the basis of the minimum number of predictors best minimizing the spatial autocorrelation of the vector of interest (Griffith and Peres-Neto 2006) yielded less accurate results than the forward selection but had the advantage to select a smaller number of MEM variables, which is an important criterion in this case. For one given SWM, this selection procedure (MIR selection) tests the significance of the Moran's *I* in the vector of interest **y** (which in this case would be the residuals of the model of **y** against **X**), then searches for the MEM variable best minimizing the value of the Moran's *I*, creates a model of **y** against this MEM variable, and tests the significance of the Moran's *I* of this new model residuals. The procedure goes on until the Moran's *I* of the model residuals is not significant anymore, hence the name of **M**inimization of the Moran's ***I*** in the **R**esiduals. If the MIR optimization criterion is chosen, our optimization procedure applies the MIR selection to each significant SWM separately and selects the SWM with the smallest

number of MIR-selected spatial predictors. In this case, again, the selected subset of MEM variables within the significant SWMs guides the selection of the SWM. Note that the MIR criterion of optimization can only be used for a univariate **y**, as the Moran's *I* is a univariate index. If **y** is multivariate, then the best choice is the forward selection (see Bauman et al. 2018).

3) In a few cases, all the spatial eigenvectors are needed for further analyses. This, for example, is the case when using Moran spectral randomizations (MSR; Wagner and Dray 2015) or smoothed MEM (Munoz 2009). In this case, the optimization should be conducted on the basis of the adjusted R2 of the whole set of eigenvectors generated from the different SWMs. Most studies are interested in contagious processes, such as dispersal limitation or spatial induced dependences related to the environment, for instance. In those cases, only the MEM variables associated to positive eigenvalues are generated, and our optimization procedure can compute an adjusted R2 for each SWM. The selected SWM is then the one displaying the highest R2. If, however, the MEM variables related to both positively and negatively spatially autocorrelated structures are needed, then our method tests the significance of the candidate SWMs for positively and negatively autocorrelated structures separately, computes the corresponding global adjusted R2, and selects the SWM for which the sum of the positive and negative model R2 is the highest.

Appendix S2 – Section 5: Comparison of the optimization method to a randomly-chosen SWM

Five contrasting SWM candidates were used in the optimization procedure: *gab* and *mst* (**B** matrices) associated with the f_{lin} and f_{down} functions (**A** matrices), and the *db*-MEM$_{\text{PCNM}}$. The power and accuracy of our optimization procedure were compared to those of the arbitrary choice of *db*-MEM$_{\text{PCNM}}$ (i.e., the most common current practice), and to those of the random selection of a SWM among a set of 57 SWMs. The latter corresponded to *del*, *gab*, *rel*, *mst*, and 10 *db* (**B** matrices) associated with f_{bin}, f_{lin}, f_{down}, f_{up}, (**A** matrices), and *db*-MEM$_{\text{PCNM}}$. The *db*-**B** matrices were generated with distance thresholds ranging regularly from the smallest distance keeping all points connected (as for *db*-MEM$_{\text{PCNM}}$), to the distance corresponding to the highest semi-variance of an empirical variogram, following Borcard et al. (2011).

References

Anderson M. J., Legendre P. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. Journal of Statistical Computation and Simulation 62:271–303.

Bauman D. et al. 2018. Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors. Ecography 41(10): 1638–1649.

Blanchet F. G. et al. 2008. Forward selection of explanatory variables. Ecology 89:2623–2632.

Borcard D. et al. 2011. Numerical Ecology with R. Springer. Springer, New-York.

Diniz-Filho J. A. F. et al. 2012. On the selection of phylogenetic eigenvectors for ecological analyses. Ecography 35:239–249.

Dray S. et al. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecological Modelling 196:483–493.

Ezekiel M. 1929. The application of the theory of error to multiple and curvilinear correlation. Journal of the American Statistical Association 24:99–104.

Griffith D. 2003. Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization. Springer, Berlin.

Griffith D. A. 2017. Spatial Weights. Pages 1–7in D. Richardson, N. Castree, M. F. Goodchild, A. Kobayashi, W. Liu, and R. A. Marston, editors.International Encyclopedia of Geography: People, the Earth, Environment and Technology. Wiley-Blackwell.

Munoz F. 2009. Distance-based eigenvector maps (DBEM) to analyse metapopulation structure with irregular sampling. Ecological Modelling 220:2683–2689.

Peres-Neto P. R., Legendre P. 2010. Estimating and controlling for spatial structure in the study of ecological communities. Global Ecology and Biogeography 19:174–184.

Šidák, Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 62:626–633.

Wagner, H. H., Dray S. 2015. Generating spatially constrained null models for irregularly spaced data using Moran spectral randomization methods. Methods in Ecology and Evolution 6:1169–1178.

Appendix S3: Tutorial to optimize the selection of a spatial weighting matrix and/or of a subset of spatial predictors in eigenvector-mapping methods.

Useful packages

```
library(adespatial)
library(spdep)
library(vegan)
```

Objective of this document

This document provides a tutorial for (1) the selection of the optimal spatial weighting matrix (SWM) among a set of candidate matrices, and/or (2) the selection of a subset of MEM variables within these SWMs. MEM variables (also further referred to as spatial predictors) are spatial eigenvectors of a doubly centered SWM whose corresponding eigenvalues are linearly related to Moran's index of spatial autocorrelation.

- The function `listw.candidates` allows building a list of one or more SWMs.
- The function `listw.select` allows optimizing the selection of the SWM from the list of candidates generated with `listw.candidates`, and optionnaly selects a subset of spatial predictors. In this latter case, it is the selection of a subset of predictors that directs the selection of the SWM.
- The function `mem.select` selects spatial predictors within one given SWM. It works as `listw.select`, but without the optimization of the SWM.

All the information presented here is detailed in the help document of the three above-mentioned functions (package *adespatial*).

I. Preparing a list of candidate spatial weighting matrices with `listw.candidates`

This function is a user-friendly way to create a list of one or more spatial weighting matrices (SWM) by selecting a set of predefined connectivity and weighting matrices (**B** and **A** matrices, respectively). The list can then be fed to the function `listw.select` to optimize the selection of the SWM and select the best eigenvector subset within this matrix while controlling the type I error rate.

The function allows to construct SWMs based on any combination of **B** and **A**. The **B** matrices are either graph-based or distance-based. The function proposes the Delaunay triangulation, Gabriel graph, relative neighbourhood graph, and the minimum spanning tree criteria to build a graph-based **B**. Distance-based SWMs can be built with the principal coordinates of neighbor matrices (PCNM; Borcard and Legendre 2002) criteria (see details below), or using another threshold distance and weighting function to define the connected pairs of site and the weight of their connection, respectively. The **A** matrix can be absent (binary SWM) or can based on a linear, concave-down, or concave-up function. The linear, concave-down, and concave-up weighting functions are defined by $1-(D/dmax)$, $1-(D/dmax)^y$, and $1/D^y$, respectively, where D is the euclidean distance between the two sites considered, $dmax$ is the maximum euclidean distance between two sites among all site pairs, and y is a user-defined parameter that can either be a single value or a vector of values. In the latter case, three values or y, for instance, will correspond to three different SWMs.

The argument $pcnm$ of the function consists in constructing a distance-based SWM based on the largest edge of the minimum spanning tree (i.e. the smallest distance keeping all sites connected), and then by weighting the links by the function $1-(D/(4*t))^2$, where D is the euclidean distance between the sites, and t is the distance threshold below which two sites are considered connected (Dray et al. 2006).

As optimizing the selection of a SWM has to be done with a p-value correction related to the number of candidate SWMs tested (see function `listw.select`), the significance threshold will decrease as the number of SWMs tested increases. This leads to a trade-off between the gain of accuracy and a loss of statistical power. As a consequence, we strongly encourage plotting the concave-down and/or concave-up weighting functions with several parameter values in order to only choose the realistic ones to build the candidate SWMs (e.g., ranging between 0.1 and 1 for the concave-up function, as values over 1 would make no ecological sense). First visualizing the connectivity schemes with the interactive `listw.explore` function may also help choosing the **B** matrices worth being generated.

See `help(listw.candidates)` for details.

Once a list of SWMs was created - be it with `listw.candidates` or not -, the list is used in the `listw.select` function. This function will optimize the selection of the SWM depending on a criterion to be selected, based either on a best subset of spatial eigenvectors within each SWM, or on the complete set of eigenvectors generated from the SWMs.

II. Optimizing the selection of a spatial weighting matrix with `listw.select`

As shown in our paper, optimizing the selection of the SWM led to inflated type I error rates if an explicit control of the number of SWMs tested was not applied. The function `listw.select` therefore applies a Sidak correction (Sidak 1967) for multiple tests to the p-value of the global test of each SWM (i.e., the model integrating the whole set of spatial predictors). The Sidak correction is computed as: $P\_corrected = 1-(1-P)^n$, where n is the number of tests performed, $P\_corrected$ is the new p-value after the correction, and P is the observed p-value. The default option of the function is to generate only MEM variables displaying positive spatial autocorrelation (i.e. the eigenvectors associated to positive eigenvalues; hereafter positive MEM variables). This option can be set to focus on negative spatially autocorrelated structures, or on both types. If the latter option is selected, then *n* - 1 MEM variables are generated from each SWM, in which case the function tests the significance of the global model separately for the positive and the negative MEM variables (see Blanchet et al. 2008). This leads to multiplying by two the number of tests, hence leading to a more severe p-value correction.

The p-value is first computed by permutations, and is then corrected according to the total number of SWMs tested. Although `listw.select` can be run without this correction, using the default value is strongly recommended to avoid highly inflated type I error rates (see Results of the paper).

`listw.select` computes spatial eigenvectors for any type of SWMs and optimizes the selection of the SWM on the basis of one out of three possible criteria. The choice of this optimization criteria has to be guided by the use one wants to make of the SWM and its spatial eigenvectors.

II.1. Optimization criteria

The objective of the optimization process can either be the entire SWM itself (i.e. all the MEM variables generated), or an optimal subset of spatial predictors within the best SWM. In this last case, the priority is to find the best subset of spatial predictors, and to choose the corresponding SWM.

II.1.1. The objective is to obtain an optimal subset of spatial predictors

In the great majority of cases, the MEM variables are used in combination with a set of actual (e.g. environmental) predictors (**X**). The objective is then either to describe as accurately as possible the spatial structures of a response dataset (**Y**) and relate them to the ones possibly present in **X** (e.g. variation partitioning, Peres-Neto and Legendre 2010), or to remove the spatial autocorrelation from the residuals of a model of **Y** against **X** to respect the condition of independence of the model residuals. In this latter case, the MEM variables are often referred to as spatial filters (Diniz-Filho and Bini 2005).

In these cases, and for any SWM, a selection of spatial predictors is necessary to avoid model overfitting and a loss of power to detect the contribution of **X** to the variability of **Y** (Griffith 2003, Dray et al. 2006, Blanchet et al. 2008, Peres-Neto and Legendre 2010, Diniz-Filho et al. 2012).

If the analysis of interest is of this kind, then the selection of the SWM should be optimized on the basis of what is actually going to be used, that is, a subset of spatial predictors. The function `listw.select` should be used with the argument `method` set either to `FWD` or to `MIR`.

`II.1.1.A. method = "FWD"`

The selection of a subset of MEM variables on the basis of the forward selection with double stopping criterion (Blanchet et al. 2008) has recently been shown to be the most powerful and accurate method of spatial eigenvector selection (Bauman et al. 2018a). If the objective of the MEM variables are the accuracy of the spatial patterns captured in **Y**, then `method = "FWD"` (default option) is the way to go. `listw.select` then tests the significance of all the candidate SWMs by permutations (i.e. global tests), corrects the p-value according to the number of matrices tested, and performs a forward selection with double stopping criterion within all the significant SWMs. The selected SWM is the one within which the highest forward-selected R-squared (R^2) is reached. The subset of MEM variables therefore guides the selection of a SWM. Note that **Y** can be a univariate or multivariate response dataset containing species abundances, for instance, but it can also be the residuals of the model of **Y** against **X**. In the latter case, using `method = "FWD"` is the most appropriate choice if the objective is to optimize the selection of the spatial structures unexplained by **X**.

`II.1.1.B. method = "MIR"`

If the objective is to estimate the coefficients of **X** in a model ordinary least squares, or generalized linear models of **Y** against **X**, then: 1) one should add as few spatial predictors as possible to the model (to avoid large standard errors), and 2) the best practice will be to optimize the selection of the SWM and its subset of spatial predictors on the basis of the residuals of the model. Bauman et al. (2018a) showed that performing the eigenvector selection on the basis of the minimum number of predictors best minimizing the spatial autocorrelation of the vector of interest (Griffith and Peres-Neto 2006) yielded less accurate results than the forward selection but had the advantage to select a smaller number of MEM variables, which is an important criterion in this case. For one given SWM, this selection procedure (called MIR selection, in Bauman et al. 2018a) tests the significance of the Moran's I in the vector of interest **Y** (which in this case would be the residuals of the model of **Y** against **X**). It then searches for the MEM variable best minimizing the value of the Moran's I, creates a model of **Y** against this MEM variable, and tests the significance of the Moran's I of this new model residuals. The procedure goes on until the Moran's I of the model residuals is not significant anymore (hence the name "**M**inimization of the Moran's **I** in the **R**esiduals").

When `method = "MIR"`, `listw.select` applies the MIR selection to each significant SWM separately (function `MEM.moransel` of Bauman et al. 2018a), and selects the SWM with the smallest number of MIR-selected spatial predictors. If two or more SWMs present the same smallest number of predictors, the choice is made among the ex-aequos on the basis of the Moran's I. Again, the selected subset of MEM variables within the significant SWMs guides the selection of the SWM.

Note that the MIR criterion of optimization can only be used for a univariate **Y**, as the Moran's I is a univariate index. If **Y** is multivariate, then the best option is the forward selection (`method = "FWD"`) (see Bauman et al. 2018a).

See `help(listw.select)` for details.

II.1.1.C. Selection of a subset of MEM variables without optimization of the SWM

The optimization of the SWM may be unnecessary when the sampling design is regular, or if one has good reasons to choose one particular SWM. However, a subset of spatial predictors may still need to be selected within the SWM. The function `mem.select` tests the significance of the SWM provided, computes, and selects the optimal subset of MEM variables on the basis of one of the three above-mentioned criteria.

See `help(mem.select)` for details.

II.1.2. The objective is the complete set of MEM variables (`method = "global"`)

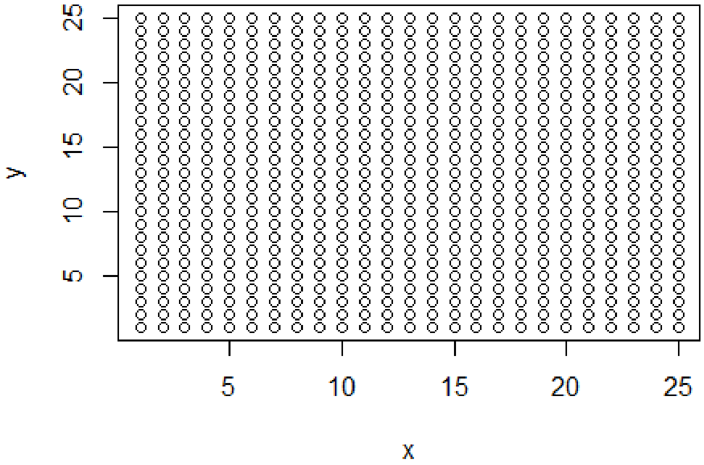
In a few cases, all the spatial eigenvectors are needed for further analyses. This, for example, is the case when using Moran spectral randomizations (MSR; Wagner and Dray 2015) or smoothed MEM (Munoz 2009).

In this case, the optimization should be conducted on the basis of the adjusted R² of the whole set of eigenvectors generated from the different SWMs. Most studies are interested in contagious processes, such as dispersal limitation or spatial induced dependences related to the environment, for instance. In those cases, only the MEM variables associated to positive eigenvalues are generated, and `listw.select` can compute an adjusted R² for each SWM. The selected SWM is then the one displaying the highest R². If, however, the MEM variables related to both positively and negatively spatially autocorrelated structures are needed, then `listw.select` tests the significance of the candidate SWMs for positively and negatively autocorrelated structures separately, computes the corresponding global adjusted R², and selects the SWM for which the sum of the positive and negative model R² is the highest.

