Rolling the Dice or Struggling for Survival

Cheating in life's casino

Edwin T. Pos

Thesis committee:

Prof. dr. Simon A. Levin
Prof. dr. Paulien Hogeweg
Prof. dr. Max Rietkerk
Prof. dr. Rampal Etienne
Prof. dr. Pieter Zuidema

Cover design
Pixabay, CCO Creative Commons, No attribution required

Photos
Opening Chapter 1: Public Domain
Opening Chapter 2: Pixabay, CCO Creative Commons, No attribution required
Opening Chapter 3: Public Domain
Opening Chapter 4: CC by 2.5, Jim Harrison
Opening Chapter 5: Public Domain (domínio público)
Opening Chapter 6: Photo courtesy of USDA ARS
Opening Chapter 7: Tericee (Flickr) CC BY-NC-SA 2.0
Opening Chapter 8: Hans ter Steege, personal permission
Opening References: Pixabay, CCO Creative Commons, No attribution required
Opening English Summary - Curriculum Vitae: E.T. Pos personal photos

Published by Utrecht University

Printed by Ipskamp Printing on FSC-certified paper

Note that all R-scripts can be found at: https://github.com/EdwinTPos/thesis/SupplementalMaterial

Printed October 2018

ISBN: 978-90-393-7068-1

Rolling the Dice or Struggling for Survival

Cheating in life's casino

Dobbelen of worstelen om te overleven
Valsspelen in het casino van het leven
(met een samenvatting in het Nederlands)

Proefschrift
ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties

in het openbaar te verdedigen op woensdag 12 december 2018
des middags te 4.15 uur

door
Edwin Theodoor Pos
geboren op 29 januari 1987
te Nunspeet

Promotoren: Prof. dr. H. ter Steege
Prof. dr. J. Ghazoul

Out of the night that covers me,
Black as the Pit from pole to pole,
I thank whatever gods may be
For my unconquerable soul.

In the fell clutch of circumstance
I have not winced nor cried aloud.
Under the bludgeoning's of chance
My head is bloody, but unbowed.

Beyond this place of wrath and tears
Looms but the Horror of the shade,
And yet the menace of the years
Finds, and shall find, me unafraid.

It matters not how strait the gate,
How charged with punishments the scroll.
I am the master of my fate:
I am the captain of my soul.

William Ernest Henley (1888)

Contents

If I have seen further, it is by standing on the shoulders of giants

Isaac Newton (1643 - 1727)

No effort of mine could avail to make the book easy reading.

Ronald A. Fisher (1890 - 1962)

Chapter One

Unravelling betadiversity of Amazonian tropical trees: rolling the dice or struggling for survival

"...Who can explain why one species ranges widely and is very numerous, and why another allied species has a narrow range and is rare? These relations are of the highest importance, for they determine the present welfare, and, as I believe, the future success and modification of every inhabitant of this world..."

In his seminal work On the Origin of Species (1), Charles Darwin asked one of the biggest and long standing questions in the field of evolutionary ecology: why are some species common and range widely while many others are rare and restricted in their distribution? We see such a pattern in almost every form of life, from trees to birds, butterflies, mammals (2), vascular epiphytes (3), frogs, bats, dung beetles (4), even bacteria (5, 6) or fungi (7) and many other groups (8). This pattern of community composition in terms of commonness and rarity can be summarized by what has been dubbed the second law of Biodiversity, stating that the "common species are rare and rare species are common" (9). But how did this seemingly universal pattern of diversity came to be and how is it is maintained? What mechanisms account for the rarity of some and the commonness of others? This question remains important as it allows for a better understanding of the driving forces of diversity, in the past, present and future. Arguably the need for understanding what determines species diversity is even more critical than ever before in light of the last decades, where we have seen an astonishing decline in species diversity proceeding at an unprecedented rate in history (10). In fact, this question was considered among the 25 most important scientific questions to be answered in the coming years (11, 12) and still is among the top 100 questions to be answered in ecology (13), even after more than 150 years since Charles Darwin first asked it.

Patterns as described above can also be found in Amazonian rainforests, where individual trees are distributed in a similar way over an incredible amount of species, even so much suggested as 16.000 (14). The Amazon, covering approximately 5.7 million square kilometers of the Earth's surface, harbors an incredible amount of biodiversity and is considered to be among the most important properties of our world. They are in a continuous dynamic state – individuals come and go. Trees fall down and their gaps are recolonized by their successors, ever changing the structure and composition of the forest. And although potential mechanisms to explain patterns in diversity at various geographical and temporal scales, such as environmental filtering or stochastic events, have been proposed over the years, no true consensus has been reached concerning the origin and maintenance of these patterns. Typically, changes in composition are studied using measurements in variation among (sampled) local communities, determining how they contribute to the regional diversity of an area and vice versa (15). For a long time it was assumed that variation in species composition was mainly the result of so called niche-assembly rules, well known from classic ecological models (16–28). According to these rules, species composition of communities is mainly the result of environmental heterogeneity and the subsequent selective pressures resulting in filtering species from the regional species pool (28). Whether species are present or absent in a community is not primarily the result of the inability to reach a site (i.e. dispersal limitation) but is more dependent on the ecological demands of each species (see also paragraph 2.1 from chapter 2). In contrast, a more neutral (and at first radical) view for explaining variation in species composition was proposed by Hubbell at the beginning of the twenty-first century in his 'Unified Neutral Theory of Biodiversity and Biogeography' (UNTB) (29), which created quite a stir in the scientific world. For readers unfamiliar with this concept, chapter two provides a light primer on neutral theory. With a simplified, yet elegant model, the UNTB was able to predict species abundance distributions, species area curves and even phylogenies. This neutral view on biological processes thus generated both interest and controversy in the field of community ecology (30–38). It has already been argued that a mechanism to maintain exact equal fitness between individuals (true neutrality) over long periods is very difficult to envision (39). To be fair, this is not what Hubbell intended when publishing his ideas. Most biologists realize the world certainly is not solely neutral and obviously driven by selection to some extent. One of the greatest biologist of all, quoted at the start of this dissertation, even wrote an entire book on the subject. However, one cannot ignore the success the UNTB has had in explaining or predicting observed patterns. The question at hand then is not how or if we can prove or defy either theory by fitting it to more and larger empirical datasets but to look more closely at both theories and to study what their relative importance is at different scales. For instance, at what temporal or spatial scales does the importance of stochastic processes in relation to more deterministic ones shift? Is the first always present as background noise, continuously acting against competition, selection and predation by "*rolling the dice of life*" (40) in concert with selection? Or is stochastic interplay more important in the long run with taxa playing this same game of life, but now using *"loaded dice"* (40) and hence cheating the otherwise stochastic game? Can we understand community dynamics using neutral theory or is het necessary to invoke more complex processes of community assembly? In other words, what is the relative importance of both neutral and niche processes and how can we disentangle these two components in the search for understanding betadiversity?

Using recent developments combined with earlier work from population genetics, constructing a novel spatially semi-explicit plot-based neutral model in combination with a large-scale dataset from the Amazonian Tree Diversity Network (ATDN) (41) and by using principles from information theory and statistical mechanics I aim at answering such questions. The goal of this thesis was to unravel Amazonian betadiversity, to create an understanding of community dynamics and to further the field of evolutionary ecology. The chapters that follow are the structure of the thesis; scaling up from detailed studies of empirical datasets to the comparisons of fundamental estimation of patterns and processes and finally to the use of a newly developed semi-spatially explicit neutral model and the mathematical machinery of information theory to provide this understanding of community dynamics:

Chapter one gives a general introduction into the subject and provides a thesis outline.

Chapter two is a light primer of neutral theory and maximum entropy, to provide readers with the necessary background for the following chapters and to set the stage between classical ecology and neutral theory. Those who are familiar with the concept of neutral theory and maximum entropy may skip this chapter.

Chapter three investigates the consequence of omitting unidentified records from empirical datasets. As ecologists are often unable to identify all field collections to a species, such indets are often not taken into account to save costly and time-consuming efforts of identifying them. The effects of this practice, however, remained fairly understudied and here we focus on its consequences on large-scale patterns in biodiversity.

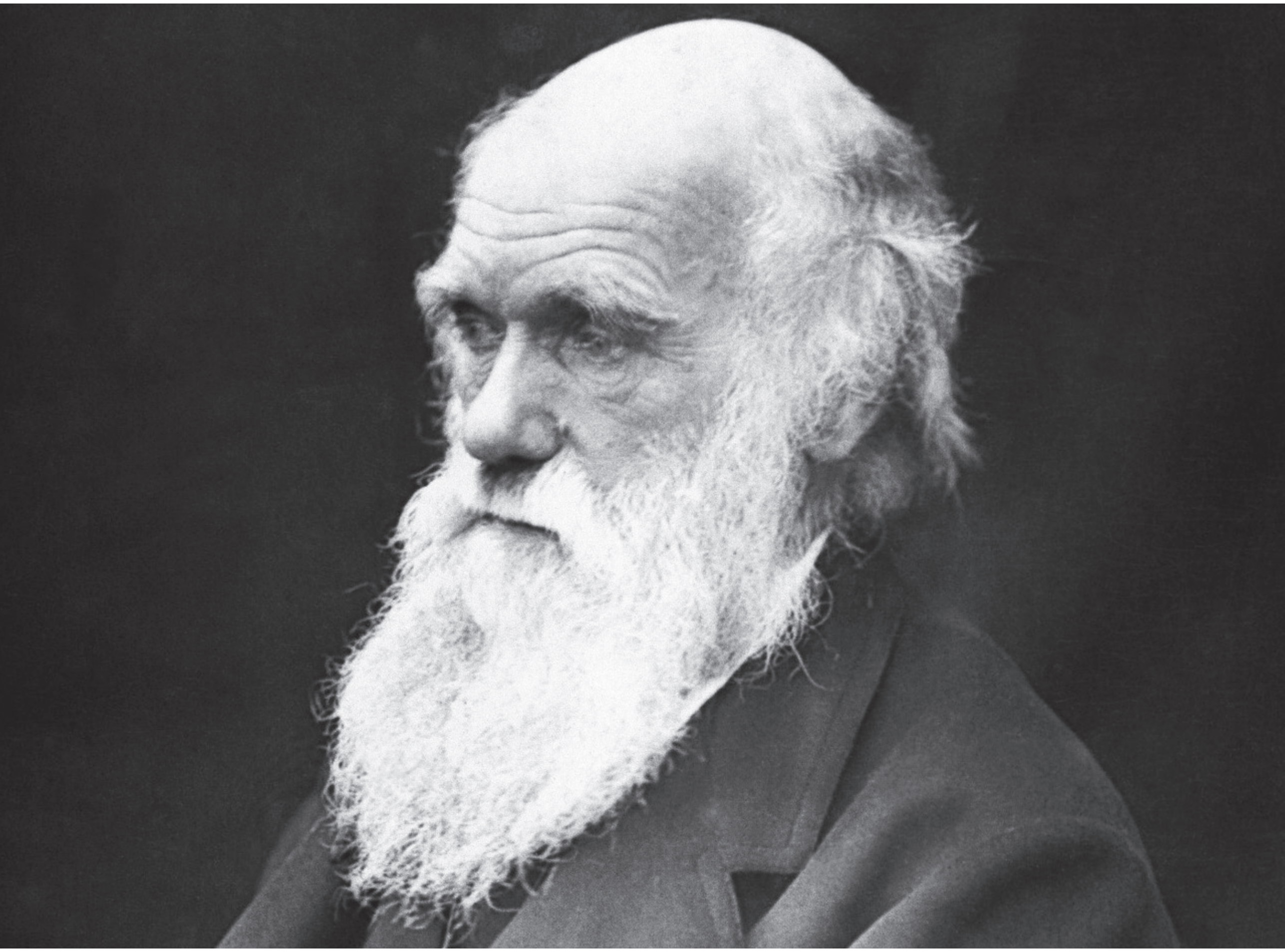
Chapter four focuses on estimation methods of one of the most widely used analyses carried out by ecologists: species richness. As nonparametric estimators are probably the most used techniques to carry out such estimations, the assumptions and results of nonparametric estimators are compared with those of a logseries approach to species richness estimation. In addition, it studies the potential of extrapolation of patterns in species richness to larger scales necessary to implement in neutral models.

Chapter five compares different estimation methods for estimating one of the core parameters of neutral models: migration. With many sophisticated methods available for estimating migration, ecologists face the difficult decision of choosing for their specific line of work. We compare and test a number of methods for their ability to estimate migration from spatially implicit and semi-explicit simulations. In addition, we provide suggestions to correct one of the methods to be implemented as estimator of migration for the newly developed semi-spatially explicit neutral model used in the next chapter.

Chapter six adds a level of biological reality to predictions from neutral theory, emphasizing the novel spatially semi-explicit neutral model and combining the model with large-scale empirical tests. It uses the three different datasets introduced in chapter two, with estimates of diversity from chapter four, and migration from chapter five and studies predictions at both local and regional scales to study the scalability of neutral theory and whether correct regional predictions follow from accurately reflected local dynamics.

Chapter seven moves away from mechanistic (neutral) models and uses the mathematical machinery from information theory and statistical mechanics (the maximum entropy formalism) to quantify the relative importance of niche and neutral processes as well as giving estimates for the actual geographic range of where potential recruits can come from in the process of community assembly.

Chapter eight is the final chapter in which I synthesize all results and put them in a wider perspective. It not only provides the closing statement of this thesis, addressing the questions posed at the start but also provides suggestions for future research and proposes a new hypothesis of the governing dynamics of communities to explain community structure.



We need not marvel at extinction; if we must marvel, let it be at our presumption in imagining for a moment that we understand the many complex contingencies, on which the existence of each species depends.

Charles Darwin, On the Origin of Species (1859: page 322)

Chapter Two

A primer on Neutral Theory and Maximum Entropy

Classical ecology states that coexistence of species and community structure follows from complex interactions determined by quantitative selection over time. In contrast, however, protagonists of neutral theory suggest more simple but strict rules of stochasticity are more important in determining community structure. This chapter provides a primer into neutral theory, providing the basis of terms, processes and paradigms that will be discussed in the following chapters. It starts by shortly introducing classical ecological theory and then moves on to neutral theory. It concludes with a short introduction to the use of Maximum Entropy, a principle from information theory used in chapter seven.