III. Illustration of the optimization methods on a simulated example

Create a grid of 25 x 25:

```r
grid <- expand.grid(x = seq(1, 25, 1), y = seq(1, 25, 1))
plot(grid)
```

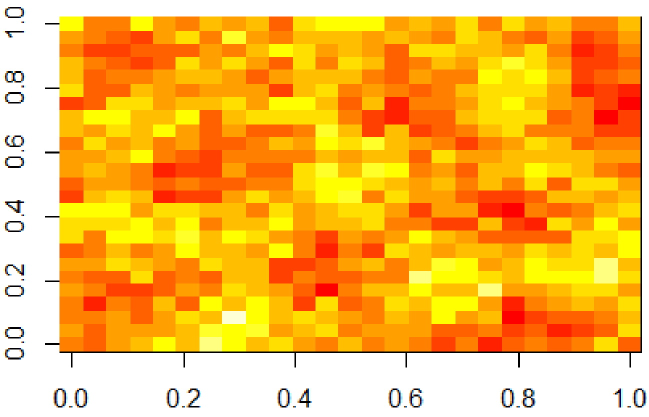


Generation of a univariate response variable **Y** structured at a medium spatial scale. **Y** can for instance be seen as the abundance of a species, or a diversity index value. **Y** is built by a linear combination of three MEM variables to which a normal noise is added:

```
nb <- cell2nb(nrow = 25, ncol = 25, "queen")
nb2 <- nb2listw(nb, style = "B")
MEM <- scores.listw(nb2, MEM.autocor = "positive")
```

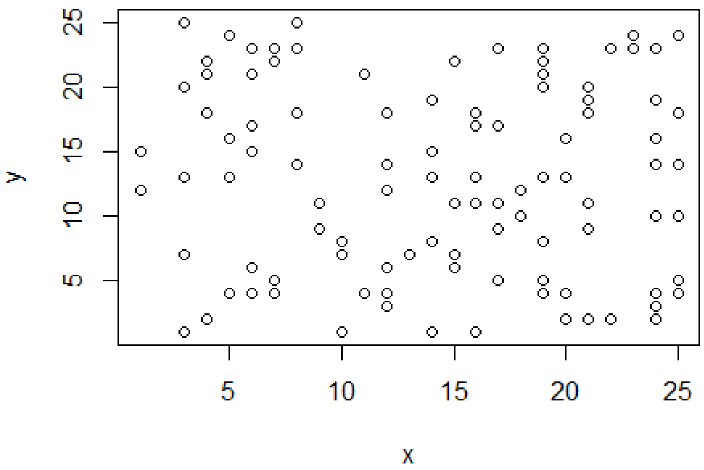
Degree of spatial autocorrelation:

```
set.seed(2018)

intensity <- 0.5
Y_space <- scale(MEM[, 20] + MEM[, 25] + MEM[, 35]) * intensity
Y_noise <- scale(rnorm(n = nrow(MEM), mean = 0, sd = 1)) * (1 - intensity)
Y <- Y_space + Y_noise
# Plot of Y at the level of the complete grid:
image(matrix(Y, ncol = 25, byrow = F))
```



Y was generated in the complete grid. To mimic a realistic sampling effort, **Y** is sampled in 100 randomly-chosen cells of the grid:

```
sample <- sample(c(1:nrow(grid)), 100, replace = FALSE)
xy <- grid[sample, ]
Y_sampled <- Y[sample]
plot(xy)
```



To determine the actual contribution of the spatially structured variability of **Y** to its overall variability (the true spatial R²), we regress the sampled **Y** against the vector `Y_space` sampled in the same cells than **Y**:

```
Y_space_sampled <- Y_space[sample]
R2_reference <- cor(Y_sampled, Y_space_sampled)^2
R2_reference

## [1] 0.4375233
```

The function `listw.candidates` is used to build the SWMs that will be tested and compared. This step only uses the x and y coordinates of the sampled cells. We choose a small number of contrasted SWMs as candidates: a Gabriel graph, a minimum spanning tree, and a distance-based connectivity defined by a threshold distance corresponding to the smallest distance keeping all sites connected (i.e., the defaut value of d2; see help of function `listw.candidates`). These **B** matrices either remain binary (no weighting), or are weighted by the linearly decreasing function (see `help(listw.candidates)`):

```
candidates <- listw.candidates(xy, nb = c("gab", "mst", "dnear"),
                               weights = c("binary", "flin"))
```

Number of candidate SWMs generated:

```
nbw <- length(candidates)
nbw

## [1] 6
```

New significance threshold value after p-value correction (Sidak correction):

```
round(1 - (1 - 0.05)^(1/nbw), 4)
```

```
## [1] 0.0085
```

This new significance threshold will be the one used in `listw.select`.

Now, we use `listw.select` together with **Y** to (1) test whether the latter displays a significant spatial structure, and (2) select the SWM yielding the subset of MEM variables best describing the spatial structure of **Y**. We therefore use the forward selection as optimization criterion (default option).

```
W_sel <- listw.select(Y_sampled, candidates, MEM.autocor = "positive", p.adjust =
TRUE)

## Procedure stopped (adjR2thresh criteria) adjR2cum = 0.415304 with 11 variables
(> 0.406564)
## Procedure stopped (adjR2thresh criteria) adjR2cum = 0.424062 with 8 variables
(> 0.414059)
## Procedure stopped (adjR2thresh criteria) adjR2cum = 0.421607 with 11 variables
(> 0.415638)
## Procedure stopped (adjR2thresh criteria) adjR2cum = 0.398848 with 8 variables
(> 0.389574)
## Procedure stopped (alpha criteria): pvalue for variable 8 is 0.064000 (> 0.05)
## Procedure stopped (adjR2thresh criteria) adjR2cum = 0.389162 with 5 variables
(> 0.368876)
```

Some characteristics of the best spatial model:

Best SWM (connectivity and weighting matrix), its selected subset of spatial predictors (forward selection), and the number of spatial predictors of this subset:

```
W_sel$best.id

## Gabriel_Linear
##              2

W_sel$best$MEM.select

## Orthobasis with 100 rows and 7 columns
## Only 6 rows and 4 columns are shown
##         MEM19      MEM13      MEM33      MEM12
## 1 -0.59754892 -1.0457151 -0.7180794 -1.2047010
## 2 -0.98000331 -0.5263510  0.8926327  1.0851139
## 3 -0.01537057 -1.3773992  0.4943144 -0.1582584
## 4 -0.91805347 -0.5274396 -0.1888308 -0.6832994
## 5  0.84657340 -1.3106319  1.0478327  0.4333857
## 6  0.82441183 -1.5888510 -0.2701254  3.4685860

W_sel$candidates$N.var[W_sel$best.id]

## [1] 7
```

Corrected p-value of the global test of the best SWM:

```
W_sel$candidates$Pvalue[W_sel$best.id]

## [1] 0.002996252
```

Adjusted R^2 of the subset of spatial predictors selected within the chosen SWM, and comparison to the real R^2 computed earlier:

```
# R² obtained with the optimization method:
round(W_sel$candidates$R2Adj.select[W_sel$best.id], 3)
```

```
## [1] 0.404

# Real R²:
R2_reference

## [1] 0.4375233

# Difference between the real R² and the R² obtained using the optimization method
:
diff <- round(R2_reference - W_sel$candidates$R2Adj.select[W_sel$best.id], 2)
diff

## [1] 0.03

# % underestimation:
perc <- round(diff / R2_reference, 2) * 100
perc

## [1] 7
```

The optimization only slightly underestimated the true R^2.

p-values of all the tested SWMs:

```
W_sel$candidates$Pvalue

## [1] 0.002397601 0.002996252 0.004192657 0.013127613 0.019047054 0.004790410
```

Adjusted R^2 of the global models and forward-selected subsets of spatial predictors, for all the significant SWMs:

```
W_sel$candidates

##                        R2Adj      Pvalue N.var R2Adj.select
## Gabriel_Binary     0.4065642 0.002397601    10    0.3962641
## Gabriel_Linear     0.4140589 0.002996252     7    0.4042502
## MST_Binary         0.4156375 0.004192657    10    0.3980359
## MST_Linear         0.3895739 0.013127613     7    0.3718948
## Dnear3.61_Binary 0.2928706 0.019047054     7    0.2418834
## Dnear3.61_Linear 0.3688757 0.004790410     4    0.3571805
```

This last result shows how different the results of different SWMs can be. However, the optimization procedure allowed the detection of the spatial structure in `Y_sampled` with a high degree of accuracy.

It is also worth mentioning that the two distance-based SWMs, although significant, yielded mitigated results in terms of R^2 estimation accuracy, confirming that graph-based MEM are more suited to irregular sampling designs than distance-based MEM (see results and discussion of the paper).

`W_sel$best$MEM.select` can be used directly as a set of spatial predictors to describe the spatial patterns of **Y** or of the common variation between **Y** and a set of environmental variables (see Peres-Neto et al. 2010, Bauman et al. 2018a).

If **Y** is univariate and the objective of the MEM variables is only to explicitly account for the spatial autocorrelation of a model of **Y** against a set of actual predictors (**X**) (spatial eigenvector mapping, Diniz-Filho et al. 2005), then the MIR criterion of optimization is more appropriate.

The following example illustrates this:

```
# For the example, we create an environmental variable which partially explains
# Y_sampled.
# We want to study the influence of this environmental variable on Y_sampled, but
# we first need to make sure that there is no spatial autocorrelation in the model
# residuals.
# If however the residuals are autocorrelated, then we need to optimize the
# selection of a small subset of spatial predictors to be integrated to the model.
set.seed(1)
environment <- (0.3 * scale(Y_sampled)) + (scale(rnorm(length(Y_sampled))) * 0.7)

# The vector we feed to listw.select has to be the residuals of the model of
# interest in this case:
mod <- lm(Y_sampled ~ environment)
summary(mod)

##
## Call:
## lm(formula = Y_sampled ~ environment)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.49123 -0.52244 -0.05192  0.53805  1.58671
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.16295    0.06978  -2.335  0.02156 *
## environment  0.31311    0.09559   3.276  0.00146 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6978 on 98 degrees of freedom
## Multiple R-squared:  0.09869,    Adjusted R-squared:  0.08949
## F-statistic: 10.73 on 1 and 98 DF,  p-value: 0.001458

res <- residuals(mod)
# We use the same set of candidate W matrices:
candidates <- listw.candidates(xy, nb = c("gab", "mst", "dnear"), weights = c("bin
ary", "flin"))
# We perform the optimization with the MIR criterion:
W_sel <- listw.select(res, candidates, MEM.autocor = "positive", p.adjust = TRUE,
                      method = "MIR")

# We can now integrate the optimized subset of spatial predictors into our model.
# The model is reliable now that the model residuals are independent:
environment2 <- cbind(environment, W_sel$best$MEM.select)
mod_final <- lm(Y_sampled ~ ., data = as.data.frame(environment2))
summary(mod_final)

##
## Call:
## lm(formula = Y_sampled ~ ., data = as.data.frame(environment2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4333 -0.4395  0.0134  0.4679  1.7863
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.16295    0.06584  -2.475 0.015060 *
## environment  0.27661    0.09076   3.048 0.002970 **
## MEM5         0.23950    0.06625   3.615 0.000479 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6584 on 97 degrees of freedom
## Multiple R-squared:  0.2057, Adjusted R-squared:  0.1893
## F-statistic: 12.56 on 2 and 97 DF,  p-value: 1.41e-05
```

IV. Generate MEM variables on the basis of a single SWM with `mem.select`

If one has a good reason to favour one particular SWM, the function `mem.select` generates MEM variables from it, tests the global model, and selects an optimal subset of spatial predictors on the basis of one of the three above-mentioned criteria (`method = "FWD"`, `"MIR"`, or `"global"`). Although `listw.select` would also work in this case, `mem.select` provides a more suited and synthesized output when using a single SWM. Using `mem.select` instead of directly using `forward.sel` or `MEM.moransel` will ensure that the global test is not skipped previous to perform a selection, which will prevent type I error rate inflation issues (Blanchet et al. 2008, Bauman et al. 2018a). If the SWM is significant, the function returns the best subset of MEM variables obtained, and optionally returns the whole set of MEM variables generated.

If, for some reason, one wants to create the MEM variables associated to one or several SWMs regardless of the significance of the global model, the argument `alpha`, `p.adjust`, and `method` can be set to 1, FALSE, and `"global"`, respectively. However, this must not be done if working on actual data, as it would highly inflate the type I error rate and likely overcorrect the spatial autocorrelation.

In order to illustrate the use of `mem.select` with a single SWM, we will now use `listw.candidates` to build a single distance-based SWM corresponding to the widely-used PCNM (Borcard and Legendre 2002) and will test its significance and accuracy with respect to the `R2_reference` computed above in the first example.

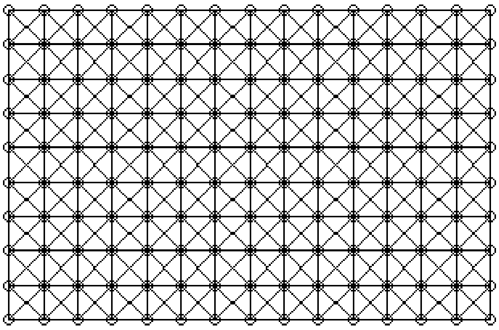
```
PCNM <- listw.candidates(xy, nb = "pcnm")
W_PCNM <- mem.select(Y_sampled, PCNM[[1]], MEM.autocor = "positive",
                     method = "FWD")

## Procedure stopped (alpha criteria): pvalue for variable 8 is 0.053000 (> 0.0500
00)

# Difference between the real R² and the R² obtained using PCNM:
diff <- round(R2_reference - max(W_PCNM$summary$AdjR2Cum), 2)
diff

## [1] 0.2

# % underestimation:
perc <- round(diff / R2_reference, 2) * 100
perc

## [1] 46
```

When used alone, the global model of **Y** against db-MEM_PCNM was significant, but performed poorly, yielding an underestimation of around 40% with respect to the real R², again illustrating that distance-based MEM are unsuitable for irregular sampling designs.

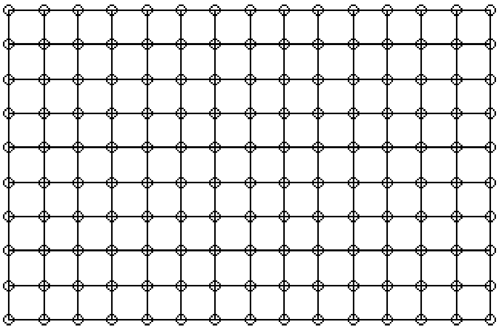
V. Regular sampling designs

Optimizing the selection of the SWM when the sampling design is roughly regular is less interesting, as the MEM variables originating from different SWMs detect roughly the same patterns (Dray et al. 2006). In those cases, rook (shared edge) or queen (shared edge or vertex) neighbor definitions or db-MEM can be used, for instance.

```
# We generate a regular grid of 25 x 25 sites:
grid <- expand.grid(x = seq(1, 15), y = seq(1, 10))
# We simulate abundance of a species distributed at random in the 150 sites:
set.seed(1)
y <- rnorm(nrow(grid))
# queen contiguity:
nb <- cell2nb(nrow = 15, ncol = 10, type = "queen")
plot(nb, grid)
```



```
# rook contiguity:
nb <- cell2nb(nrow = 15, ncol = 10, type = "rook")
plot(nb, grid)
```



```
# Distance-based criterion of connectivity:
# Let us consider that the sites distant from less than 2 distance units are
# connected:
DB <- listw.candidates(grid, nb = "dnear", d2 = 2, weights = "binary")
# We generate a SWM based on the rook criterion of connectivity (without weighting
# the connections), and we test the significance of the global model with
# `mem.select`:
lw2 <- nb2listw(nb, style = "B")
mem <- mem.select(y, lw2, MEM.autocor = "all")

## No significant positive spatial structure

## No significant negative spatial structure
```

No spatial structure could be detected in the response variable, as was expected from a random distribution.

References

Bauman D. et al. 2018a. Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors. Ecography, 41(10), 1638–1649.

Bauman D. et al. 2018b. Optimizing the choice of a spatial weighting matrix in eigenvector-based methods. Ecology: doi: 10.1002/ecy.2469.

Blanchet G. et al. 2008. Forward selection of explanatory variables. Ecology, 89(9), 2623–2632

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

Diniz-Filho J.A.F., Bini L.M. 2005. Modelling geographical patterns in species richness using eigenvector-based spatial filters. Global Ecology and Biogeography, 14(2), 177–185

Diniz-Filho J.A.F. et al. 2012. On the selection of phylogenetic eigenvectors for ecological analyses. Ecography, 35, 239–249

Dray S. et al. 2006. Spatial modeling: a comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM). Ecological Modelling, 196, 483–493

Griffith D. 2003. Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization. Springer, Berlin

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

Munoz, F. 2009. Distance-based eigenvector maps (DBEM) to analyse metapopulation structure with irregular sampling. Ecological Modelling, 220, 2683–-2689

Peres-Neto P., Legendre P. 2010. Estimating and controlling for spatial structure in the study of ecological communities. Global Ecology and Biogeography, 19, 174–184

Sidak Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association, 62(318), 626–633

Wagner H., Dray S. 2015. Generating spatially constrained null models for irregularly spaced data using Moran spectral randomization methods. Methods in Ecology and Evolution, 6, 1169–1178

Data S1: R code used for the simulations of the study.

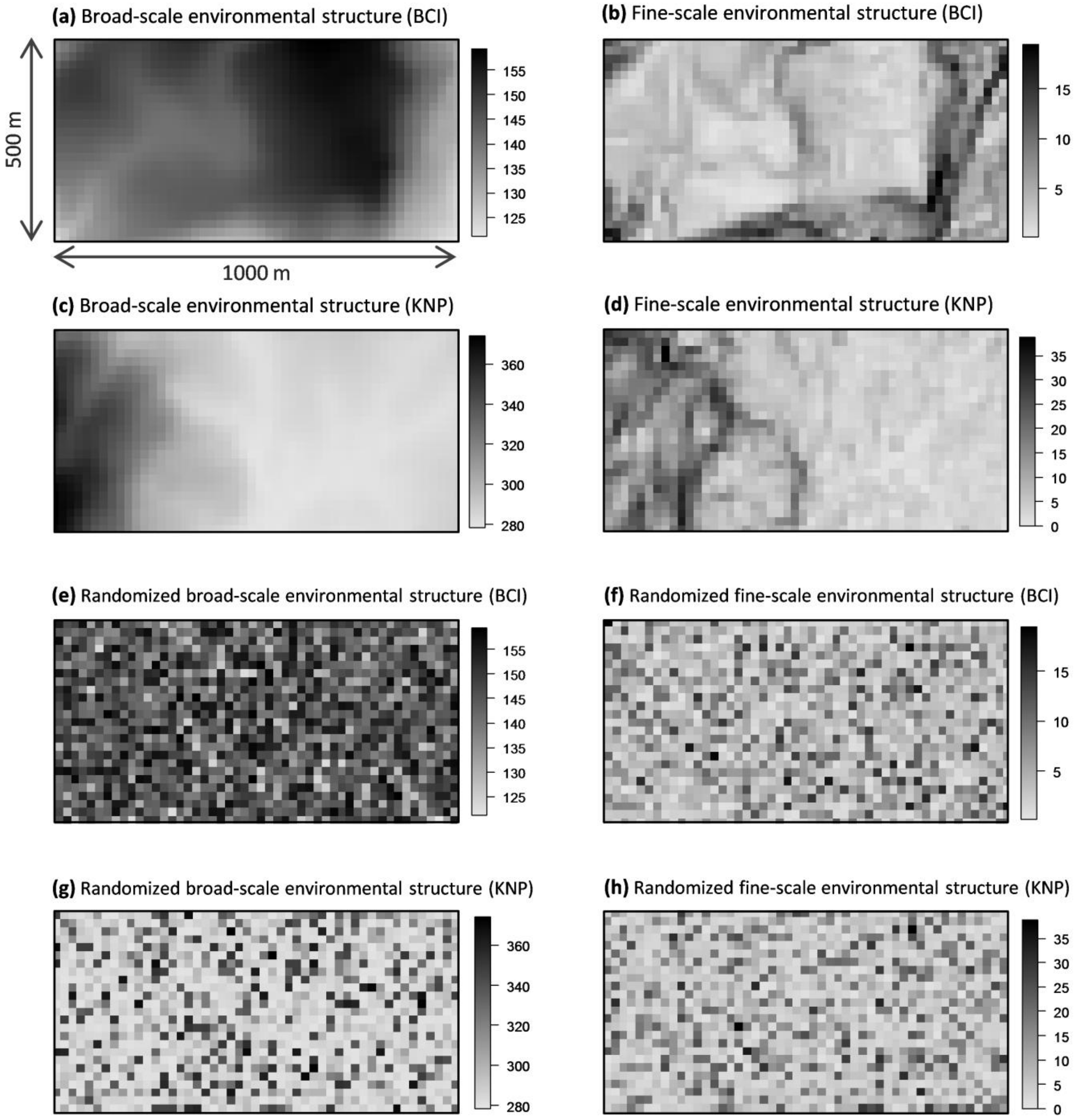
This appendix is too big to be presented here. It is available at *Ecology* online.

Appendix S4: Type I error rates of the different types of spatial weighting matrices considering a clustered and a random sampling design. Uniform, normal (mean of 0 and standard deviation of 1), exponential and cube exponential are the different types of random distributions that were used to test the type I error rate of the different SWMs.

Random distribution		**Clustered sampling design**				**Random sampling design**			
		uniform	**normal (0, 1)**	**exponent 1**	**exponent cube**	**uniform**	**normal (0, 1)**	**exponent 1**	**exponent cube**
Matrix B	*Matrix A*	*type I error rate*				*type I error rate*			
del	**Binary**	0.041	0.052	0.030	0.039	0.043	0.051	0.047	0.048
del	**Linear**	0.054	0.054	0.033	0.045	0.048	0.055	0.054	0.047
del	**Concave-down**	0.049	0.053	0.034	0.044	0.042	0.054	0.055	0.045
del	**Concave-up**	0.060	0.053	0.035	0.049	0.052	0.053	0.055	0.052
gab	**Binary**	0.048	0.041	0.045	0.041	0.058	0.039	0.050	0.048
gab	**Linear**	0.056	0.045	0.040	0.033	0.051	0.040	0.044	0.045
gab	**Concave-down**	0.059	0.040	0.044	0.037	0.056	0.039	0.045	0.045
gab	**Concave-up**	0.046	0.046	0.046	0.040	0.059	0.039	0.043	0.051
rel	**Binary**	0.040	0.032	0.047	0.046	0.063	0.047	0.045	0.043
rel	**Linear**	0.052	0.035	0.047	0.048	0.056	0.048	0.048	0.045
rel	**Concave-down**	0.049	0.040	0.042	0.049	0.061	0.049	0.043	0.044
rel	**Concave-up**	0.044	0.036	0.047	0.039	0.058	0.050	0.049	0.043
mst	**Binary**	0.051	0.042	0.039	0.047	0.058	0.046	0.045	0.036
mst	**Linear**	0.049	0.044	0.035	0.045	0.059	0.039	0.046	0.032
mst	**Concave-down**	0.051	0.042	0.038	0.051	0.059	0.051	0.050	0.040
mst	**Concave-up**	0.042	0.047	0.042	0.038	0.058	0.043	0.048	0.036
db	**Binary**	0.047	0.055	0.048	0.055	0.045	0.062	0.043	0.052
db	**Linear**	0.053	0.033	0.038	0.034	0.057	0.043	0.043	0.041
db	**Concave-down**	0.036	0.035	0.035	0.045	0.061	0.050	0.042	0.042
db	**Concave-up**	0.047	0.053	0.045	0.053	0.052	0.045	0.043	0.041
db-MEM$_{PCNM}$	**1-(D/4t)^2**	0.047	0.054	0.043	0.049	0.045	0.057	0.048	0.049

V. Chapitre V : Supporting information

Figure S1: Heat maps (1250 plots of 20 x 20 m each) showing the spatial structure of the eight environmental variables used in our study. (a) and (c): elevation (in m) in Barro Colorado Island (BCI) and Korup National Park (KNP), respectively. (b) and (d): topographical slope (BCI, KNP). (e) to (h): artificial variables created by randomly mixing plot from maps (a) to (d), respectively.


(a) Broad-scale environmental structure (BCI)
500 m
1000 m
155
150
145
140
135
130
125
(b) Fine-scale environmental structure (BCI)
15
10
5
(c) Broad-scale environmental structure (KNP)
360
340
320
300
280
(d) Fine-scale environmental structure (KNP)
35
30
25
20
15
10
5
0
(e) Randomized broad-scale environmental structure (BCI)
155
150
145
140
135
130
125
(f) Randomized fine-scale environmental structure (BCI)
15
10
5
(g) Randomized broad-scale environmental structure (KNP)
360
340
320
300
280
(h) Randomized fine-scale environmental structure (KNP)
35
30
25
20
15
10
5
0

Appendix S1: Methodological complements.

Appendix S1 – Section 1: Forward selection with double stopping criterion

Blanchet et al. (2008) highlighted two serious issues related to classical model selections, that is, inflated type I error rates and model overfitting. To overcome the former, they introduced a mandatory global test of significance prior to performing any kind of model selection. A model section can only be performed if the global model (i.e. the model including all explanatory variables) is significant at a predefined significance threshold. The overfitting problem was solved by introducing a second stopping criterion to the classical forward selection. While the classical procedure selected one by one the variables maximising the adjusted R^2 of the model and used the new *p*-value to confirm the integration of the variable, the forward selection with double stopping criterion adds the adjusted R^2 of the global model as a second stopping criterion. The selection procedure therefore stops either if the inclusion of the next variable yields a non-significant *p*-value or if the adjusted R^2 of the model exceeds the adjusted R^2 value obtained with all explanatory variables (global value). This improved forward selection has been shown to be unbiased and to have a correct type I error rate for both orthogonal (i.e. MEM) and non-orthogonal (i.e. environmental) variables (see details in Blanchet et al. 2008).

Appendix S1 – Section 2: Constructing a subset of spatial and environmental variables for the variation partitioning

The spatial variables used in the partitioning (also referred to as MEM variables) are a subset of selected spatial eigenvectors generated by Moran's eigenvector maps (MEM, Dray et al. 2006). While the method generates n – 1 spatial eigenvectors by diagonalising a doubly-centred spatial weighting matrix constructed from the spatial coordinates of the n sampled units (see details in Dray et al. 2006, Bauman et al. 2018b), the selection of a subset of MEM variables is necessary prior to performing the variation partitioning to avoid model overfitting (see Griffith 2003, Dray et al. 2006, Blanchet et al. 2008, Peres-Neto and Legendre 2010). The forward selection with double stopping criterion (Blanchet et al. 2008; see details in the Section 2 of this appendix) has recently been shown to be the most powerful and accurate eigenvector-selection method when the objective is to capture the spatial patterns as accurately as possible, as it is the case in variation partitioning (Bauman et al. 2018a). It has also been shown that optimising the selection of the subset of MEM variables by searching for the best one among a set of candidate spatial weighting matrices increased the power and accuracy of the method (Bauman et al. 2018b; see details Section 3). However the significance threshold of the global models tested for each spatial weighting matrix is penalised by the total number of candidates compared, so that the optimisation has to be conducted from a parsimonious set of realistic candidates (details in Bauman et al. 2018b).

A subset of environmental variables can also – but does not have to – be selected using the above-mentioned forward selection, or any other unbiased method, providing that the global model is significant (Blanchet et al. 2008).

Appendix S1 – Section 3: Optimising the selection of a spatial model

MEM variables can be generated from a multitude of spatial weighting matrices (SWM) (Dray et al. 2006). Although different SWMs yield similar MEM variables for regular sampling designs, the selection of a SWM is the most crucial step of MEM for irregularly-distributed sampled units (Dray et al. 2006, Bauman et al. 2018b). Bauman et al. (2018b) have recently shown that in the latter case, optimising the selection of the SWM from a set of candidates significantly improved both statistical power and spatial R² estimation accuracy (see details about different criteria of optimisation when they should be used in Bauman et al. 2018). The selection of the SWM should be optimised separately for the response variable(s) and the environmental variables to maximise both power and accuracy (see details in the illustrated tutorial of Appendix S4).

Appendix S1 – Section 4: Simulation Scenarios

For each of the ten distribution patterns (see Fig. 2, step 2), we generated 1000 replicates of tree populations independent of the environment, hereafter "*generalist populations*". In parallel, we also generated 1000 replicates of tree populations forced by the filtering of one of the three selected environmental variables displaying contrasted spatial structures (see first section of the methods and Fig. 2, step 1), hereafter "*specialist populations*". Thus, this yielded a total of 40 simulation scenarios (SS; see Table 1 and supplementary Table S1). A specialist population was created by first generating a generalist one, and then by keeping each individual according to a Gaussian probability density function, *f*(*x*), depending on the value of the "filtering environmental variables" in its quadrat,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad \text{Eqn. (1)}$$

where μ is an arbitrary optimum for the filtering environmental variable, and x is the value of this variable observed in the quadrat where the individual stands. μ was arbitrarily fixed at zero, after standardizing environmental variables to reach minimum and maximum values of zero and 1, respectively. As a consequence, the abundance was negatively correlated with the topographical slope or the elevation. σ was used as a selection coefficient that was fixed at 0.25, a value that allowed a strong selection of individuals (here, 15 to 40% of all the individuals were selected). To ensure identical sampling efforts between a specialist and a generalist population, we removed r randomly chosen individuals from the latter population, where r was the difference of abundance between the two populations before removal.

Appendix S1 – Section 5: Additional simulation scenarios using Moran's eigenvector maps

MEM variables were also used to build three additional SS to complete the 40 SS described in the main text. To do so, linear combinations of MEM variables to which a random noise was added were used to create spatially structured species and environmental variables. The orthogonality of MEM variables allowed for example creating a total spatial dependence or independence between the response and the environmental variable, by generating them using the same or different MEM variables. The first two additional SS corresponded to extreme cases in which (i) the environment and the abundance displayed independent spatial patterns, (ii) and the environment was entirely spatially structured and totally correlated to the spatial structure of the abundance. The third additional SS (iii) corresponded to a response variable strongly correlated to an environmental variable, itself negatively correlated to three MEM variables positively correlated to the response variable. This SS therefore simulated negative SSEF corresponding to an ISD (Legendre and Legendre 2012), hence allowing testing the statistical performances of our SSEF testing procedure for this situation.

The general strategy to build a spatially structured variable ***x*** (abundance or environmental variable) was to standardize (i.e., subtract the mean and divide by the standard deviation; STD in equation 2 below) (i) the sum of three to five MEM variables and (ii) a random normal noise, using equation 2. The variable was generated – following Bauman et al. (2018b) – as the sum of these two elements multiplied by α and 1 – α, respectively, where α is the degree of spatial autocorrelation, as illustrated by the following equation:

$$\boldsymbol{x} = \underbrace{[\alpha * (\mathrm{MEM_i} + \mathrm{MEM_j} + \mathrm{MEM_k})_{\mathrm{STD}}]}_{\boldsymbol{x}_{spatial}} + \underbrace{[(1-\alpha) * (\mathrm{noise}\ N(0,1))_{\mathrm{STD}}]}_{\boldsymbol{x}_{noise}} \qquad \text{Eqn. (2)}$$

The variables were all generated on the complete grid of 50 × 25 cells (i.e., the 1250 quadrats). The spatial weighting matrix used to generate the MEM variables considered, as connected, the quadrats sharing either an edge or a vertex (i.e. queen contiguity). We only used MEM variables associated to positive eigenvalues (i.e., modeling positive spatial autocorrelation) to build the response or environmental variables.

To generate the SS in which the environment and the abundance were independently spatially structured (i.e., no fraction [b] expected; **additional scenario 1**), the abundance was built using $\mathrm{MEM_3}$, $\mathrm{MEM_4}$, $\mathrm{MEM_5}$, $\mathrm{MEM_6}$, and $\mathrm{MEM_7}$, while the filtering environmental variable was built with $\mathrm{MEM_{37}}$, $\mathrm{MEM_{55}}$, $\mathrm{MEM_{75}}$, and $\mathrm{MEM_{85}}$. For the environmental variable (***env*** in equation 3 below) to be related to the abundance (***y*** in equation 4) while remaining spatially independent, the abundance was correlated to the noise component of the environmental variable (i.e., $\boldsymbol{env}_{noise}$):

$$\boldsymbol{env} = [0.8 * (\mathrm{MEM_{37}} + \mathrm{MEM_{55}} + \mathrm{MEM_{75}} + \mathrm{MEM_{85}})_{\mathrm{STD}}] + [0.2 * (\mathrm{noise}\ N(0,1))_{\mathrm{STD}}] \qquad \text{Eqn. (3)}$$

$$\boldsymbol{y} = [0.45 * (\mathrm{MEM_3} + \mathrm{MEM_4} + \mathrm{MEM_5} + \mathrm{MEM_6} + \mathrm{MEM_7})_{\mathrm{STD}}] + [0.05 * (\mathrm{noise}\ N(0,1))_{\mathrm{STD}})] + [0.5 * (\boldsymbol{env}_{noise})_{\mathrm{STD}}] \qquad \text{Eqn. (4)}$$

To generate the SS in which ***env*** is entirely comprised in the spatial structure of ***y*** (**additional scenario 2**), ***env*** was built using a subset of the MEM variables used to build ***y*** (equations 5 and 6, below):

$$\boldsymbol{env} = [\ 0.8 * (\mathrm{MEM_4} + \mathrm{MEM_5} + \mathrm{MEM_6})_{\mathrm{STD}}] + [0.2 * (\mathrm{noise}\ N(0,1))_{\mathrm{STD}}] \qquad \text{Eqn. (5)}$$

$$\boldsymbol{y} = [0.8 * (\mathrm{MEM_3} + \mathrm{MEM_4} + \mathrm{MEM_5} + \mathrm{MEM_6} + \mathrm{MEM_7})_{\mathrm{STD}}] + [0.2 * (\mathrm{noise}\ N(0,1))_{\mathrm{STD}})] \qquad \text{Eqn. (6)}$$

In the last additional scenario (**additional scenario 3**), ***env*** was strongly negatively correlated to the MEM variables to which ***y*** was positively correlated (see equations 9 and 10). In addition, the identity of the MEM variables used to generate ***y*** was selected randomly between MEM4 and MEM50 so that ***y*** displayed a varying broad, intermediate, or fine-scaled spatial pattern.

$$\boldsymbol{y} = [0.4 * (\mathrm{MEM}_x + \mathrm{MEM}_y + \mathrm{MEM}_z)_{\mathrm{STD}}] + [0.4 * (\boldsymbol{env})_{\mathrm{STD}})] + [0.2 * (\text{noise } N(0, 1))_{\mathrm{STD}})] \qquad \text{Eqn. (9)}$$

$$\boldsymbol{env} = [-0.4 * (\mathrm{MEM}_x + \mathrm{MEM}_y + \mathrm{MEM}_z)_{\mathrm{STD}}] + [0.6 * (\text{noise } N(0, 1))_{\mathrm{STD}})] \qquad \text{Eqn. (10)}$$

The variables ***env*** and ***y*** generated on the complete grid (1250 cells) with the three above-described additional SS were then sampled in the same 153 quadrats as in the other 40 SS (Fig. 2, step 2). Distance-based MEM variables were generated on the basis of the sampled quadrats, using the same spatial weighting matrix described in the main text, and the performance of the SSEF testing procedure was then assessed as described in the section *Statistical performance of the SSEF testing procedure* in the Materials and Methods, after 1000 simulations.

The construction of the three additional scenarios can be reproduced using the R code provided in Appendix S3.