2.1 Species coexistence in classical ecology

The niche forms a fundamental aspect of traditional ecological thinking. It tells us that each species inhabits a specific part of the ecosystem not only defined by a physical location but also the interactions with biotic and abiotic elements within the ecosystem. Environmental heterogeneity, inter-specific competition and resource partitioning between species work together with niche differentiation to allow coexistence of multiple species. This principle of coexistence was formalized more than a century ago by Ernst Haeckel in 1869 (42), who coincidently was the first to coin the term *Oecologie*. It comes as no surprise that our observations of nature, in which species seem to be almost perfectly adapted in both morphological and functional appearance lead to these early conclusions on the seemingly *"lock and key"* principles of evolution and the occupation of specific habitats of species (43). The idea of the niche, however, did not come into existence until the beginning of the twentieth century when Grinnell (44) proposed the so-called pre-interactive or potential niche as it was later known (45). It was defined by the overlap between abiotic and biotic elements and the fundamental requirements of organisms for living and reproduction in the absence of competition or predation. Shortly thereafter, a more functional niche concept also incorporated trophic relations between organisms and their place in the food web (46). The current view of the

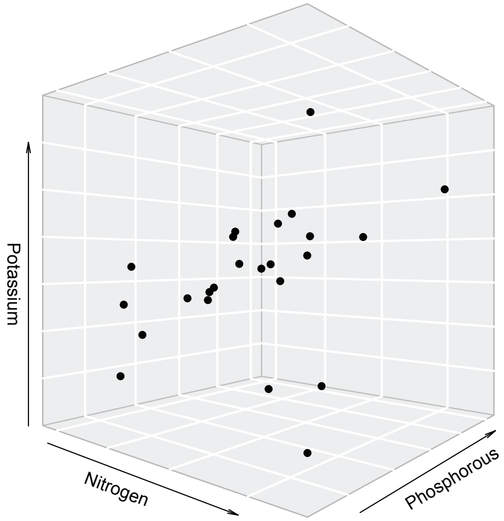


Figure 2.1
The niche hypervolume
Based on three chemical soil properties from a dry *Pinus silverstris* forest in Sweden. Adapted from (49) and created using the package *plot3D* (50) in the R statistical environment (51).

niche is a product of Gause's axiom, the statement that if two species are (almost) identical in their niche characteristics they simply cannot occupy the exact same location (47) and Hutchinson who moved the niche concept from the environment to the organisms themselves, making the niche an attribute of species. Thereby viewing it more as a continuum (48). According to the latter view, an ecological niche should be considered as a multidimensional space of environmental variables, termed a hyper volume (Fig. 2.1).

Within the niche, a particular species can flourish, for values above and below this hyper volume its performance in terms of survival and reproduction decreases. Such niche based models and theories can provide great insights, e.g. in ecosystem functioning, the dynamics of invasive species but also the evolution of adaptation, whereas other theories as of yet cannot (54). Intuitively, this idea of niches is also very straightforward. The quantification of niches and determining whether species diversity is a direct result of non-overlapping niches in which species coexist by interspecific competition and resource partitioning is, however, another matter. It suffers from being overly complex. These same complexities of niche theory lead many scientists to develop alternative models to study coexistence of species and observed patterns in species abundance patterns, such as Neutral Theory.

2.2 Neutral theory of biodiversity

If we scoop a few liters of water from the ocean we find a huge variety of planktonic species in a seemingly very homogenous environment and at first glance appearing to have a very similar niche. According to Gause's axiom this should be nearly impossible or at the very least very improbable and it was dubbed the "*plankton paradox*" (55). Such observations argued whether there was a niche for niches in community ecology at all. They also lead to the earliest equilibrium versus non-equilibrium discussions, precursors to the niche versus neutral debate. According to the first, species diversity is maintained due to functional differences such as life history strategy (56), habitat preference (57) or pathogens, pests and predators (58, 59). Here, due to forces of selection, composition roughly stays the same (in equilibrium). Non-equilibrists, however, put more emphasis on processes as speciation, immigration and extinction to maintain diversity (53, 60–62), with composition continuously changing (hence the non equilibrium).

During the early sixties, MacArthur and Wilson provided one of the first mechanistic neutral models for ecology with the equilibrium theory of insular zoogeography (53). These ideas were furthered by analyses of neutral models from Caswell and Hubbell (60, 61). Kimura had, however, already paved the way for neutral theory in ecology with his Neutral Theory of Molecular Evolution, leading the way for other such models where alleles were replaced by individuals in an ecological context (63, 64). Hubbell formalized these ideas in his Unified Neutral Theory of Biodiversity and Biogeography at the turn of the twenty-first century (29). For the unfamiliar reader, box one on the next page provides the necessary background in the workings of neutral theory.

From box one it is clear that in the original model it is only migration that is determining the relative abundance distribution patterns of local communities, whereas speciation and extinction (both random processes as well) regulate diversity in the metacommunity. And although the results of neutral theory showed good fits to actual field data, there is one fundamental assumption that poses a serious problem for ecologists: panmixis, or the ability to move around freely without any restriction. This comes from the fact that although there is an implicit spatial relation between the local and the metacommunity as is defined by the migration parameter m, within each there is no spatial dependency. In other words, any immigrants or local recruits can come from any location within their respective community. From numerous field experiments and studies on betadiversity we know it is likely that offspring will be recruited close by the parent and that there is a strong distance decay of similarity in composition, even within homogeneous environments (65). Charles Darwin himself also noted that: *"migration is likely as important for selection, and hence eventual composition, as the environment itself"* (1).

Box 1 Classical Neutral Theory

Classic neutral theory models communities at two different spatial and temporal scales: a large metacommunity consisting of *Jm* individuals operating at evolutionary timescales and a smaller local community having *J* individuals operating at ecological timescales. It is quite similar to original Mainland-island models by Wright, MacArthur and Wilson (55, 62) where the metacommunity represents the main land and the local community is the island (Fig. 2.2). Both communities are governed by a zero sum assumption of a saturated landscape meaning that they are saturated with individuals and replacement is immediate, not allowing for gaps in the community. All individuals in the neutral model also have equal demographic probabilities (i.e. the same chance of birth and death) and therefor recruitment is proportional to the species abundance.

Local community dynamics. Each time step, an individual in the local community dies and is immediately replaced. Replacements can either be an immigrant from the metacommunity with probability *m* or the offspring from a randomly chosen individual within the local community at probability *1-m*. The parameter *m* can be considered a measure of dispersal, a probability between zero and one that a replacement will be an immigrant rather than a local recruit. At values between zero and one it can either be severely or hardly dispersal limited, depending on whether values approximate zero or one. When *m* is equal to one there is no dispersal limitation and all replacements are from the metacommunity. In this special limiting case, it represents, in essence, a direct (random) sample from the metacommunity. On the other extreme, if *m* equals zero this means that for every replacement, none are coming from the metacommunity but from the local community itself. This inevitably leads to a closed community with an absorbing state: monodominance of a single species due to the process of ecological drift (the analog of genetic drift).