Appendix S1 – Section 6: Sampling Design

For each SS, the environmental and the abundance data was sampled in 153 regularly-spaced quadrats among the 1250 quadrats of the 50-ha study area (12.24% of the whole area's surface, 6.12 ha; see Fig. 2, steps 2), a realistic sampling effort that fits the range of sampling sizes found in the literature (e.g., Kadavul and Parthasarathy 1999, Condit et al. 2004, Zent and Zent 2004, Biwolé et al. 2015, Muledi et al. 2017). These sampled quadrats were regularly disposed according to a 9 by 17 grid-like structure (Fig. 2, step 2), a spatial configuration that allowed performing torus-translations. Adjacent sampled quadrats were distant of 60 m along both the X and Y axes of the grid. Under this sampling scheme, the parameter *m* of the Poisson Cluster Process (PCP, Plotkin et al. 2000) was randomly modulated but within limited range of values, so that that the total number of individuals eventually sampled per simulation of aggregated tree distribution (in the 153 quadrats) reached an average of 200, 400, and 600 ± 50 (min/max) individuals, for the three increasing values of *ρ* (see Fig. 2), respectively. For the distribution pattern 10 (spatially randomized tree distribution), the average number of individuals was fixed at the intermediate value of 400 ± 50.

References

Bauman D. et al. 2018a. Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors. – Ecography 41(10): 1638–1649.

Bauman D. et al. 2018b. Optimizing the choice of a spatial weighting matrix in eigenvector-based methods. – Ecology: doi: 10.1002/ecy.2469.

Biwolé A. B. et al. 2015. New data on the recent history of the littoral forests of southern Cameroon: an insight into the role of historical human disturbances on the current forest composition. – Plant Ecol. Evol. 148: 19–28.

Blanchet F. G. et al. 2008. Forward selection of explanatory variables. – Ecology 89: 2623–2632.

Condit R. et al. 2004. Tropical forest dynamics across a rainfall gradient and the impact of an El Niño dry season. – J. Trop. Ecol. 20: 51–72.

Dray S. et al. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). – Ecol. Model. 196: 483–493.

Griffith D. 2003. Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization. – Springer.

Kadavul, K., Parthasarathy N. 1999. Plant biodiversity and conservation of tropical semi-evergreen forest in the Shervarayan hills of Eastern Ghats, India. – Biodivers. Conserv. 8: 419–437.

Legendre, P. and Legendre, L. 2012. Numerical Ecology, 3rd English edn. Elsevier Science BV, Amsterdam.

Muledi J. I. et al. 2017. Fine-scale habitats influence tree species assemblage in a Miombo Forest. – J. Plant Ecol. 10: 958–959.

Peres-Neto P. R., Legendre, P. 2010. Estimating and controlling for spatial structure in the study of ecological communities. – Global Ecol. Biogeogr. 2: 174–184.

Plotkin J. B. et al. 2000. Species-area Curves, Spatial Aggregation, and Habitat Specialization in Tropical Forests. – J. Theor. Biol. 207: 81–99.

Zent E. L., Zent S. 2004. Floristic composition, structure, and diversity of four forest plots in the Sierra Maigualida, Venezuelan Guayana. – Biodivers. Conserv. 13: 2453–2483.

Appendix S2: Suppressor variables in variation partitioning between environmental and spatial components

I. Negative shared space-environment fractions

In a variation partitioning of a response variable (*y*) or a matrix of response variables between a set of environmental and a second set of MEM variables (*env* and *space*, respectively), a negative adjusted shared space-environment fraction (SSEF, or fraction [b]) can arise for three different reasons:

- because of the adjustement procedure (Peres-Neto et al. 2006) applied to the fractions used to calculate the SSEF,
- when *env* is negatively correlated to *space* (MEM variables), but both components are positively correlated to *y* (Legendre and Legendre 2012),
- when a suppressor effect is present.

While a negative SSEF caused by the first point is ecologically meaningless and should be considered as a zero, a negative SSEF generated by the second reason mentioned above bears real ecological information and needs to be tested. These two points are tackled explicitly in the paper and are well separated by the function *envspace.test*, statistically speaking (see *Results* section).

However, while negative SSEF caused by the adjustment procedure have a positive unadjusted value and can therefore be easily detected, the negative SSEF caused by both the second and the third points mentioned above have negative unadjusted values, and will therefore be tested by *envspace.test* (see *Material and methods*). Since only negatively correlated explanatory components correlated to *y* have real ecological meaning, suppressor effects must be discarted prior to performing the variation partitioning to avoid misinterpretations of the SSEF.

This appendix details how suppression effects can be avoided.

A suppressor variable (i.e. a suppressor) is an explanatory variable uncorrelated (classic suppression) or very little correlated (negative suppression) to *y*, but correlated to one or severall other explanatory variables, themselves correlated to *y* (the predictors) (Azen and Budescu 2003, Beckstead 2012, Ray-Mukherjee et al. 2014). The portion of predictor variability the suppressor is correlated to is the portion of variability of the predictors that is not correlated to *y*. As a consequence, the inclusion of a suppressor in a model enhances the overall explanatory power of the model as a result of a suppression of a portion of irrelevant variability in the predictor(s). In this case, although a model of *y* against the suppressor is generally not significant and explains nearly nothing, the inclusion of the suppressor in the model including the predictors explains a (much) larger portion of *y* than the model of *y* against the predictors without the suppressor.

This could happen, for example, if:

- *env* is related to one or several MEM variables of *space* that describe its spatial structure but not those of *y*,
- the spatial structure of *y* is explained by other MEM variables of *space* (by definition uncorrelated to the MEM variables related to *env*, as MEM variables are orthogonal),
- *env*, alone, is not related to *y*.

A simple way to avoid suppression is to perform the partitioning if and only if both explanatory components (*env* and *space*), tested separately, are significantly related to *y*. This condition is already automatically tested by *envspace.test*.

However, *env* may be slightly significant, in the case of a negative suppression caused by a small correlation between the suppressor and *y*.

Therefore, a complementary way to be sure to avoid suppression effects is to perform an appropriate selection of MEM variables to be integrated to the *space* component, prior to performing the partitioning. Indeed, if all the MEM variables of *space* are related to *y*, the suppression effect described in the three bullets above will not happen, as the MEM variables describing structures of *env* but not *y* will have been removed from *space*.

This is illustrated in the two following examples.

II. Simulation of a suppression effect

```
library(adespatial)
library(spdep)
library(vegan)
```

Creation of sampled grid of 17 x 9 regularly-spaced sampled units:

```
grid <- expand.grid(x = seq(1:17), y = seq(1:9))
```

Creation of a distance-based spatial weighting matrix:

```
w_grid <- listw.candidates(grid, nb = "pcnm")
```

Generation of the positively autocorrelated MEM variables (*space* component):

```
MEM <- scores.listw(w_grid[[1]], MEM.autocor = "positive")
```

The response variable (*y*) is generated to be correlated to a few MEM variables of *space*:

```
set.seed(1)
sel.mem.y <- seq(2, 4, 1)
mem.y <- MEM[, 1]
for (i in sel.mem.y) mem.y <- mem.y + MEM[, i]
y <- 0.4 * scale(mem.y) + 0.6 * scale(rnorm(length(mem.y)))
```

Creation of an environmental variable very slightly correlated to *y* and highly correlated to MEM variables of *space* that are not correlated to *y*:

```
sel.mem.env <- seq(7, 20, 1)
mem.env <- MEM[, 6]
for (i in sel.mem.env) mem.env <- mem.env + MEM[, i]

env <- 0.17 * scale(y) + 0.8 * scale(mem.env) +
       0.13 * scale(rnorm(length(mem.env)))
```

Is *env* significantly related to *y*?

```
(tes.env <- anova(rda(y, env)))

## Permutation test for rda under reduced model
## Permutation: free
## Number of permutations: 999
##
## Model: rda(X = y, Y = env)
##           Df Variance      F Pr(>F)
## Model      1  0.01702 4.8581  0.033 *
## Residual 151  0.52892
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The environment is slightly significantly related to *y*.

Unadjusted and adjusted SSEF:

```
# Unadjusted value ([ab] - [a]; see Fig. 1 of the paper):
RsquareAdj(rda(y, env))$r.squared - RsquareAdj(rda(y, env, MEM))$r.squared

## [1] -0.1353896

# SSEF calculated from the adjusted fractions [ab] and [a]:
(varpart(y, env, MEM)$part$indfract[2, 3])

## [1] -0.3044937

# Test of the SSEF:
SSEF.test <- envspace.test(y, env, grid, MEM, w_grid[[1]],
                           MEM.autocor = "positive", summary = TRUE)

##
## ***********************************************************************
## The R2 value of the shared space-environment fraction (R2adj) and the correspon
ding
## p-value computed with the MSR test are:
## R2adj = -0.304, p-value MSR = 0.001.
## ***********************************************************************
```

The SSEF is substantial and detected as significant despite the fact that *env* is nothing more than a suppressor with a very slight effect on *y*. Although this seems very much like an induced spatial dependence as the one generated by the second cause of negative SSEF presented at the beginning of this appendix, it has nothing to do with it. We therefore understand how suppression effects could lead to wrong ecological interpretations of the SSEF.

Let us see if, as expected, the selection of an appropriate subset of MEM variables makes the suppression disappear. The forward selection with double stopping criterion (Blanchet et al. 2008) is the most powerful and accurate selection method when the objective is to capture as much spatial structure of *y* as possible (Bauman et al. 2018).

```
MEM.FWD <- mem.select(y, w_grid[[1]], MEM.autocor = "positive",
                      method = "FWD")$MEM.select

## Procedure stopped (adjR2thresh criteria) adjR2cum = 0.392437 with 6 variables
(> 0.376512)
```

Unadjusted and adjusted SSEF:

```
# Unadjusted value:
RsquareAdj(rda(y, env))$r.squared - RsquareAdj(rda(y, env, MEM.FWD))$r.squared

## [1] 0.0213239

# SSEF calculated from the adjusted fractions [ab] and [a]:
(varpart(y, env, MEM.FWD)$part$indfract[2, 3])

## [1] 0.01877471
```

As expected, the suppression effect has be “suppressed” and there are no more risk to mistaken the suppression effect for an induced spatial dependence.

```
SSEF.test <- envspace.test(y, env, grid, MEM.FWD, w_grid[[1]],
                           MEM.autocor = "positive", summary = TRUE)

## 
## ************************************************************************
## The R2 value of the shared space-environment fraction (R2adj) and the correspon
ding
## p-value computed with the MSR test are:
## R2adj = 0.019, p-value MSR = 0.007.
## ************************************************************************
```

The effect is still detected as significant, but it is so small that it bears no important knowledge.

Commonality analysis can also be used in multiple regressions as a very useful tool to further explore potential suppressor variables (see details in Ray-Mukherjee et al. 2014).

References

Azen R., Budescu D.V. 2003. The dominance analysis approach for comparing predictors in multiple regression. Psychological Methods, 8(2), 129–148

Peres-Neto P. et al. 2006. Variation partitioning of species data matrices: estimation and comparison of fractions. Ecology, 87(10), 2614–2625

Beckstead J. 2012. Isolating and examining sources of suppression and multicollinearity in multiple linear regression. Multivariate Behavioral Research, 47(2), 224–246

Legendre P., Legendre L. 2012. Numerical ecology. Elsevier, Amsterdam

Ray-Mukherjee J. et al. 2014. Using commonality analysis in multiple regressions: A tool to decompose regression effects in the face of multicollinearity. Methods in Ecology and Evolution, 5(4), 320–328

Appendix S3: R code used for the simulation study.

This appendix is too big to be presented here. It is available online.

Appendix S4: Tutorial for the test of the shared space-environment fraction of a variation partitioning

Useful packages

```
library(adespatial)
library(spdep)
library(vegan)
```

Objective of this document

This document is an appendix of the paper entitled *Testing and interpreting the shared space-environment fraction in variation partitioning analyses of ecological data*, by Bauman D., Vleminckx J., Drouet T. and Hardy O. (2018). It provides a tutorial to illustrate the test of the shared space-environment fraction of a variation partitioning (SSEF, or fraction [b]), using the function `envspace.test` of the `adespatial` package. The information presented in this tutorial is detailed in the help document of the above-mentioned function.

The oribatid mite data will be used as an example (Borcard et al. 1992, 1994 for details on the dataset).

I. Concept of the method

This document focuses on the case of a variation partitioning of a response variable or matrix (`Y`) against an environmental component (`env`) and a spatial component (i.e. spatial eigenvectors, hereafter MEM variables, `MEM`; Borcard and Legendre 2002, Dray et al. 2006).

The adjusted R2 (Peres-Neto et al. 2006; R2adj) of the SSEF is not an actual R2, as it is computed by subtracting the R2 of other fractions (adjusted following Peres-Neto et al. 2006) of the analysis and therefore has zero degree of freedom. The SSEF can therefore not be tested in the classical way (permutation of the model residuals, Anderson and Legendre 1999). The function `envspace.test` provides two ways of testing this fraction using constrained null models: torus-translations (or torus randomisation, Upton and Fingleton 1985) (TT), for regular sampling designs only, and Moran spectral randomisations (MSR, Wagner and Dray 2015), for any kind of sampling design.

A condition prior to performing a variation partitioning is that both the global environmental and the global spatial models must be significant. Otherwise, the partitioning would make no sense.

If, these two preliminary conditions are respected, then the function `envspace.test` first checks whether `env` displays a significant spatial structure (using function `mem.select`), and the function proceeds to the test of the SSEF if, and only if `env` is spatially structured. Otherwise, the SSEF should not be considered (and is therefore not tested).

`envspace.test` is based on the variation partitioning and on TT and/or MSR. These methods can all be equally used for univariate and multivariate response data (e.g. abundance of one species, or community composition data, respectively). As a consequence, `envspace.test` can also be used for both univariate and multivariate response data. In this tutorial, we illustrate the function on a multivariate `Y`, hence using redundancy analyses (RDA).

For more details:

```
help(envspace.test)
help(varpart)
help(msr)
```

II. Data input

Input of the community data (Y) and Hellinger-transformation to prepare the data for the redundancy analyses (RDA) of the partitioning (see Legendre and Gallagher 2001):

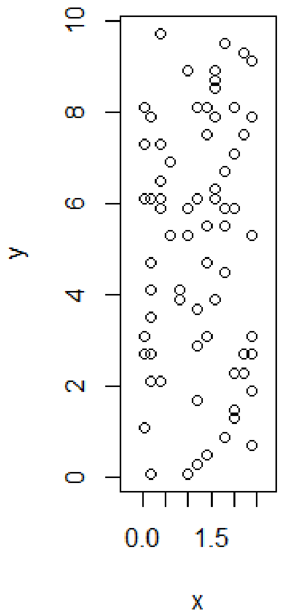
```
data(mite)
Y <- decostand(mite, method = "hellinger")
dim(Y)

## [1] 70 35
```

Environmental explanatory dataset (here, we only use the two quantitative variables, i.e., the subtrate density and water content of the substrate):

```
data(mite.env)
env <- mite.env[, 1:2]
dim(env)

## [1] 70  2
```

Coordinates of the 70 sites:

```
data(mite.xy)
coord <- mite.xy
```



Spatial coordinates of the 70 sampled sites

III. Generation of MEM variables

MEM variables should be constructed separately for Y and env, as the spatial weighting matrix (SWM) allowing an optimal detection of the spatial patterns of Y and env may not be the same. Here, since the objective is to capture the spatial structures of Y as accurately as possible, the optimisation of the selection of a subset of MEM variables for Y is based on the forward selection with double stopping criterion of Blanchet et al. (2008), following Bauman et al. (2018a,b). For env, the SWM must be optimised on the basis of the global models of the different SWMs, because the MSR will use all MEM variables to randomise `env` (see details on the optimisation of a SWM and a subset of MEM variables in Bauman et al. 2018a, b; see also `help(mem.select)`).

The MEM variables generated for `env` will allow 1) testing for a spatial structure in `env` (preliminary condition for testing the SSEF), and 2) using MSR to test the SSEF.

The MEM variables generated on the basis of `Y` will allow 1) testing for a spatial structure in `Y` (preliminary condition for testing the SSEF), and 2) will be used as the spatial component for the variation partitioning.

III.1. Optimisation of a SWM for `Y` and `env`

The optimisation is performed in two steps, following Bauman et al. 2018a: 1) generate a list of the candidate SWMs to be compared, then 2) test all candidates while correcting the significance threshold according to the number of candidates tested, and keep the one yielding the maximum coefficient of determination (R2; see details in the help files of functions `listw.candidates` and `listw.select`).

We create five candidate SWMs: a connectivity matrix based on a Gabriel graph, and on a minimum spanning tree (i.e., two contrasted graph-based SWMs), either with no weights on the connexions (binary SWMs), or weighted by a function decreasing linearly with the distance. We also add a distance-based SWM:

```
candidates <- listw.candidates(coord, nb = c("gab", "mst", "pcnm"),
                               weights = c("binary", "flin"))
```

Optimising the selection of a subset of MEM variables and a SWM (see details in Bauman et al. 2018a):

For Y:

```
modsel.spe <- listw.select(Y, candidates, MEM.autocor = "positive",
                           method = "FWD")
```

For `env` (selection on standardised ‘env’ because the standardised env will be used to test the SSEF, when using envspace.test):

```
modsel.env <- listw.select(scale(env), candidates, MEM.autocor = "positive",
                           method = "global")
```

Note that, if needed for some reason, MEM can be generated from any of the other candidate SWMs, using function *scores.listw* or *mem.select*.

```
# Generation of the positively autocorrelated MEM variables from the SWM
# consisting of a minimum spanning tree weighting by the linear function:
mem_mst.flin <- scores.listw(candidates$MST_Linear, MEM.autocor = "positive")
```

The function *mem.select* should be prefered, however, to generate the MEM variables from one given SWM. Indeed, it uses *scores.listw*, but has the advantage to perform a global test of significance on the whole set of MEM variable, and to generate the MEM variables if and only if the global model is significant. This control avoids proceeding to the selection of a subset of spatial predictors with an inflated type I error rate (Blanchet et al. 2008, and see details in Bauman et al. 2018b). The function *mem.select*, in addition, provides the option of performing the selection of a subset of MEM variables according to two possible selection methods (if the global model is significant), that is, maximising the adjusted R², or minimising the residual spatial autocorrelation with a minimum number of MEM variables (details in Bauman et al. 2018a). The function *listw.candidates* - which optimises the selection of a SWM and optionally of a subset of MEM variables among the candidate SWMs - is based on *mem.select*.

```
# Example: We test for Y the SWM 'Gabriel_linear', generate the MEM and use the
# forward selection with double stopping criterion (Blanchet et al. 2008) to
# select a subset ofpredictors:
select <- mem.select(Y, candidates$Gabriel_Linear, MEM.autocor = "positive")
```

The best practice yielding the maximum statistical power and accuracy of MEM has however recently been shown to consist in optimising the selection of a SWM among a set of carefully-chosen candidates (function *listw.select*, see details in Bauman et al. 2018a).

IV. Variation partitioning and test of the SSEF

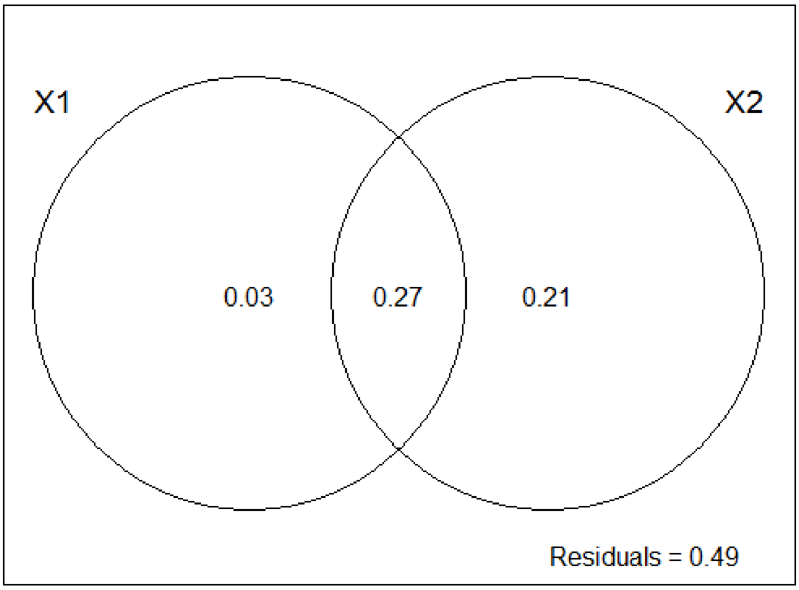
We checked that both Y and env displayed significant spatial structures. We still need to check whether Y is significantly related to env prior to perform the partitioning.

```
(global.env <- anova(rda(Y, env))$Pr[1])

## [1] 0.001
```

The global environmental model is significant.

We partition the variation of Y into an environmental component and a spatial component.

```
# Subset of selected MEM variables within the best SWM:
MEM.spe <- modsel.spe$best$MEM.select

VP <- varpart(Y, env, MEM.spe)
# SSEF:
(SSEF <- VP$part$indfract$Adj.R.square[2])

## [1] 0.2727156
```



Venn diagram of the variation partitioning results

Test of the SSEF using `envspace.test`: (Since we generated MEM variables associated to positive eigenvalues only for the species, we do the same here for the environment)

```
SSEF.test <- envspace.test(spe = Y, env = env, coord = coord, MEM.spe = MEM.spe,
                           listw.env = candidates[[modsel.env$best.id]],
                           MEM.autocor = "positive", regular = FALSE, nperm = 999)