Metacommunity dynamics. The metacommunity represents the major source pool of species. A process similar to that of the local community determines composition of the metacommunity. However, as it operates at a different spatial and temporal scale, neutral theory assumes the species abundance distribution does not change on local community timescales and it does not receive immigrants but instead allows for speciation and extinction events. It is assumed the abundance distribution of the metacommunity is in equilibrium at local community timescales. The expected species richness and relative species abundance in the metacommunity under the preceding assumptions are controlled by a single parameter θ (theta), the fundamental biodiversity number. At small speciation rates it is approximated by *2Jmv*, with *Jm* being the total number of individuals in the metacommunity and *v* being a constant rate of speciation. Theta here is analogous to Kimura's theta, defining the homozygosity of a population in a stable equilibrium (60). Although in general *Jm* is quite large, speciation rates are very low, resulting in a θ of intermediate size. High values of theta can be the result of either a large metacommunity size or high speciation rates. If θ is small the predicted dominance-diversity curve becomes steep, representing high rates of dominance within the metacommunity. However, as θ becomes larger it starts predicting the often-observed logseries like species abundance curve (29). It is of note that when metacommunity size increases, in the absence of migration, theta becomes equal to fishers alpha (29).

As a consequence, Hubbell already recognized early on that this original spatially implicit model lacks any form of functional betadiversity and is a poor reflection of the real world (29). In the field of population genetics, this same issue had already been tackled by providing solutions for the correlation between allelic states among individuals that were separated by a certain distance (*r*), for instance between islands and the mainland (63, 52). Several solutions to calculate *F(r)*, the probability that two randomly chosen alleles are the same, have been proposed over the following years (63, 66–69) which also found their way into neutral theory (35, 70). Following this, spatial dependency was incorporated into neutral models in various ways. From grid like models incorporating differential probabilities of migration depending on distance (71) to individual based models in which dispersal ability could be varied and life history strategy differences among species could be added, as well as conspecific density dependence for each species (72). And although these brought neutral theory closer to the real world, they inevitably also made it more complex. The first step towards an analytically tractable model for the quantification of beta diversity came with the extension of the original neutral model by a hyper-cubic lattice with d dimensions where each site of the model represents a single individual (73). Such developments incorporated at least one extra level of spatial dependency, namely that within the local community but not the overarching dependency between communities.

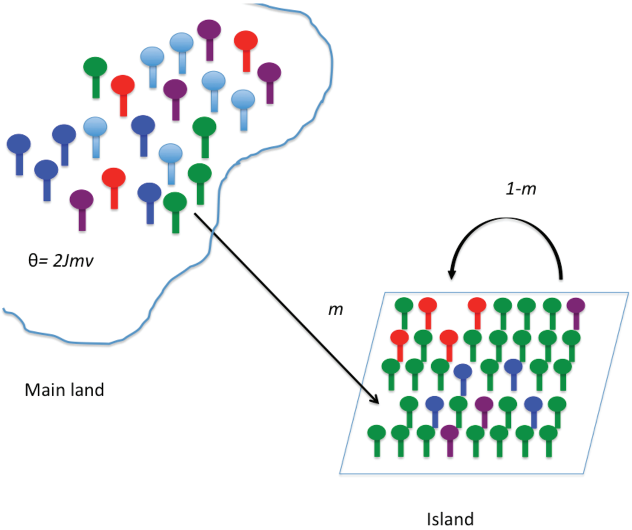


Fig. 2.2 The original spatially implicit neutral model. Adapted from original Mainland-Island models (52, 53) where a larger metacommunity is represented by the mainland and a smaller local community by the island. Migration determines the relation between the metacommunity and local community with 1 minus the migration probability giving the probability of a local recruit after a death in the local community.

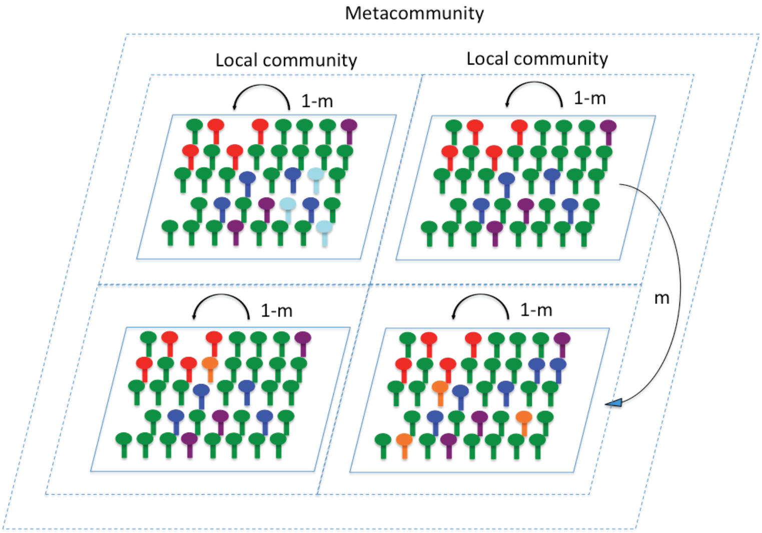
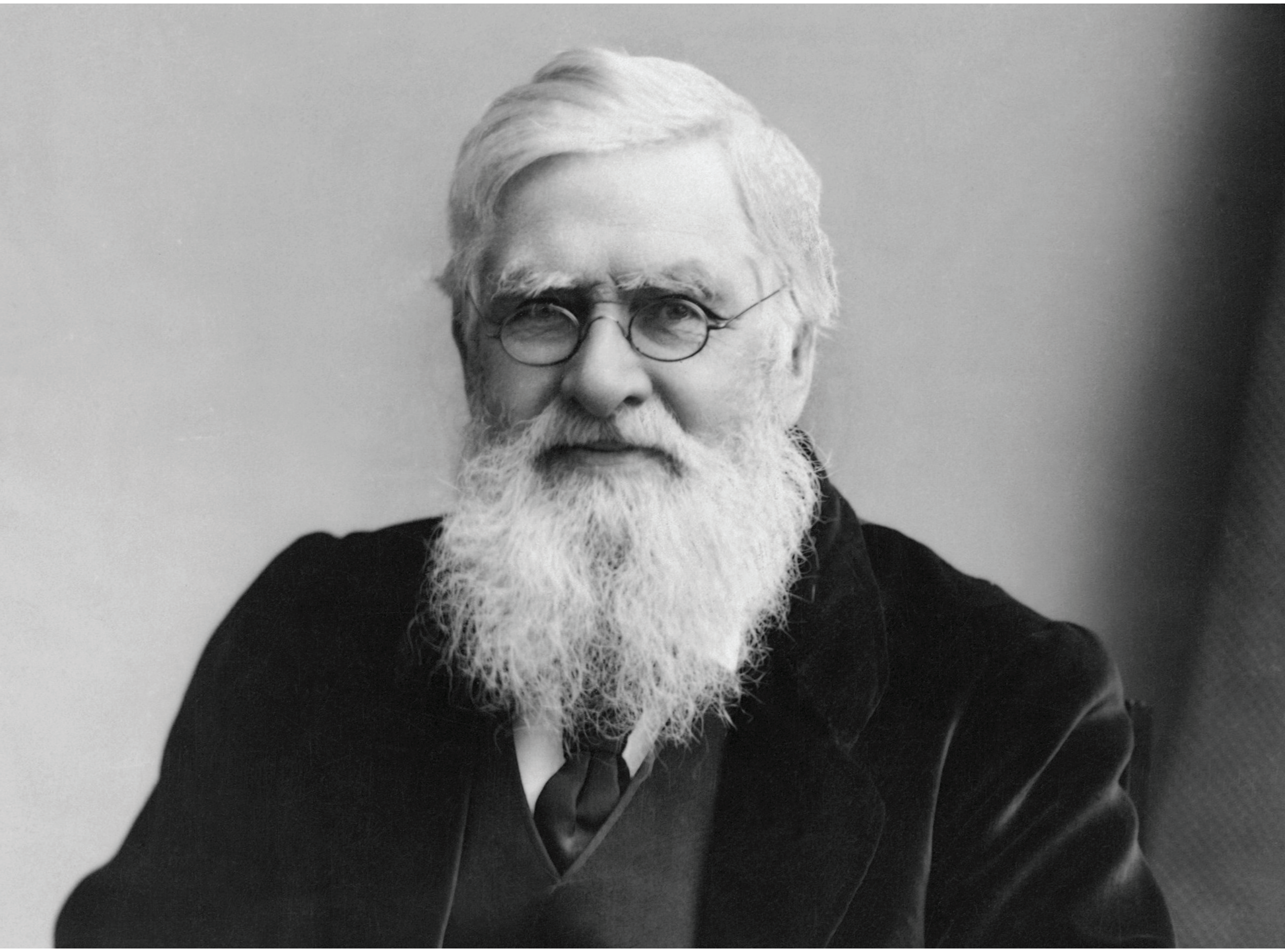