# SSEF:
SSEF.test$obs

## [1] 0.2727156

# Significance of the permutation test performed with MSR:
SSEF.test$pval

## [1] 0.001
```

The SSEF value is therefore very unlikely to have been produced by chance, and we can proceed to the interpretation of this portion of variation of Y correlated to a spatially structured environment.

This fraction may consist in an induced spatial dependence (Legendre and Legendre 2012), when the environmental variable causing [b] is among the variables measured and included in env. In this case, mapping the spatial structure(s) shared by Y and env (to which corresponds the amount of variation captured by fraction [b]) can help indicating at what spatial scale the environmental variable induces a structure in Y.

It is worth reminding that the purpose of variation partitioning is not to establish causal relations, so that inferring causal relations from the SSEF or the other fractions of the partitioning should always be done with caution. In addition, a priori hypotheses and ecological theory or knowledge should underlie the choice of the environmental variables to be measured and included in the analysis to allow strong interpretations of the SSEF and the other fractions of the analysis (see McIntire and Fajardo 2009). It is also worth mentioning that part of an actual induced spatial dependence may still remain in fraction [c] (i.e. the pure spatial component), in the case where a variable responsible for a pattern in Y has not been accounted for.

References

Anderson M., Legendre P. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. Journal of Statistical Computation and Simulation, 62, 271–303

Bauman D. et al. 2018a. Optimizing the choice of a spatial weighting matrix in eigenvector-based methods. Ecology, doi: 10.1002/ecy.2469.

Bauman D. et al. 2018b. Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors. Ecography, 41(10), 1638–1649.

Blanchet G. et al. 2008. Forward selection of explanatory variables. Ecology, 89(9), 2623–2632

Borcard D. et al. 1992. Partialling out the spatial component of ecological variation. Ecology, 73(3), 1045–1055

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

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

Dray S. et al. 2006. Spatial modeling: a comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM). Ecological Modelling, 196, 483–493

Legendre P., Gallagher E.D. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia, 129(2), 271–280

Legendre P., Legendre L. 2012. Numerical ecology. Elsevier, Amsterdam

McIntire E. J. B., Fajardo A. 2009. Beyond description: the active and effective way to infer processes from spatial patterns. – Ecology 90: 46–-56

Peres-Neto P. et al. 2006. Variation partitioning of species data matrices: estimation and comparison of fractions. Ecology, 87(10), 2614–2625

Wagner H., Dray S. 2015. Generating spatially constrained null models for irregularly spaced data using Moran spectral randomization methods. Methods in Ecology and Evolution, 6, 1169–1178

Upton G.J.G., Fingleton B. 1985. Spatial data analysis by example. Vol. 1: Point pattern and quantitative data. Wiley, NewYork.

Table S1: Standardised values (centred on their mean and divided by their standard deviation) of the eight environmental variables used in the simulations.

The table is too big to be included here. It is available online.

Table S2. Detailed characteristics of the 40 main simulation scenarios (SS). The first column refers to the number of the SS described. Column 2 indicates whether no environmental filtering is simulated (SS 1 to 10) or if there is an environmental filtering by a spatially randomised variable (SS 11 to 20) or by a variable displaying fine-scale (SS 21 to 30) or broad-scale spatial structure (SS 31 to 40) (See also Fig. 2, step 1). Column 3: Type of tree distribution pattern (DP; see Fig. 2, step 2): aggregated (DP 1 to 9) or spatially random (DP 10). Column 4 indicates the type of statistical performance investigated: type I error rate if the SS does not model an induced spatial dependence effet (ISD; SS 1 to 20), or statistical power if it does model an ISD (SS 21 to 40). Column 5 indicates whether the SS simulates an ISD (SS 21 to 40) or not (SS 1 to 20).

SS	Environmental filter	DP	Aggregated DP	Performance	ISD simulated
1	None	1	yes	Type I error rate	No
2		2			
3		3			
4		4			
5		5			
6		6			
7		7			
8		8			
9		9			
10		10	no		
11	Spatially random	1	yes		
12		2			
13		3			
14		4			
15		5			
16		6			
17		7			
18		8			
19		9			
20		10	no		
21	Structured (fine-scale)	1	yes	Statistical power	Yes
22		2			
23		3			
24		4			
25		5			
26		6			
27		7			

28		8	
29		9	
30		10	no
31	Structured (broad-scale)	1	yes
32		2	
33		3	
34		4	
35		5	
36		6	
37		7	
38		8	
39		9	
40		10	no

Table S3: Frequency of selection (over 1000 replicated SS) of each MEM modelling positively autocorrelated tree distribution patterns in the 30 SS corresponding to specialist tree distributions

The table is too big to be presented here. It is available online.

Table S4. Type I error rate and/or statistical power (indicated in column 2) of the SSEF test (TT/MSR: SSEF test using torus-translations and Moran Spectral randomisations, respectively) associated to the three additional scenarios. Additional scenarios 1 to 2 represent the extreme cases where: (i) the environment and the abundance displayed totally independent spatial patterns (additional scenario 1), and (ii) the environment was entirely spatially structured and totally correlated to the spatial structure of the abundance (additional scenario 2). Additional scenario 3 aimed at testing the type I error rate and the statistical power of the SSEF test when simulating negative unadjusted R^2 values for this fraction (tests performed using MSR only; *np*: not performed).

	Statistical performance	MSR	TT
Additional scenario 1	Type I error rate	0	0
Additional scenario 2	Statistical power	0.984	0.972
Additional scenario 3	Type I error rate	0.007	*np*
	Statistical power	0.784	*np*

Table S5: Mean and standard deviation values (calculated over the 1000 replicated simulations) of the SSEF in each simulation scenario

		No environmental filtering									
		SS 1	**SS 2**	**SS 3**	**SS 4**	**SS 5**	**SS 6**	**SS 7**	**SS 8**	**SS 9**	**SS 10**
GENERALIST populations	**standard deviation**	0.0197	0.0232	0.0208	0.0216	0.0331	0.0278	0.0367	0.0333	0.0305	0.0168
	mean	-0.0009	0.0057	0.0061	0.0022	0.0109	0.0139	0.0194	0.0182	0.0179	-0.0009

		Spatially random filtering environmental variable									
		SS 11	**SS 12**	**SS 13**	**SS 14**	**SS 15**	**SS 16**	**SS 17**	**SS 18**	**SS 19**	**SS 20**
SPECIALIST populations	**standard deviation**	0.0241	0.0252	0.0199	0.0229	0.0252	0.0177	0.0343	0.0250	0.0264	0.0137
	mean	0.0071	0.0034	-0.0103	-0.0109	-0.0051	-0.0123	-0.0035	-0.0093	-0.0124	-0.0085

		Fine-scale filtering environmental variable									
		SS 21	**SS 22**	**SS 23**	**SS 24**	**SS 25**	**SS 26**	**SS 27**	**SS 28**	**SS 29**	**SS 30**
SPECIALIST populations	**standard deviation**	0.0462	0.0502	0.0503	0.0455	0.0394	0.0391	0.0537	0.0399	0.0587	0.0572
	mean	0.0407	0.0875	0.1117	0.0775	0.1463	0.1784	0.1017	0.1671	0.1981	0.1816

		Broad-scale filtering environmental variable									
		SS 31	**SS 32**	**SS 33**	**SS 34**	**SS 35**	**SS 36**	**SS 37**	**SS 38**	**SS 39**	**SS 40**
SPECIALIST populations	**standard deviation**	0.0645	0.0867	0.0771	0.0707	0.0587	0.0679	0.0584	0.0595	0.0633	0.0781
	mean	0.1372	0.2309	0.2932	0.2514	0.3353	0.3860	0.3204	0.3676	0.4605	0.1169

		additional scenario 1	**additional scenario 2**
SPECIALIST populations	**standard deviation**	0.0000	0.0732
	mean	0.0000	0.3744

VI. Chapitre VI : Supporting information

Figure S1: Schematic map of Treignes forest plot on a background of the elevation. The north is indicated by the thick black arrow in the upper-right corner. The × are the limits of the 143 12.5 × 12.5 m quadrats. The zoom on a quadrat represents the locations of the five soil cores taken in all quadrats (see *Material and methods* section).

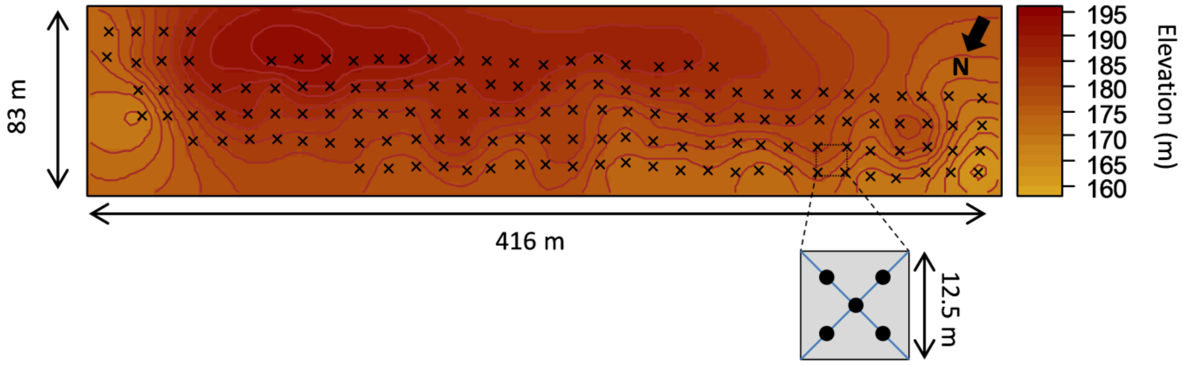
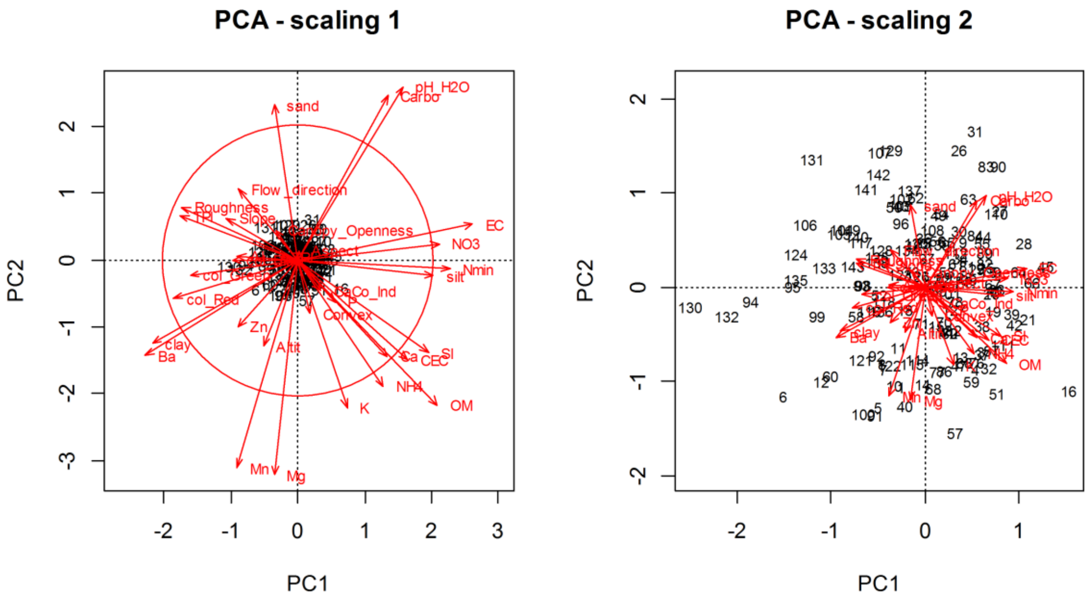


Figure S2: Principal component analysis (PCA) on the correlations of the soil, light intensity, and topography variables. The red circle of the left-handed plot is a circle of equilibrium contribution, whose radius is equal to $\sqrt{\boldsymbol{d/p}}$, where d is the number of axes displayed (i.e. 2), and p is the number of dimensions of the PCA (i.e. 29). The radius represents the length of a vector that would contribute equally to all dimensions, so that the vectors longer than the radius can be interpreted confidently in the represented plane.



Appendix S1: Material and method complements

Appendix S1 – Section 1: Environmental variables

Soil: For each soil sample, a stoniness index (SI) was estimated on the field after sieving (2-mm mesh) by a discreet quantitative index taking the values of 0 (null SI), 1 (half of the sieve area or less covered by gravels) or 2 (between half to all the sieve area covered). The pH-H2O and the electrical conductivity (EC) were respectively measured with glass electrodes (Mettler-Toledo) and a conductimetre (VWR EC300) in a soil suspension in distilled water (ratio 1:5 between soil and water). Organic matter content (OM) was calculated by mass loss of a sample after dry ashing at 550 °C. The RGB colour components of soils were determined from scans of dampened samples spread on a sheet. The carbonate index is a discreet quantitative index depending on the intensity of the effervescence reaction of a soil sample in contact with HCl (0, 1, or 2). The next soil parameters were measured on a composite sample of the five soil cores of the quadrat. Soil texture was determined by wet sieving and the pipette method after OM destruction with H_2O_2 and clay dispersion by Na citrate. The plant-available elements (P, Ca, Mg, K, Al, Fe, Mn, Zn, Ba) were extracted with a 0.016 M cobaltihexamine trichloride solution, following the ISO norm 23470:2007, and were measured by inductively coupled plasma optical emission spectroscopy (ICP-OES) with CCD detector (Varian, Vista MPX). The effective cation exchange capacity (CEC) was determined on the same extract by spectrophotometry at 475 nm.

Topography: The altitude of a quadrat was measured as the average altitude of the four corners. For each quadrat, the slope was the mean angular deviation from the horizontal of the four triangular planes obtained by the connexion of three of the corners, and the convexity was the mean altitude difference between a quadrat and its eight (or less, for the edge quadrats) neighbours (Harms et al. 2001). The aspect (i.e. slope orientation) was calculated using Horn's algorithm (1981). The flow direction (of water) is the direction of the higher elevation difference. Terrain ruggedness index (TRI) is the mean of the absolute differences between the elevation of a quadrat of that of the eight surrounding ones (Wilson et al. 2007), while terrain position index (TPI) is the difference the elevation of a quadrat and the mean elevation of the eight surrounding quadrats (Weiss 2001). Roughness is the difference between the maximum and the minimum elevation of the quadrat and its eight surrounding quadrats (Wilson et al. 2007).

Appendix S1 – Section 2: Environmental variable selection

First, we excluded one variable of each pair of variables displaying a Pearson correlation ≥ 0.7 (see Table S2) to avoid collinearity issues (Dormann et al. 2013). Then, a selection approach based on Information Criterion (IC) was adopted to select the final subset of environmental variables (**E**) to be used in the variation partitioning.

While stepwise selections depend on the starting point and stopping rule (e.g. forward and backward selection generally yield different final models, Grafen and Hails 2002), IC-based approaches give a weight (i.e. based on the IC) to alternative models, rank the models according to their IC-value, and use all the models (or the n best-ranked ones) for inference (Burnham and Anderson 2003). Unlike iterative procedure selections, IC-based approaches explicitly consider model-selection uncertainties and ensure that the best model (according the IC) will be considered (Burnham and Anderson 2003, Calcagno and Mazancourt 2010). As performing all possible combinations of variables may quickly become very time-prohibitive, we used the genetic algorithm developed by Calcagno and Mazancourt (2010) (package *glmulti*) to explore only a subset of all possible models, on the basis of the corrected version of Akaike information criterion (AIC_c) (details in Appendix S2: Section 1). This procedure was applied separately for all species, as this is a univariate approach. When the model converged (see Appendix S2: Section 1), the first 100 best models were saved. The procedure was repeated 20 times for each species and a consensus of the genetic algorithms was made to improve convergence. We then performed a model averaging of the first 100 best models (here, $\Delta AIC_c \approx 2$), and kept the environmental variables displaying an importance value higher than 0.8, following Calcagno and Mazancourt (2010). The sum of the variables selected with this procedure for all species constituted the final environmental matrix (**E**).

Appendix S1 – Section 3: Spatial predictor selection

We selected an optimal subset of spatial predictors to be used in the variation partitioning following Bauman et al. (2018b, 2018a) and using the functions of the package *adespatial* (Dray et al. 2018). To do so, we first generated four candidate spatial weighting matrices (SWMs) with two connectivity matrices and two weighting matrices. The connectivity matrices were a Gabriel graph (graph-based MEM) and a distance-based criterion based on the minimum distance that kept all quadrats connected (i.e. 13 m) (distance-based MEM). These matrices were either not weighted (binary SWMs) or were weighted by a function decreasing linearly with the distance (see Bauman et al. 2018b). A global test of the community composition (**Y**) against the whole set of MEM variables was performed separately for each SWM, and the *p*-value of those tests were corrected for multiple testing using the Šidak correction (Šidák 1967) to avoid type I error rate inflation issues (Bauman et al. 2018b). Since we chose to look for both positively and negatively spatially autocorrelated patterns in **Y**, the MEM variables associated to positive and negative eigenvalues were tested separately, hence doubling the total number of tests (i.e. $4 \times 2 = 8$). A forward selection with double-stopping criterion (Blanchet et al. 2008) was performed on those SWMs whose global model was significant, and the subset of spatial eigenvectors yielding the highest forward-selected adjusted canonical coefficient of determination (R^2) was selected for the variation partitioning. The optimised subset had 30 spatial predictors and corresponded to a SWM built from a Gabriel's graph with unweighted links.

Appendix S1 – Section 4: Test of the variation partitioning fractions

While the fractions of the VP resulting from actual RDAs were tested by the classical residual permutation procedure of Anderson and Legendre (1999), the global environmental fraction ([adfg] in Fig. 1) as well as the fractions resulting from the subtraction of other fractions ([d], [f], and [g] in Fig. 1) cannot be tested by this procedure. Regarding fraction [adfg], the permutation of the residuals of the global environmental model would have an inflated type I error rate due to the presence of spatial autocorrelation (SAC) in the model residuals (Dutilleul et al. 2008). The fractions obtained by subtraction of other fraction have zero degrees of freedom and can therefore not be tested by the classical procedure. Bauman et al. (2018c) have recently devised an original test for these shared fractions using constrained null models generated with torus randomisation tests or Moran spectral randomisations (MSR; Wagner and Dray 2015). Since the method was originally devised for a VP of two components, we proceeded as follows. A joint space-environment fraction (JSEF) was tested separately and jointly for the broad-scale and fine-scale MEM variables, yielding three tests: $\text{JSEF}_{\text{Broad}}$, $\text{JSEF}_{\text{Fine}}$, and $\text{JSEF}_{\text{Broad+Fine}}$. To test $\text{JSEF}_{\text{Broad}}$, we first generated the following model: RDA (Y ~ E | MEM_{F}) ([ad] on Fig. 1). The fitted portion of the model (i.e. all constrained axes) was then used together with MEM_{B} to test the $\text{JSEF}_{\text{Broad}}$ (using function *envspace.test* of the *adespatial* package; Dray et al. 2018). The test of $\text{JSEF}_{\text{Fine}}$ and $\text{JSEF}_{\text{Broad+Fine}}$ was based on the same logic.

Appendix S1 – Section 5: Spatial point pattern analysis

To evaluate whether dispersal was likely to be an ecological process underlying some of the unexplained patterns of the variation partitioning, we fitted an inhomogeneous Thomas point process (i.e. a Cox process) to the observed pattern of each species (Baddeley et al. 2015).

This point process first uses an algorithm to estimate the intensity of the point pattern (i.e. the number of individuals by unit of area) as a function of the environmental covariates (same variables used in **E** for the variation partitioning). Then, the clustering parameters of the model are fitted by the method of minimum contrast (Diggle and Gratton 1984), that is, by matching a theoretical summary statistics (here, the *K*-function, Ripley 1981) to the corresponding empirical one, computed from the observed point pattern. The clustering parameter of interest was the standard deviation of random displacement of a point from its cluster centre (i.e. σ^2). Here, we decided to consider this parameter multiplied by two as the estimated scale of a dispersal-limited cluster.

References

Anderson M. J., Legendre P. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. Journal of Statistical Computation and Simulation 62:271–303.

Baddeley A. et al. 2015. Spatial point patterns: methodology and applications with R. Chapman & Hall/CRC.

Bauman D. et al. 2018a. Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors. Ecography 41(10), 1638–1649.

Bauman D. et al. 2018b. Optimizing the choice of a spatial weighting matrix in eigenvector-based methods. Ecology: doi: 10.1002/ecy.2469.

Bauman D. et al. 2018c. Testing and interpreting the shared space-environment fraction in variation partitioning analyses of ecological data. Oikos, doi: 10.1111/oik.05496.

Blanchet F. G. et al. 2008. Forward selection of explanatory variables. Ecology 89:2623–2632.

Burnham K. P., Anderson D. R. 2003. Model selection and multimodel inference: a practical information-theoretic approach. Springer Science & Business Media.

Calcagno V., De Mazancourt C. 2010. glmulti: An R package for easy automated model selection with (Generalized) linear models. Journal of Statistical Software 34:1–29.

Diggle P. J., Gratton R. J. 1984. Monte Carlo methods of inference for implicit statistical models. Journal of the Royal Statistical Society, Series B (Methodological) 46:193–227.

Dormann C. F. et al. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36:27–46.

Dray S. et al. 2018. adespatial: Multivariate multiscale spatial analysis. R package version 0.2-0.

Dutilleul P. et al. 2008. Modifying the t test for assessing the correlation between two spatial processes 49:305–314.

Grafen A., Hails R. 2002. Modern Statistics for the Life Sciences. Oxford University Press, Oxford.

Harms K. E. et al. 2001. Habitat associations of trees and shrubs in a 50-ha neotropical forest plot. Journal of Ecology 89:947–959.

Horn B. K. P. 1981. Hill shading and the reflectance map. Proceedings of the IEEE 69:14–47.

Ripley B. D. 1981. Spatial statistics. John Wiley and Sons.

Šidák Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 62:626–633.

Wagner H. H., Dray S. 2015. Generating spatially constrained null models for irregularly spaced data using Moran spectral randomization methods. Methods in Ecology and Evolution 6:1169–1178.

Weiss A. D. 2001. Topographic Positions and Landforms Analysis (poster), ESRI International User Conference, July 2001. San Diego, CA: ESRI.

Wilson M. F. J. et al. 2007. Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope. Marine Geodesy 30:3–35.

Table S1: Description of the tree community of the Moesia north hillside (living individuals only). The species are listed by decreasing basal area. The numbers in brackets are percentages of the total abundance of basal area of the site. % quadrats is the percentage of the 143 quadrats in which the species was present. Bold species are the species that were considered in the quadrat-based and individual-based analyses.

Species	Code	Family	Abundance	Basal area (m^2/ha)	dbh (cm) (mean ± sd)	% quadrats
Quercus robur	QueRob	Fagaceae	690 (8.9%)	1578.99 (31.62%)	18.9 ± 9.9	90
Pinus sylvestris	PinSyl	Pinaceae	226 (2.9%)	1024.26 (20.51%)	28.8 ± 8.7	54
Betula pendula	BetPen	Betulaceae	349 (4.5%)	737.04 (14.76%)	18.8 ± 8.1	64
Crataegus monogyna	CraMon	Rosaceae	2408 (31.1%)	437.74 (8.77%)	4.1 ± 4.4	100
Carpinus betulus	CarBet	Betulaceae	387 (5%)	361.04 (7.23%)	10.2 ± 9.1	57
Corylus avellana	CorAve	Betulaceae	1547 (19.9%)	332.63 (6.66%)	5.1 ± 4.1	96
Quercus petraea	QuePet	Fagaceae	79 (1%)	173.1 (3.47%)	17.6 ± 11.3	33
Fagus sylvatica	FagSyl	Fagaceae	198 (2.6%)	154.38 (3.09%)	8.9 ± 8.7	55
Populus tremula	PopTre	Salicaceae	49 (0.6%)	67.91 (1.36%)	15.5 ± 5.9	12
Prunus avium	PruAvi	Rosaceae	132 (1.7%)	42.91 (0.86%)	3.9 ± 7.1	22
Salix caprea	SalCap	Salicaceae	26 (0.3%)	28.78 (0.58%)	13.8 ± 5.6	14
Acer campestre	AceCam	Sapindaceae	77 (1%)	17.43 (0.35%)	2.3 ± 6.3	28
Prunus spinosa	PruSpi	Rosaceae	708 (9.1%)	9.74 (0.2%)	1.1 ± 1.2	50
Cornus sanguinea	CorSan	Cornaceae	443 (5.7%)	9.57 (0.19%)	1.4 ± 1.5	49
Fraxinus excelsior	FraExc	Oleaceae	78 (1%)	8.13 (0.16%)	2 ± 4.1	22
Acer pseudoplatanus	AcePse	Sapindaceae	21 (0.3%)	3.01 (0.06%)	2.6 ± 4.8	8
Crataegus laevigata	CraLae	Rosaceae	9 (0.1%)	1.83 (0.04%)	5.6 ± 3.2	4
Pyrus pyraster	PyrPyr	Rosaceae	3 (0%)	1.62 (0.03%)	9 ± 6.2	1
Ligustrum vulgare	LigVul	Oleaceae	168 (2.2%)	0.97 (0.02%)	0.8 ± 0.7	24
Clematis vitalba	CleVit	Ranunculaceae	9 (0.1%)	0.61 (0.01%)	3 ± 2.3	3
Sorbus aucuparia	SorAuc	Rosaceae	9 (0.1%)	0.57 (0.01%)	2 ± 3.1	6
Malus sylvestris	MalSyl	Rosaceae	1 (0%)	0.51 (0.01%)	10.1	1
Viburnum lantana	VibLan	Adoxaceae	40 (0.5%)	0.27 (0.01%)	0.8 ± 0.8	15
Lonicera periclymenum	LonPer	Caprifoliaceae	57 (0.7%)	0.23 (0%)	0.8 ± 0.5	13
Juglans regia	JugReg	Juglandaceae	5 (0.1%)	0.21 (0%)	2.7 ± 1.2	3
Euonymus europaeus	EuoEur	Celastraceae	9 (0.1%)	0.02 (0%)	0.6 ± 0.2	2
Viburnum opulus	VibOpu	Adoxaceae	19 (0.2%)	0.02 (0%)	0.4 ± 0.1	8
Ilex aquifolium	IleAqu	Aquifoliaceae	3 (0%)	0.01 (0%)	0.8 ± 0.6	1
Sorbus torminalis	SorTor	Rosaceae	1 (0%)	0.01 (0%)	1.0	1
Ribes rubrum	RibRub	Grossulariaceae	3 (0%)	0 (0%)	0.5 ± 0.2	1
Tilia platyphyllos	TilPla	Tiliaceae	1 (0%)	0 (0%)	1.0	1
Total			7755	4993.55		

Table S2: Descriptive statistics of the soil, light, and topography parameters measured. sd: standard deviation; CV: coefficient of variation (i.e. sd/mean). The range is the distance beyond which the variable is not spatially autocorrelated anymore. It was computed through the fitting of a semivariogram model to the empirical semivariogram. See *Material and methods* for variable abbreviations.

Environmental variables		**Unit**	**Min**	**Max**	**Mean ± sd**	**CV (%)**
Soil	Stoniness index	-	0.00	2.00	1.1 ± 0.56	50.91
	Sand	%	1	41	8 ± 7	79
	Silt	%	44	86	67 ± 10	15
	Clay	%	9	45	24 ± 9	37
	col_{Red}	-	149.20	204.60	177.8 ± 9.52	5.35
	$\text{col}_{\text{Green}}$	-	129.80	177.00	153.17 ± 8.77	5.72
	col_{Blue}	-	107.80	145.40	125.58 ± 7.17	5.71
	Carbonate index	-	0.00	2.00	0.8 ± 0.58	71.54
	pH_{H2O}	-	6.14	8.31	7.5 ± 0.45	6.06
	EC	µS	33.0	197.9	105.7 ± 25.4	24.0
	OM	%	7.11	27.65	12.47 ± 3.71	29.74
	CEC	cmol_{c}/kg	14.21	32.20	23.42 ± 3.04	12.97
	NO_3^-	µg/g	6	131	42 ± 22	52
	NH_4^+	µg/g	3	24	11 ± 4	39
	N_{total}	µg/g	14	148	53 ± 23	44
	P	µg/g	0.005	0.078	0.026 ± 0.012	47.06
	Ca	µg/g	3269	6417	4779 ± 587	12
	Mg	µg/g	42	230	116 ± 31	27
	K	µg/g	45	284	105 ± 36	34
	Al	µg/g	0.34	12.33	1.68 ± 2.2	131.49
	Fe	µg/g	0.02	6.23	0.77 ± 1.08	141.38
	Mn	µg/g	0.43	24.62	2.97 ± 4.79	161.33
	Zn	µg/g	0.01	1.15	0.15 ± 0.19	123.29
	Ba	µg/g	3.5	34.9	10.9 ± 4.4	40.8
Light	CaCo	%	42	98	83 ± 10	12
	Canopy openness	-	0.13	7.46	2.04 ± 1.29	63.29
Topography	Elevation	m	161.0	193.6	180 ± 6.4	3.5
	Slope	°	9.0	52.1	24.9 ± 6.6	26.3
	Convex	-	-3.73	4.16	0.15 ± 1.55	1001.60
	Aspect	-	42	272	145 ± 36	25
	TPI	-	1.32	6.81	3.05 ± 1.03	33.91
	TRI	-	-3.52	2.21	-0.09 ± 0.67	-714.38
	Roughness	-	3.6	20.0	8.5 ± 2.7	31.6
	Flow direction	-	3	59	20 ± 11	54

Table S3: Pearson correlations of the soil, light, and topographical environmental measured in the 143 quadrats of the forest plot. See Material and methods for the full names corresponding to the acronyms.

Pearson correlation	SI	Carbo	sand	silt	clay	R	G	B	pH_H2O	EC	OM	NO3-	NH4+	Nmin	CEC	Al
Carbo	0.15	1.00														
sand	-0.16	0.56	1.00													
silt	0.29	0.05	-0.42	1.00												
clay	-0.33	-0.49	-0.21	-0.75	1.00											
R	-0.10	-0.18	-0.02	-0.13	0.20	1.00										
G	-0.05	-0.01	0.18	-0.21	0.15	0.93	1.00									
B	0.06	0.13	0.27	-0.18	0.04	0.79	0.94	1.00								
pH_H2O	0.03	0.71	0.45	0.10	-0.42	-0.28	-0.13	0.03	1.00							
EC	0.34	0.42	0.02	0.36	-0.42	-0.33	-0.22	-0.08	0.46	1.00						
OM	0.70	-0.06	-0.34	0.29	-0.17	-0.19	-0.16	-0.04	-0.08	0.35	1.00					
NO3-	0.26	0.20	0.06	0.22	-0.30	-0.35	-0.21	-0.09	0.25	0.54	0.26	1.00				
NH4+	0.44	-0.08	-0.14	0.16	-0.10	0.00	0.01	0.04	-0.16	0.21	0.57	0.22	1.00			
Nmin	0.34	0.19	0.03	0.24	-0.31	-0.33	-0.20	-0.07	0.21	0.56	0.37	0.98	0.41	1.00		
CEC	0.21	-0.06	-0.31	0.25	-0.06	-0.25	-0.26	-0.20	0.07	0.38	0.46	0.15	0.26	0.20	1.00	
Al	0.00	0.00	0.13	-0.16	0.09	0.10	0.11	0.04	-0.11	-0.14	-0.08	-0.08	-0.14	-0.10	-0.25	
Ba	-0.37	-0.50	-0.02	-0.38	0.51	0.19	0.14	0.00	-0.50	-0.50	-0.31	-0.30	-0.04	-0.32	-0.17	0.24
Ca	0.26	0.04	-0.17	0.21	-0.13	-0.27	-0.24	-0.17	0.09	0.28	0.42	0.16	0.15	0.18	0.51	0.08
K	0.33	-0.11	-0.13	0.11	-0.02	0.10	0.17	0.24	-0.12	0.14	0.40	0.21	0.32	0.26	0.37	0.09
Mg	0.14	-0.53	-0.40	-0.03	0.32	0.04	0.01	-0.02	-0.55	-0.19	0.37	0.01	0.32	0.06	0.34	0.03
Mn	0.16	-0.57	-0.53	-0.01	0.33	0.23	0.10	-0.01	-0.70	-0.31	0.22	-0.20	0.22	-0.16	0.18	0.05
P	0.10	-0.03	-0.09	0.07	-0.05	0.03	0.04	0.07	0.02	0.11	0.15	0.12	0.05	0.13	0.05	-0.22
Zn	-0.03	-0.21	-0.05	-0.14	0.21	0.17	0.13	0.07	-0.31	-0.19	0.00	-0.11	0.08	-0.08	-0.08	0.01
CaCo_Ind	0.25	0.01	0.01	0.08	-0.08	-0.05	-0.04	0.01	-0.10	0.11	0.13	-0.04	0.06	-0.04	-0.02	-0.06
Canopy Openness	-0.25	-0.02	-0.01	-0.07	0.07	0.04	0.02	-0.03	0.10	-0.12	-0.13	0.04	-0.06	0.03	0.03	0.06

Elevation	0.24	-0.08	0.23	-0.37	0.20	0.20	0.28	0.26	-0.19	-0.07	0.18	-0.09	0.36	-0.02	-0.07	0.10
Convex	0.19	0.01	-0.04	-0.07	0.06	-0.02	0.01	0.03	-0.02	-0.02	0.14	0.04	0.12	0.07	0.07	0.01
Slope	-0.14	0.03	0.06	-0.11	0.10	0.02	0.02	0.00	-0.01	-0.09	-0.22	0.01	-0.17	-0.03	-0.10	0.13
Aspect	-0.03	-0.01	-0.09	-0.01	0.03	0.00	0.00	0.02	0.06	-0.04	-0.04	0.05	-0.08	0.04	0.04	-0.07
TPI	-0.28	-0.01	0.16	-0.30	0.26	0.09	0.12	0.07	-0.07	-0.22	-0.30	-0.11	-0.25	-0.16	-0.12	0.17
TRI	0.00	-0.03	0.02	0.06	-0.07	0.08	0.07	0.05	-0.03	-0.09	0.01	-0.02	0.00	-0.02	-0.13	0.02
Roughness	-0.23	-0.01	0.13	-0.25	0.21	0.09	0.09	0.02	-0.08	-0.26	-0.32	-0.09	-0.26	-0.14	-0.19	0.23
Flow_direction	-0.29	0.04	0.26	-0.08	-0.03	0.03	0.08	0.05	0.00	-0.14	-0.37	0.04	-0.09	0.02	-0.23	0.12

Pearson correlation	Ba	Ca	K	Mg	Mn	P	Zn	CaCo_Ind	Canopy Openness	Elevation	Convex	Slope	Aspect	TPI	TRI	Roughness
Ba	1.00															
Ca	0.04	1.00														
K	0.15	0.45	1.00													
Mg	0.47	0.47	0.56	1.00												
Mn	0.47	0.16	0.35	0.70	1.00											
P	-0.08	0.14	0.09	0.04	0.18	1.00										
Zn	0.18	-0.20	0.04	0.21	0.28	0.18	1.00									
CaCo_Ind	0.00	0.01	0.05	0.04	0.03	-0.03	-0.01	1.00								
Canopy Openness	0.01	-0.02	-0.05	-0.03	-0.02	0.02	0.01	-0.99	1.00							
Altit	0.28	0.05	0.20	0.21	0.14	0.00	0.14	0.07	-0.07	1.00						
Convex	0.06	0.12	0.19	0.11	0.08	-0.03	-0.09	0.09	-0.08	0.47	1.00					
Slope	0.15	-0.02	-0.03	0.04	0.00	-0.06	0.01	0.06	-0.06	-0.15	-0.03	1.00				
Aspect	-0.09	-0.05	-0.01	-0.09	0.04	0.26	0.09	-0.02	0.02	-0.21	-0.08	0.06	1.00			
TPI	0.30	0.01	-0.08	0.10	-0.01	-0.16	0.01	-0.01	0.00	-0.03	-0.04	0.70	0.03	1.00		
TRI	0.04	0.07	-0.04	0.04	-0.03	-0.11	0.05	-0.01	0.00	0.04	-0.10	0.07	0.22	0.15	1.00	
Roughness	0.27	-0.06	-0.18	0.03	0.01	-0.11	0.04	-0.01	0.00	-0.04	0.01	0.72	0.10	0.89	0.15	1.00
Flow_direction	0.18	-0.16	-0.12	-0.11	-0.13	-0.16	0.02	-0.01	0.01	-0.01	0.05	0.13	-0.18	0.24	-0.12	0.22

Table S4: Adjusted clustering parameters of the inhomogeneous Thomas point process for all analysed species. κ is the intensity of clusters (i.e. number of individuals per m^2), while σ^2 is the standard deviation of random displacement of a point from its cluster centre. The pattern spatial scale is defined as the value of σ^2 multiplied by two.

Species	**κ**	**σ^2 (m)**	**Pattern scale (m)**
Prunus spinosa	0.306045371	0.1	0.3
Ligustrum vulgare	0.003255101	1.0	2.0
Cornus sanguinea	0.001319327	1.7	3.3
Carpinus betulus	0.006131472	1.9	3.9
Crataegus monogyna	0.012521506	2.8	5.5
Fagus sylvatica	0.004241511	2.9	5.8
Quercus robur	0.016263451	3.7	7.3
Corylus avellana	0.009804125	4.0	8.0
Prunus avium	1.64E-07	4.1	8.3
Betula pendula	0.000523843	4.2	8.4
Fraxinus excelsior	0.000580204	4.5	9.1
Quercus petraea	0.000650333	8.7	17.4
Viburnum lantana	15911.25491	59.0	118.0
Populus tremula	18515.98488	59.5	119.0
Acer campestre	23710.78053	68.7	137.4
Pinus sylvestris	9.51E-06	105.7	211.4

Table S5: Spatial association index (A_{ij}) of the species pairs that fell out of the global simulation enveloppes within neighbourhood of 3 and 10 m. The green and orange cells are the values below -0.3 and beyond 0.3, respectively. The full names corresponding to the species acronyms are provided in Table S1.

d = 3 m	AceCam	BetPen	CarBet	CorAve	CorSan	CraMon	FagSyl	FraExc	LigVul	PinSyl	PopTre	PruAvi	PruSpi	QuePet	QueRob	VibLan
AceCam									1.11				0.86			
BetPen			-0.26			0.11										
CarBet		-0.25		-0.12	-0.28	-0.13	-0.16		-0.39	-0.24		-0.23	-0.31		0.19	
CorAve			-0.32		0.15	-0.58	-0.39	-0.24		0.89			-0.57		-0.53	
CorSan		0.77	-0.13	0.56		0.46	0.19			0.15	-0.96	0.81	-0.23	-0.69	-0.96	
CraMon		0.36	-0.45	-0.29	0.20								0.39		0.18	
FagSyl												0.65				
FraExc													0.86			
LigVul			-0.32			0.19										
PinSyl								1.13							-0.12	
PopTre																
PruAvi																
PruSpi		-0.35	-0.68	-0.58	-0.46	0.16	-0.98	0.84	0.12	0.17		0.84		-0.22	-0.28	
QuePet																
QueRob		0.23		0.25	0.79					-0.29			0.47			
VibLan																

d = 10 m	AceCam	BetPen	CarBet	CorAve	CorSan	CraMon	FagSyl	FraExc	LigVul	PinSyl	PopTre	PruAvi	PruSpi	QuePet	QueRob	VibLan
AceCam			0.86	0.94	0.43	0.18	0.36	0.52	0.53			0.12	0.43		0.39	
BetPen			-0.12	0.99	-0.53	-0.66	0.16	0.19		-0.23	0.29	-0.28	-0.15	0.16	0.57	-0.35
CarBet	-0.40	-0.53		-0.80	-0.13	-0.77	-0.84	-0.16	-0.14	-0.15		-0.11	-0.22	-0.12	0.59	
CorAve	0.26	0.21	0.15		0.22	0.13		0.65	0.22			0.21	-0.54			0.19
CorSan	0.30	0.23	-0.22	0.52		0.19	0.14	0.25	0.36	0.18	-0.45	0.22	-0.12	-0.74	-0.11	0.35
CraMon	0.23	0.15	0.12	0.15	0.23		0.22	0.28	0.20	0.26	0.73	0.24	0.25	0.24		0.28
FagSyl	0.23	0.12	-0.23	-0.19	0.33	0.15		0.49		0.13		0.27	0.23	0.29	0.46	
FraExc	0.39	-0.14	-0.16	-0.27	0.89	0.12	0.24		-0.85	0.38	-0.55	-0.67	0.35	-0.27	-0.15	
LigVul	0.94	-0.20	-0.35	-0.19	0.22	-0.18	-0.32	-0.23		-0.17	-0.69	0.12	0.46	-0.40	-0.25	0.35
PinSyl	-0.15	-0.12	-0.25	-0.19	0.18	0.19	0.18	0.41			-0.52	0.36			-0.20	0.27
PopTre	-0.16	0.52	-0.16	-0.18	-0.38	-0.24	-0.17	-0.48	-0.56	-0.39		-0.45	-1.20		-0.19	
PruAvi	0.41	-0.41	-0.24	-0.19	-0.74	-0.12	-0.16	-0.83	0.17	-0.95	-0.63		-0.15	-0.22	-0.27	0.99
PruSpi	0.13	-0.33	-0.35	-0.29	0.24	0.60	0.11	0.39	0.15	0.62	-1.37	0.16		-0.14	-0.12	
QuePet		0.27					0.33		-0.27							
QueRob			0.21							0.91			0.29			
VibLan					0.49	0.25			0.69	0.34		0.47				

VII. Discussion générale : Supporting information

Appendix S1: Comparison of detrending *versus* not detrending data before performing MEM analysis

Objective of the document

This R code simulates spatially structured response variables (*y*) consisting of the sum of one or two trends, one finer spatial pattern (i.e. one MEM variable), and a random noise. Then, the statistical power and spatial R^2 estimation accuracy of two procedures are compared: (1) detrending the response variable prior to using MEM, and (2) using MEM directly without first detrending the response variable.

The degree of spatial autocorrelation of *y* varies between simulations to mimic strongly and weekly spatially autocorrelated species. In addition to the varying trend(s) simulated, the identity of the MEM variables used to structure *y* also varies to add a spatial structure either structured at broad or at intermediate to fine spatial scale.

Simulations