Fig 2.3 A collection of local communities, with the summation forming the larger metacommunity. Local communities are connected by the same migration parameter although this is now migration from plot to plot. It is, however, still an approximation of migration, as the intermediate plots are not taken into account and migration still acts as an ecological aggregated parameter, incorporating not only dispersal but also filtering and recruitment.

To solve the latter, one attempt was to create a metacommunity model in which a local community was embedded within the metacommunity instead of connected to it via an abstract parameter of migration (Fig. 2.3) (74). The local community, however, was still a separate entity from the metacommunity. Later, local communities were truly embedded in the metacommunity by introducing a continuous landscape (75) and this idea was further developed providing an analytical approach to examine a network structure of communities (76), making neutral theory more and more biologically realistic. Although the above developments reflect progress in neutral models approaching the real world by incorporating some form of spatial dependency, which is far from done, they still lack a good connection to empirical data as this is also subject to sampling schemes and sampling errors usually not incorporated in such models. This thesis provides a first step towards solving these issues in chapters five and six.

2.3 A short primer on Maximum Entropy

Chapter 7 of my dissertation focuses on the use of principles from information theory and statistical mechanics: the *Maximum Entropy Formalism (MEF)*. Moving away from mechanistic models for which parameters need to be estimated and assumptions need to be made, the MEF allows for objective inference of relative importance of various aspects of ecological communities without invoking such assumptions and parameters. This formalism has its foundation in two different fields of science: information theory and statistical mechanics. Edwin Jaynes combined both to construct the MEF in 1957 (77, 78). After more than 20 years the MEF was applied to ecological problems (79, 80). Here, the process of community assembly is viewed from a statistical mechanic viewpoint. In other words, the idea that a macroscopic pattern (e.g. a species abundance distribution) arises from random microstate allocations forced to obey certain macroscopic constraints. These constraints in classical statistical mechanics are given by physical assumptions not appropriate for ecological communities (81–83) but the principle remains the same. In short, the allocation of resources at the microstate level interacts with natural selection between entities (being species, reproductively isolated genotypes or any other taxonomic identifiable group). This results in a balance between random and deterministic allocation. In other words, fitness differences between entities bias these random allocations. If fitness differences are heritable and repeatable in time and space, this will generate repeatable community structures and ultimately will lead to observed patterns of community structure. This also led to the title and cover of this thesis, in which we draw a comparison between species in life and individuals in a casino. To quote Shipley: *"The analogy that emerges is of Nature as an immense casino. The species play craps with loaded dice for resource payoffs. There is no guarantee of success, only a probability of success. The dice that each species uses are biased due to the unique traits that each possesses, thus weighting the probability one way or another, but whether or not the bias helps or hurts the species depends on the nature of the tables (environments) on which the dice are thrown"* (40). The dice on the front of this thesis represent exactly these loaded dice that species are playing with in the casino of life. What we then try to do using the MEF is to find how these dice are biased and how this depends on the table (e.g. environment) on which entities are playing. This is done by calculating the importance of functional traits relative to for instance the abundance in the regional species pool independent of these traits and demographic stochasticity. In other words, we try to find out in what way species are cheating an otherwise fair and stochastic game.



If the traveler notices a particular species and wishes to find more like it, he must often turn his eyes in vain in every direction. Trees of varied forms, dimensions, and colors are around him, but he rarely sees any of them repeated. Time after time he goes towards a tree which looks like the one he seeks, but a closer examination proves it to be distinct.

Alfred Russel Wallace, Equitorial Vegetation (1891)

Chapter Three

Are all species necessary to reveal ecologically important patterns?

(Published in Ecology and Evolution 4.24 (2014): 4626-4636)

Edwin Pos[1,2], Juan Ernesto Guevara Andino[3], Daniel Sabatier[4], Jean-François Molino[4], Nigel Pitman[5,6], Hugo Mogollón[7], David Neill[8], Carlos Cerón[9], Gonzalo Rivas[10], Anthony Di Fiore[11], Raquel Thomas[12], Milton Tirado[13], Kenneth R. Young[14], Ophelia Wang[15], Rodrigo Sierra[13], Roosevelt García-Villacorta[16,17], Roderick Zagt[18], Walter Palacios[19], Milton Aulestia[20], Hans ter Steege[1,2]

[1]Naturalis Biodiversity Center, Section Botany, Leiden, The Netherlands
[2]Ecology and Biodiversity Group, Utrecht University, The Netherlands
[3]Department of Integrative Biology, University of California, Berkeley, CA 94720-3140, USA
[4]IRD, UMR AMAP, Montpellier, France
[5]The Field Museum, 1400 S. Lake Shore Drive, Chicago, IL 60605-2496, USA
[6]Center for Tropical Conservation, Nicholas School of the Environment, Duke University, Durham, NC 27708, USA
[7]Endangered Species Coalition, 8530 Geren Rd., Silver Spring, MD 20901, USA
[8]Universidad Estatal Amazónica, Puyo, Ecuador
[9]Universidad Central Herbario Alfredo Paredes, Escuela de Biología Herbario Alfredo Paredes, Ap. Postal 17.01.2177, Quito, Ecuador
[10]University of Florida, Wildlife Ecology and Conservation & Quantitative Spatial Ecology, 110 Newins-Ziegler Hall, PO Box 110430, Gainesville, Florida, USA
[11]Univ. of Texas at Austin, Department of Anthropology, SAC 5.150, 2201 Speedway Stop C3200 Austin, Texas 78712, USA
[12]Iwokrama International Programme for Rainforest Conservation, Georgetown, Guyana
[13]GeoIS, El Día 369 y El Telégrafo, 3° Piso, Quito, Ecuador
[14]University of Texas, Geography and the Environment, Austin, Texas, 78712, USA
[15]Northern Arizona University, Flagstaff, Arizona, 86011, USA
[16]University of Edinburgh, Institute of Molecular Plant Sciences, Mayfield Rd, Edinburgh, EH3 5LR, United Kingdom
[17]Royal Botanic Garden of Edinburgh, 20a Inverleith Row, Edinburgh EH3 5LR, UK
[18]Tropenbos International, Lawickse Allee 11, PO Box 232, Wageningen, 6700 AE, the Netherlands
[19]Universidad Técnica del Norte, Herbario Nacional del Euador, Quito, Ecuador
[20]Herbario Nacional del Ecuador, Casilla 17-21-1787, Avenida Río Coca E6-115, Quito, Ecuador