```
library(adespatial)
library(spdep)
```

We build a grid of 50 x 25 cells from which we generate positively autocorrelated MEM variables with a graph-based spatial weighting matrix (SWM) consisting of an unweighted Gabriel graph.

```
grid <- expand.grid(x = seq(1:50), y = seq(1:25))
w_grid <- listw.candidates(grid, nb = "gab", weights = "binary")
MEM <- scores.listw(w_grid[[1]], MEM.autocor = "positive")
```

The simulations will be conducted on a sample of 153 cells to mimic a realistic sampling effort. The latter will be either regularly or randomly distributed:

```
# Regular:
tx <- seq(from = 1, to = 50, by = 3)
ty <- seq(from = 1, to = 25, by = 3)
grid_sub  <- grid[grid[, 1] %in% tx, ]
grid_sub  <- grid_sub[grid_sub[, 2] %in% ty, ]
subreg <- as.vector(sapply(rownames(grid_sub), function (x) match(x,
                                                                 rownames(grid))))

# Random:
set.seed(1)
subran <- sample(c(1:1250), 123, replace = FALSE)
```

Number of simulations:

```
nsim <- 1000
```

Result matrix to gather the simulation results (power, spatial R^2 estimation accuracy):

```
results.reg <- data.frame(detrended_MEM = rep(NA, nsim),
                          undetrended_MEM = rep(NA, nsim),
                          detrended_PCNM = rep(NA, nsim),
                          undetrended_PCNM = rep(NA, nsim))
results.ran <- data.frame(detrended_MEM = rep(NA, nsim),
                          undetrended_MEM = rep(NA, nsim),
                          detrended_PCNM = rep(NA, nsim),
                          undetrended_PCNM = rep(NA, nsim))
```

Vector to gather the simulated spatial R^2 of reference (i.e. the real explanatory power of the spatial structure to explain the response variable):

```
R2.ref.all_reg <- c()
R2.ref.all_ran <- c()
```

The spatial eigenvectors will be either a selection of MEM variables resulting from the optimization procedure of Bauman et al. (2018b) used with the criterion of the forward selection with double stopping criterion (Blanchet et al. 2008), or a subset of PCNM variables obtained by forward selection (Blanchet et al. 2008).

Here, the optimisation procedure is run on the basis of five candidate SWMs, using the forward selection with double stopping criterion (Blanchet et. al. 2008) (see Bauman et al. 2018a, b for a review on the selection of spatial eigenvectors in eigenvector-based methods, and for a discussion on the selection of a spatial weighting matrix, respectively). The five candidate SWMs are two graph-based SWM (Delaunay triangulation, Gabriel graph) unweighted or weighted by a function decreasing with the distance following a concave-down curve (exponent = 5, see help(listw.candidates)), and a distance-based SWM.

Begining of the simulations:

```
for (h in 1:2) {   # For the regular and random sampling designs
  results <- results.ran
  R2.ref.all <- c()

  if (h == 1) sub <- subreg else sub <- subran

  grid.sub <- grid[sub, ]

  # Here, we already generate the candidate spatial weighting matrices (SWM)
  # that will be used for the optimization. We also generate the PCNM variables.
  candid <- listw.candidates(grid.sub, nb = c("del", "gab", "pcnm"),
                             weights = c("binary", "fdown"))
  # PCNM:
  pcnm <- PCNM(dist(grid.sub))

  for (i in 1:nsim) {
    set.seed(i)
    nb.trend <- sample(c(1, 2), 1)          # Number of trends: either 1 or 2
    samp.trend <- sample(c(1, 2), nb.trend) # Selected trends: x, y, or x and y

    # If only one MEM variable is used to generate the "complex spatial pattern"
    # component of y:
    samp.fine <- sample(seq(5, 50, 1), 1)   # One randomly-sampled MEM for y
    if (nb.trend == 1) {
      mem <- scale(grid[, samp.trend]) + scale(MEM[, samp.fine])
    } else mem <- scale(grid[, 1]) + scale(grid[, 2]) + scale(MEM[, samp.fine])

    # Response variable on the complete grid:
    alpha <- sample(seq(0.4, 0.5, 0.01), 1)
    y <- (alpha * scale(mem)) + ((1-alpha) * scale(rnorm(length(mem))))
```

```
# Subsampling in the 153 cells:
y.sub <- y[sub]
mem.sub <- mem[sub]

# Real R² (reference):
R2.ref <- cor(y.sub, mem.sub)^2
R2.ref.all <- c(R2.ref.all, R2.ref)

# Detrending:
# ***********
mod.trend <- lm(y.sub ~., as.data.frame(grid.sub))
trend <- which(summary(mod.trend)$coefficients[c(2, 3), 4] <= 0.05)
if (length(trend) != 0) {
  det <- lm(y.sub ~ ., as.data.frame(grid.sub[, trend]))
  # The trend(s) is/are removed:
  y.det <- residuals(det)
  # Adjusted R² of the trend:
  R2.trend <- summary(det)$adj.r.squared
  # Optimization of the SWM and of a best subset of MEM variables from the
  # fivecandidate SWMs generated above with 'listw.candidates':
  # (More details in help(listw.candidates))

  modsel.det <- listw.select(y.det, candid, method = "FWD",
                             MEM.autocor = "positive")