Abstract

Aim While studying ecological patterns at large scales, ecologists are often unable to identify all collections, forcing them to either omit these unidentified records entirely, without knowing the effect of this, or pursue very costly and time-consuming efforts of identifying them. These indets may be of critical importance but as of yet their impact on the reliability of ecological analyses is poorly known. We investigated the consequence of omitting the unidentified records and provide an explanation for the results.

Location South America (Guyana, Suriname, French Guiana and Ecuador)

Methods We used three large-scale independent datasets, each consisting of records having been identified to a valid species name (Identified Morpho-Species - IMS) and a number of unidentified records (Unidentified Morpho-Species - UMS). A subset was created for each dataset containing only the IMS, which was compared with the complete dataset containing All Morpho-Species (AMS: = IMS + UMS) for the following analyses: species diversity (Fishers alpha), similarity of species composition, Mantel test and ordination (NMDS). In addition we also simulated an even larger number of unidentified records for all three datasets and analysed the agreement between similarities again with these simulated datasets.

Results For all analyses results were extremely similar when using the complete datasets or the truncated subsets. IMS predicted ≥ 91% of the variation of AMS in all tests/analyses. Even when simulating a larger fraction of UMS, IMS predicted the results for AMS rather well. Using only IMS also out-performed using higher taxon data (genus level identification) for similarity analyses.

Main conclusions Finding a high congruence for all analyses when using IMS rather than AMS suggests that patterns of similarity and composition are very robust. In other words, having a large number of unidentified species in a dataset may not affect our conclusions as much as is often thought.

Keywords: Beta-diversity, Fishers alpha, species richness, indets, morpho-species, large-scale ecological patterns, similarity of species composition, Mantel test, NMDS, spatial turnover.

Introduction

In comparative ecology, the proper naming of species is essential. Historically, ecological studies have assigned a particular name to a particular entity based on the Darwinian species concept, which uses morphological characters to separate clusters of individuals into species (1, 84). While studying ecological patterns at large scales, ecologists are often unable to identify all individuals encountered in the field to species. This leads to a potential problem: individuals that are recorded in a dataset but which have no valid species name (hereafter *indets*). As databases grow larger, so does the number of indets, with each plot added to a database also adding a number of new unidentified morpho-species (UMS), which ecologists must either incorporate or ignore in analyses. Both of these options potentially introduce errors of some sort, and there is no agreement among ecologists how indets should be handled or to what degree they might compromise the results of large-scale analyses. These questions have been addressed on multiple occasions. Pitman et al. (1999), comparing tree species communities, also posed the question what would be the result of eliminating species that lacked taxonomic identification. In their view the only variable that would substantially change with more individuals identified to a species was the geographic range of a species (85). Following this statement, Ruokolainen et al. (2002) focused on the geographical ranges of identified vs. unidentified species previously mentioned by Pitman et al. (1999) and agreed that this bias has the potential to greatly distort analyses and added that it is not necessarily confined to distributional patterns (86). Some might be more obvious than others; species richness will be underestimated when unidentified specimens belong to new species and this will also affect the relative abundance distribution. Similarities of species composition may also be affected, which will affect subsequent analyses that depend on these similarities, importantly Mantel tests and ordinations, tests that are often used by ecologists.

Many studies have sought a middle ground between high-cost, taxonomically precise analyses and more cost-effective methods without losing valuable ecological information, for instance by relaxing taxonomic resolution ((87) and references therein) or by randomly reassigning UMS to identified species present in other plots or to itself again, in which case it was considered a new species (88). This, however, unintentionally increases similarity between plots. In several studies, correlations were in fact found between different taxon-level approaches and the patterns in abundance and composition in both marine and terrestrial habitats (89–92). In an attempt to abbreviate forest inventories, Higgins and Ruokolainen also made use of higher taxon level analyses by elinimінating species identifications (93). While promising, these studies mostly dealt with unidentified species by decreasing taxonomic resolution, allowing the use of more individuals from a dataset without identification up to species-level. However, as Terlizzi et al. (2003) have noted, many large-scale ecological questions (e.g., species loss or the degradation of forest diversity) require species-level analyses (87). And, while new analytical tools offer some help in standardizing ecological datasets, removing synonyms, and checking the validity of names (e.g., the Taxonomic Name Resolution Service or TNRS (94) and the R packages *taxize* (95) and *Taxonstand* (96)), they cannot help solve the indet problem. In a theoretical approach, it was shown that by subsampling datasets at random, thereby simulating a random sampling at a lower intensity, and by making subsamples based on the difficulty in identifying them, the outcome of analyses on species richness and composition do not necessarily change (97). The first probably being the result of the relative abundance distribution theoretically remaining identical even with smaller subsamples, because of the random sampling. To our knowledge the effect of omitting unidentified species has not yet been tested with actual data containing unidentified records at a scale as presented here.

Here we use three independent large-scale harmonized and standardized tree inventory datasets (Guyana/Suriname, French Guiana and Ecuador) to test whether ecological patterns such as species diversity, richness, composition and underlying gradients in the full datasets, using all morpho-species differ from those in subsets of identified morpho-species. This was done using three often-used analyses: Species richness and Fishers alpha (98), to study patterns in tree species diversity, the similarity of species composition between samples for studying patterns in species turnover (65) and non-metric multidimensional scaling (NMDS), an ordination technique designed to search for patterns in community composition. We also tested the similarities using a higher taxon level, in this case genus-level, against results generated by the complete dataset (i.e. all morpho-species, the sum of the identified morpho-species and unidentified morpho-species included). These tests have significant practical implications, because a finding of no difference between using only identified morpho-species or all morpho-species would suggest a simple solution to the indet problem: omitting them altogether. In turn, this might make it possible to use large datasets that are currently underutilized in ecology because they contain large numbers of indets.

Table 3.1 The number of one hectare plots for each forest type listed by country. Guyana and Suriname are used as one dataset. Type abbreviations are igapó (IG), podzol (PZ), swamp (SW), terra firme (TF) and várzea (VA). Minimum Diameter at Breast Height (DBH) as limit for measurement was 10 centimeters for all plots.

	IG	PZ	SW	TF	VA	Min. DBH	Nr. 1 Ha plots
Guyana/Suriname	0	21	0	45	1	10	67
Ecuador	2	3	4	53	10	10	72
French Guiana	0	0	0	63	0	10	63
Total	2	24	4	161	11	NA	202

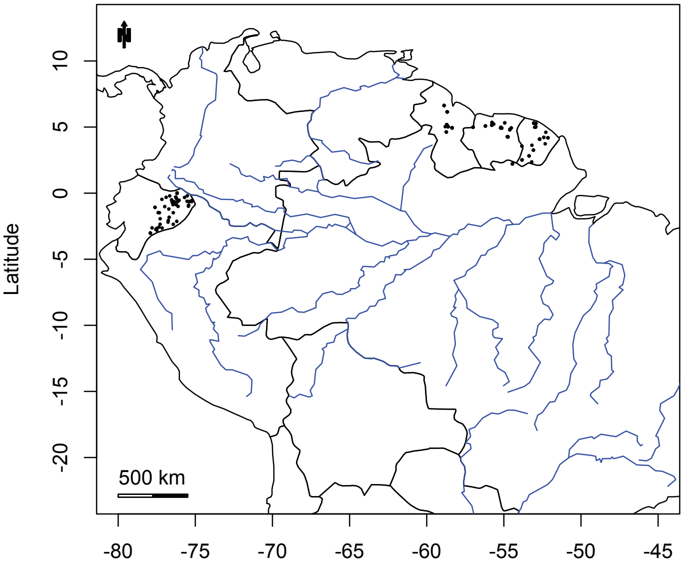


Fig. 3.1 Map showing location of all 202 plots belonging to the Ecuador (blue), Guyana/ Suriname (red) and French Guiana (black) datasets.

Methods

Species composition data. Three independent, non-overlapping, tree inventory datasets were assembled: one from Guyana and Suriname, one from Ecuador, and one from French Guiana (Fig. 3.1). Each dataset consisted of 63-72 one-hectare plots, in which all trees ≥10 cm DBH had been inventoried (see Table 3.1 for details). Within each dataset, one or two persons responsible for the majority of the collected material harmonized all species names. Olaf Bánki and Juan Ernesto Guevara performed harmonization for the Guyana/Suriname and Ecuador datasets, respectively, while Daniel Sabatier and Jean-François Molino together harmonized the French Guianan dataset (hereafter referred to as OSB, JEG, S-M). Harmonization was done by morphological comparison of collections with reference to a 'morpho-holotype' for each morpho-species. Species names of all subsets were standardized with the W3 Tropicos database, using TNRS (94). The three datasets were harmonized independently of each other; no attempt was made to harmonize the three datasets into one. Three types of common ecological analyses (described below) were performed for each dataset, twice: once for the all morpho-species (hereafter AMS) and once for a subset composed of only identified morpho-species (IMS), omitting the unidentified morpho-species of this dataset (UMS – thus AMS = IMS + UMS). All tests were performed in the R statistical and programming environment (51). To calculate the Mantel statistics and metaMDS (a variant of NMDS) we used the package *vegan* (99). All linear models were tested for significance with a permutation procedure from the package *lmperm* (100).

Diversity analyses. To test how UMS influence analyses of alpha and beta diversity, we calculated Fisher's alpha values (98) for every one-hectare plot twice: once with AMS and once for only IMS. We then performed a linear regression analysis between Fisher's alpha calculated for AMS and IMS to determine whether diversity patterns remain the same when datasets are truncated like this. Fisher's alpha is a widely used diversity index, specifically suited for species abundances following a logseries distribution. Fisher's alpha has been shown to be a very efficient diversity index for discriminating between sites (101). This is a consequence of Fisher's alpha being theoretically independent of sample size and therefore much less influenced by the abundances of the more common species (102, 103). If UMS can safely be excluded from the dataset we expect to find no deviation from the pattern predicted by using only IMS or AMS and high R^2 values from the linear regression analysis. We do expect, however, as UMS are especially common among the rare species, that omitting UMS may result in a significant decrease in Fisher's alpha, which was tested by a paired sample t-test.

Similarity in species composition. To examine whether floristic similarity between plots differed when using AMS or only IMS we constructed floristic similarity matrices for each dataset and a geographical distance matrix between the plots. Again, this was done twice for each dataset: once for AMS and once for IMS. We calculated the Mantel statistic (104) as the matrix correlation between the two similarity matrices (in this case the floristic and the geographical matrix). Random permutation of both rows and columns of the species similarity matrix is then used to evaluate the significance of the performed test (105). We performed a linear regression between the pairwise similarities between all plots of each dataset to assess the prediction of similarity values based on only the IMS. Because the two similarity matrices (i.e. based on IMS only or AMS) are not independent, this should be interpreted as underestimates of the risk to abandon the null hypothesis of no dependence between the matrices. However, we need to stress that despite the non-independence, this is exactly the test we need to perform, as we are interested if IMS are a good predictor of AMS. Floristic similarity values were first calculated with the *Bray-Curtis* index of similarity, which is based on both species occurrence and abundances at each site (106). For comparison, we also used the *Jaccard* index and the *Sørensen* index to calculate similarities. The Jaccard index is only based on species presence or absence, i.e. ignoring differences in species abundance (107) and calculates similarity as the number of shared species between two sites divided by the total number of species of the two sites combined. The Sørensen index (108) is in essence much the same as the Jaccard index with the exception of giving double the weight to the shared species. To test the degree to which pairwise communities are more different or more similar than expected by chance we used the *Raup-Crick* distance metric and repeated the above analyses. The Raup-Crick metric (β_{RC}) was previously used in Paleontological studies and just recently in some works related to variation in beta diversity and species turnover (109–111). The β_{RC} metric calculates the similarity between two communities under a null model. Assuming that $SS_{1,2}$ is the number of shared species between two communities with values of alpha diversity α_1 and α_2 respectively, the β_{RC} is obtained by random draws of α_1 and α_2 species from a determined species pool to estimate the probability of observing the shared species. The Mantel statistic was first calculated based on the standard distance matrix function in vegans vegdist (99). We then used the Raup-Crick method, under a null-model assuming that the occurrence probability of species is frequency dependent, and performed the Mantel's statistic and linear regression on the matrices of pairwise similarities again. Similar to the diversity analyses, if omitting UMS from our datasets indeed makes no difference we again expect to find high R^2 values from the regression between analyses performed on IMS and AMS. In addition we also tested for the deviation from a slope of 1 belonging to the relationship of $y = x$ (i.e. when IMS and AMS generate the exact same results).

Table 3.2 Adjusted R² coefficients from the linear regression for each analysis; listed for all three datasets. All regression coefficients were found significant at a 0.001-signficance level after 5000 permutation iterations. Results of stratification were averaged over 50 runs for each diversity index.

	Valid vs Morpho Guyana/Suriname	Ecuador	French Guiana
Fishers Alpha	0.967	0.959	0.970
Mantell Bray-Curtis	0.983	0.998	0.999
Mantell Bray-Curtis (genus level)	0.739	0.805	0.904
Mantell Jaccard	0.983	0.998	0.999
Mantell Sørensen	0.966	0.995	0.996
Raup-Crick	0.918	0.955	0.967
NMDS axis 1	0.979	0.998	0.9997
NMDS axis 2	0.991	0.988	0.998
Stratification (50%) Bray Curtis	0.80 (SD 0.17)	0.92 (SD 0.042)	0.92 (SD 0.05)
Stratification (50%) Sørensen	0.60 (SD 0.073)	0.85 (SD 0.02)	0.81 (SD 0.051)
Stratification (50%) Jaccard	0.78 (SD 0.19)	0.91 (SD 0.04)	0.92 (SD 0.05)
Stratification (25%) Bray Curtis	0.59 (SD 0.2)	0.81 (SD 0.07)	0.82 (SD 0.09)
Stratification (25%) Sørensen	0.51 (SD 0.12)	0.75 (SD 0.06)	0.71 (SD 0.097)
Stratification (25%) Jaccard	0.59 (SD 0.19)	0.79 (SD 0.072)	0.81 (SD 0.095)

To test whether using a higher taxon approach would yield similar results as the approach based on AMS as above, we also tested results from a similarity analysis based on only genera against the results of the AMS dataset. Agreement between similarities was analysed using the same procedure as above.