  # If a significant spatial structure is detected in y_sub, then the
  # difference between the R² of reference and the sum of the R² of the
  # trend and of the MEM variables is computed:
  if (any(modsel.det$candidates$Pvalue <= 0.05)) {
   if (length(modsel.det) > 1) {
    R2.det <- RsquareAdj(rda(y.sub, modsel.det$best$MEM.select))$adj.r.squared
    R2.final <- R2.trend + R2.det
    delta_det <- R2.final - R2.ref
    results$detrended_MEM[i] <- delta_det
   }
  }
  # PCNM:
  p <- anova(rda(y.det, pcnm$vectors))$Pr[1]
  if (p <= 0.05) {
    R2.det <- RsquareAdj(rda(y.det, pcnm$vectors))$adj.r.squared
    class <- class(try(fwd <- forward.sel(y.det, pcnm$vectors,
                                          adjR2thresh = R2.det), TRUE))
    if (class != "try-error") {
      modsel.det <- pcnm$vectors[, sort(fwd$order)]
      R2.det <- RsquareAdj(rda(y.sub, modsel.det))$adj.r.squared
      R2.final <- R2.trend + R2.det
      delta_det <- R2.final - R2.ref
      results$detrended_PCNM[i] <- delta_det
    }
  }
}


# No detrending:
# **************
# We optimize the selection of a subset of MEM variables among the five SWM
# candidates directly, without first detrending y.
```

```
    modsel.undet <- listw.select(y.sub, candid, method = "FWD",
                                 MEM.autocor = "positive")

    if (any(modsel.undet$candidates$Pvalue <= 0.05)) {
      if (length(modsel.undet) > 1) {
        R2.undet <- modsel.undet$candidates$R2Adj.select[modsel.undet$best.id]
        delta_det <- R2.undet - R2.ref
        results$undetrended_MEM[i] <- delta_det
      }
    }
    # PCNM:
    p <- anova(rda(y.sub, pcnm$vectors))$Pr[1]
    if (p <= 0.05) {
      R2.undet <- RsquareAdj(rda(y.sub, pcnm$vectors))$adj.r.squared
      class <- class(try(fwd <- forward.sel(y.sub, pcnm$vectors,
                                            adjR2thresh = R2.undet), TRUE))
      if (class != "try-error") {
        modsel.undet <- pcnm$vectors[, sort(fwd$order)]
        R2.undet <- RsquareAdj(rda(y.sub, modsel.undet))$adj.r.squared
        delta_det <- R2.undet - R2.ref
        results$undetrended_PCNM[i] <- delta_det
      }
    }
  }
  if (h == 1) {
    R2.ref.all_reg <- R2.ref.all
    results.reg <- results
  } else {
    R2.ref.all_ran <- R2.ref.all
    results.ran <- results
  }
}
```

Calculation of the statistical power and R² estimation accuracy:

```
# Regular sampling design:
# ************************
# Delta_R2 (accuracy):
accuracy.det_MEM.mean.reg <- mean(na.omit(results.reg$detrended_MEM))
accuracy.det_MEM.sd.reg <- sd(na.omit(results.reg$detrended_MEM))
accuracy.undet_MEM.mean.reg <- mean(na.omit(results.reg$undetrended_MEM))
accuracy.undet_MEM.sd.reg <- sd(na.omit(results.reg$undetrended_MEM))

accuracy.det_PCNM.mean.reg <- mean(na.omit(results.reg$detrended_PCNM))
accuracy.det_PCNM.sd.reg <- sd(na.omit(results.reg$detrended_PCNM))
accuracy.undet_PCNM.mean.reg <- mean(na.omit(results.reg$undetrended_PCNM))
accuracy.undet_PCNM.sd.reg <- sd(na.omit(results.reg$undetrended_PCNM))

# Power:
power.det_MEM.reg <- length(which(results.reg$detrended_MEM != "NA")) / nsim
power.undet_MEM.reg <- length(which(results.reg$undetrended_MEM != "NA")) / nsim
power.det_PCNM.reg <- length(which(results.reg$detrended_PCNM != "NA")) / nsim
power.undet_PCNM.reg <- length(which(results.reg$undetrended_PCNM != "NA")) / nsim

# Random sampling design:
# ***********************
# Delta_R2 (accuracy):
accuracy.det_MEM.mean.ran <- mean(na.omit(results.ran$detrended_MEM))
accuracy.det_MEM.sd.ran <- sd(na.omit(results.ran$detrended_MEM))
accuracy.undet_MEM.mean.ran <- mean(na.omit(results.ran$undetrended_MEM))
accuracy.undet_MEM.sd.ran <- sd(na.omit(results.ran$undetrended_MEM))
accuracy.det_PCNM.mean.ran <- mean(na.omit(results.ran$detrended_PCNM))
accuracy.det_PCNM.sd.ran <- sd(na.omit(results.ran$detrended_PCNM))
accuracy.undet_PCNM.mean.ran <- mean(na.omit(results.ran$undetrended_PCNM))
accuracy.undet_PCNM.sd.ran <- sd(na.omit(results.ran$undetrended_PCNM))

# Power:
power.det_MEM.ran <- length(which(results.ran$detrended_MEM != "NA")) / nsim
power.undet_MEM.ran <- length(which(results.ran$undetrended_MEM != "NA")) / nsim
power.det_PCNM.ran <- length(which(results.ran$detrended_PCNM != "NA")) / nsim
power.undet_PCNM.ran <- length(which(results.ran$undetrended_PCNM != "NA")) / nsim
```

Results of the 1000 simulations with and without detrending:

```
##                           Power    Accuracy_mean  Accuracy_sd
## det. + MEM - regular      0.421         0.024        0.064
## MEM - regular             0.973        -0.005        0.062
## det. + MEM - random       0.379         0.029        0.085
## MEM - random              0.917        -0.025        0.073
## det. + PCNM - regular     0.305         0.068        0.068
## PCNM - regular            0.815         0.010        0.010
## det. + PCNM - random      0.271         0.054        0.094
## PCNM - random             0.736        -0.042        0.091
```

Negative and positive delta_R^2 indicate underestimation and overestimation, respectively.

Graphical outputs

Here are a few graphical outputs of four examples from the simulations:

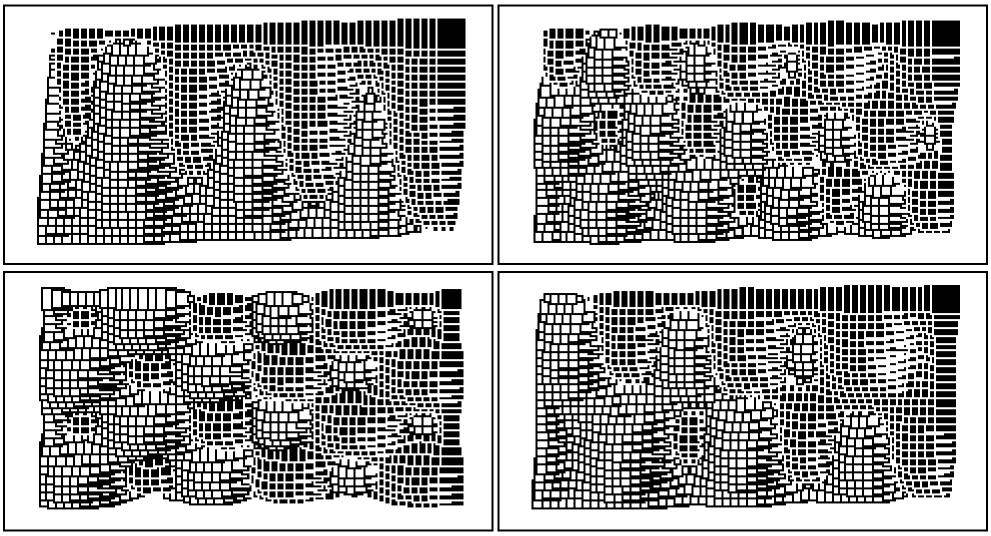


Figure 2: Examples of spatial patterns resulting from one or two trends and one MEM of broad of intermediate scale, on the complete grid

These maps highlight spatial patterns resulting from trends and MEM. They were used with a random noise to generate response variables *y* (Fig. 3).

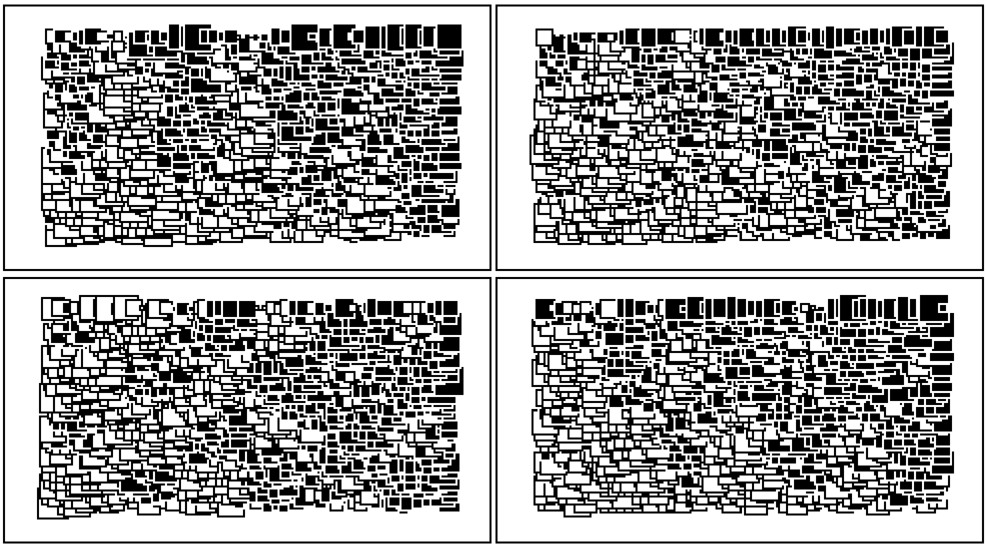


Figure 3: Examples of response variables resulting from the linear combination of the spatial maps of Figure 2 and a random noise. The big black and white squares can be viewed as cells displaying high and low abundance of the considered species, respectively.

The reference spatial patterns of *y* after sampling the complete grid in 153 regularly-distributed cells are displayed in Fig. 4.

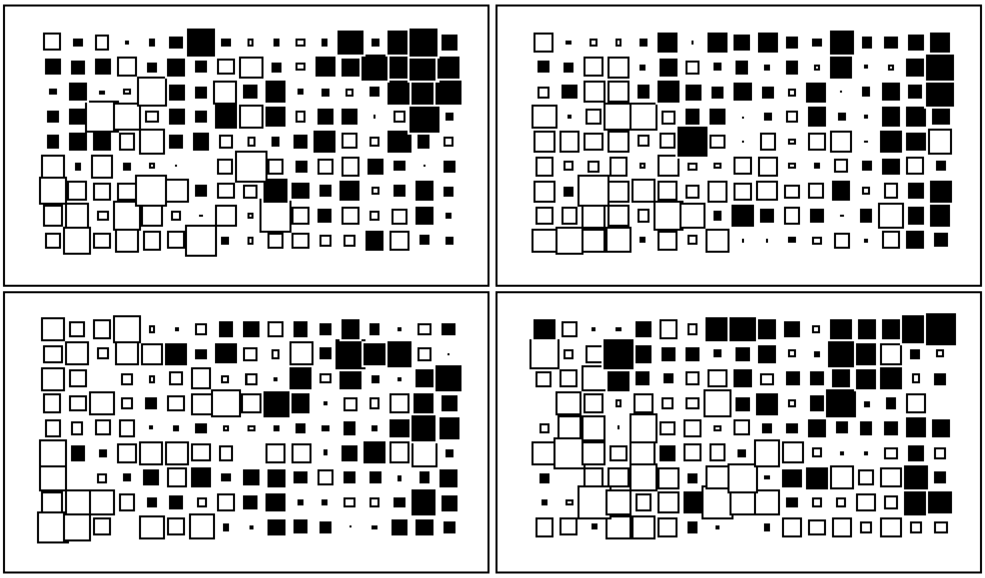


Figure 4: Spatial patterns of Figure 3 after sampling of the 153 cells. These are the patterns from which the reference R² were calcultated, and that the detrending + MEM and MEM only procedures aimed to retrieve

References

Bauman D. et al. 2018a. Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors. Ecography, 41(10), 1638–1649.

Bauman D. et al. 2018b. Optimizing the choice of a spatial weighting matrix in eigenvector-based methods. Ecology: doi: 10.1002/ecy.2469.

Blanchet G. et al. 2008. Forward selection of explanatory variables. Ecology, 89(9), 2623–2632

Borcard D. et al. 2004. Dissecting the spatial structure of ecological data at multiple scales. Ecology, 85, 1826–1832

Borcard D. et al. 2018. Numerical ecology with R, second edition. Use R! Series, Springer International Publishing, Cham, Switzerland.

Legendre P. Legendre L. 2012. Numerical ecology. Elsevier, Amsterdam

Publications et communications

Publications

Fayolle A, Swayne MD, Aleman J, Azihou AF, **Bauman D**, te Beest M et al. A sharp floristic discontinuity revealed by the biogeographic regionalization of African savannas. **Journal of Biogeography** – in press.

Bauman D*, Vleminckx J*, Hardy O, Drouet T. 2018. Testing and interpreting the shared space-environment fraction in variation partitioning analyses of ecological data. **Oikos**: doi: 10.1111/oik.05496.

Cuma FM, **Bauman D**, Mujinya BB, Mleci Y, Kalenga M, Shutcha MN et al. 2018. Edaphic specialisation in relation to termite mounds in Katanga (DR. Congo): a reciprocal transplant experiment with congeneric tree species. **Journal of Vegetation Science**: doi: 10.1111/jvs.12675.

Bauman D, Drouet T, Fortin M-J, Dray S. 2018. Optimizing the choice of a spatial weighting matrix in eigenvector-based methods. **Ecology**: doi: 10.1002/ecy.2469.

Davila F, Botteaux A, **Bauman D**, Cherasse S, Aron S. 2018. Antibacterial activity of male and female sperm-storage organs in ants. **Journal of Experimental Biology** : doi: 10.1242/jeb.175158.

Bauman D, Drouet T, Dray S, Vleminckx J. 2018. Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors. **Ecography** 41(10): 1638–1649.

Kuhn A, **Bauman D**, Darras H, Aron S. 2017. Sex-biased dispersal creates spatial genetic structure in a parthenogenetic ant with a dependent-lineage reproductive system. **Heredity** 119(4): 207-213.

Muledi JI*, **Bauman D***, Drouet T, Vleminckx J, Jacobs A, Meerts P et al. 2017. Fine-scale habitats influence tree species assemblage in a miombo forest. **Journal of Plant Ecology** 10(6): 958-969.

Vleminckx J, Doucet J-L, Morin-Rivat J, Biwolé AB, **Bauman D**, Hardy O et al. 2017. The influence of spatially structured soil properties on tree community assemblages at a landscape scale in the tropical forests of southern Cameroon. **Journal of Ecology** 105(2): 354-356.

Bauman D, Raspé O, Meerts P, Degreef J, Muledi JI, Drouet T. 2016. Multiscale assemblage of an ectomycorrhizal fungal community: the influence of host functional traits and soil properties in a 10-ha miombo forest. **FEMS Microbiology Ecology** 92(10) fiw151.

* *Equally contributing authors*

Communications orales

Bauman D, Drouet T. 2017. Habitat filtering, dispersal limitation and local tree species coexistence: a matter of scale. **54th Annual Meeting of the Association of Tropical Biology and Conservation** – Merida, Mexico.

Bauman D, Drouet T. 2017. Combining Moran's eigenvector maps (MEM) to spatial point pattern analysis: a complementary approach for highlighting spatial patterns and processes in communities. **Young Modellers in Ecology Meeting 2017** – Buchenbach, Germany.

Bauman D, Drouet T. 2017. Improving the detection of dispersal limitation in plant communities. **European Conference of Tropical Ecology** – Brussels, Belgium.

Bauman D, Raspé O, Degreef J, Meerts P, Muledi JI, Drouet T. 2014. Disentangling deterministic and neutral processes driving species assemblages in fungal and vegetal ectomycorrhizal communities in a dry woodland (DRC). **Second Annual Meeting on Plant Ecology and Evolution** – Louvain-La-Neuve, Belgium.

La compréhension des mécanismes écologiques qui sous-tendent l'assemblage des communautés végétales et la coexistence des espèces est un objectif central en écologie. Ces mécanismes sont nombreux et contrastés. La composition d'une communauté de plantes dépend ainsi de mécanismes déterministes liés aux conditions environnementales abiotiques (climat, conditions physiques et chimiques du sol, lumière) et d'interactions biotiques complexes, positives (facilitation, symbioses) comme négatives (compétition, prédation, pathogènes). En outre, les communautés sont influencées par des effets stochastiques (capacité de dispersion limitée, dérive écologique). Ces mécanismes, aussi différents soient-ils, ont néanmoins en commun la génération de motifs (*patterns*) spatiaux de distribution des espèces dans les communautés. L'analyse détaillée de ces structures spatiales est ainsi une voie privilégiée vers la compréhension des mécanismes qui les ont générées.

Les méthodes basées sur des vecteurs propres spatiaux, telles que les *Moran's eigenvector maps* (MEM), sont aujourd'hui parmi les plus puissantes pour détecter les patterns complexes et multi-échelles propres aux distributions des espèces. Ces vecteurs propres, combinés à un ensemble de variables environnementales dans un cadre de partition de variation, permettent de démêler les effets uniques et l'effet conjoint des variables environnementales et spatiales sur la variation de composition d'une communauté. Cette partition mène ainsi à une quantification de l'action des mécanismes déterministes et des effets stochastiques sur la communauté.

Cette thèse présente des solutions puissantes et précises à plusieurs importantes limitations méthodologiques compromettant la détection de patterns spatiaux de distributions d'espèces et l'inférence des mécanismes écologiques sous-jacents. Ces apports méthodologiques sont également illustrés au travers de trois cas d'études fournis par deux communautés d'arbres – tropicale et tempérée – et une communauté de champignons symbiotiques des arbres.