Multivariate analyses. To evaluate underlying structures of floristic composition within the three datasets we performed Non-Metric Multidimensional Scaling (NMDS) using MetaDMS. Two NMDS were performed separately for each dataset: once for AMS and once for IMS. The scores of the first and second axes were then compared separately by linear regression. NMDS is an ordination technique, which attempts to find the best rank-order agreement between actual similarities in floristic similarity and interpoint distance in the computed ordination space (112–114). NMDS therefore does not try to fit axes based on eigenvalues but instead represents a coordinate system for the ordination space. We used MetaMDS, a NMDS procedure that centers the origin on the averages of the axes and uses principal components to align the scores in such a way that most variation is projected along the first axis (99). We tested the hypothesis that the patterns produced by the NMDS on the first and second axis are similar using either the IMS or AMS and hence that linear regressions will yield high R^2 values. Here we also tested for the deviation from a slope of 1 belonging to the relationship of $y = x$.

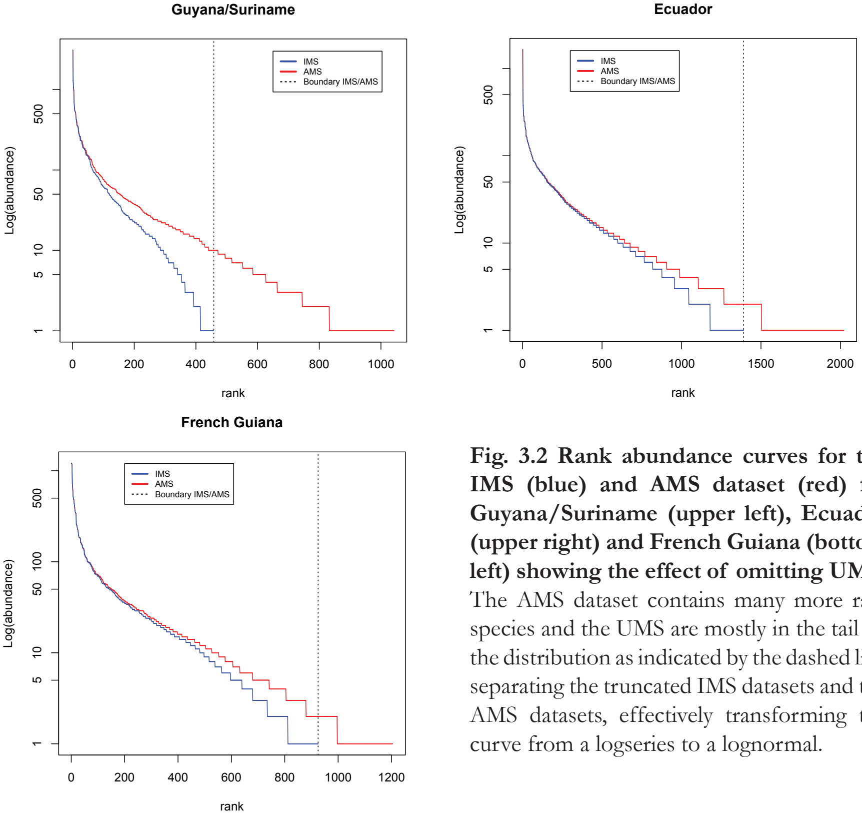


Fig. 3.2 Rank abundance curves for the IMS (blue) and AMS dataset (red) for Guyana/Suriname (upper left), Ecuador (upper right) and French Guiana (bottom left) showing the effect of omitting UMS. The AMS dataset contains many more rare species and the UMS are mostly in the tail of the distribution as indicated by the dashed line separating the truncated IMS datasets and the AMS datasets, effectively transforming the curve from a logseries to a lognormal.

Data stratification. To test for the robustness of predictions based on IMS, we created random smaller subsets to perform the same Mantel test as explained above. A random subset of respectively 50% and 25% was selected from the Guyana/Suriname, French Guiana, and Ecuador IMS pool. In making the IMS dataset even smaller in comparison with the complete dataset (by randomly omitting IMS), we simulated a larger proportion of UMS. This was repeated for 50 iterations from which mean values were calculated for the similarity matrices using the same three indices as used for the similarity analyses described above.

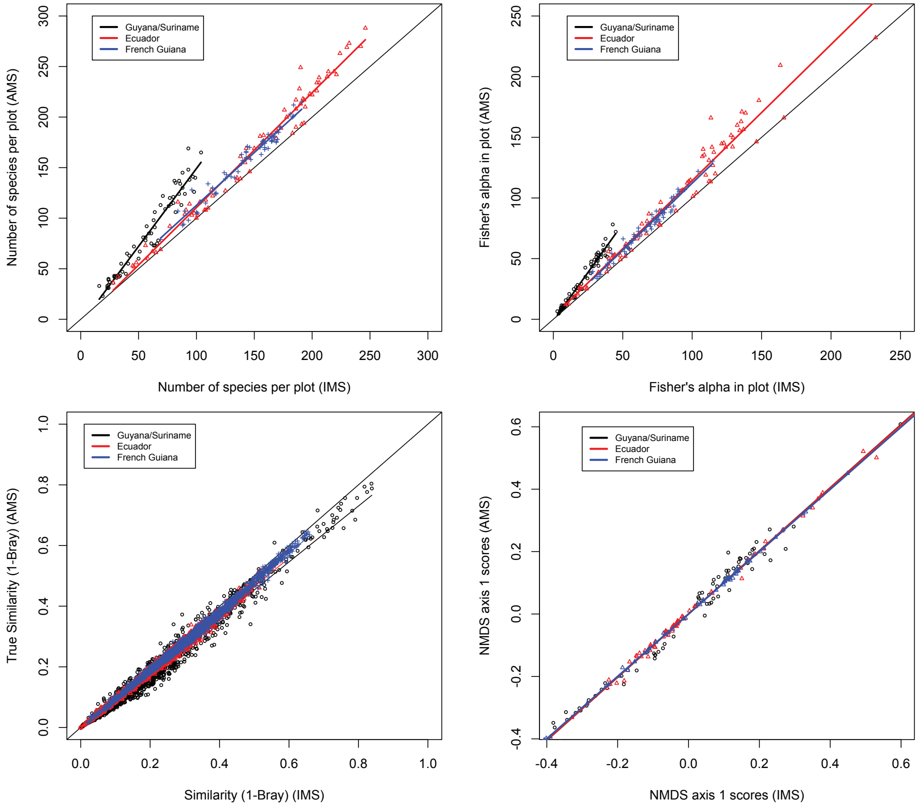


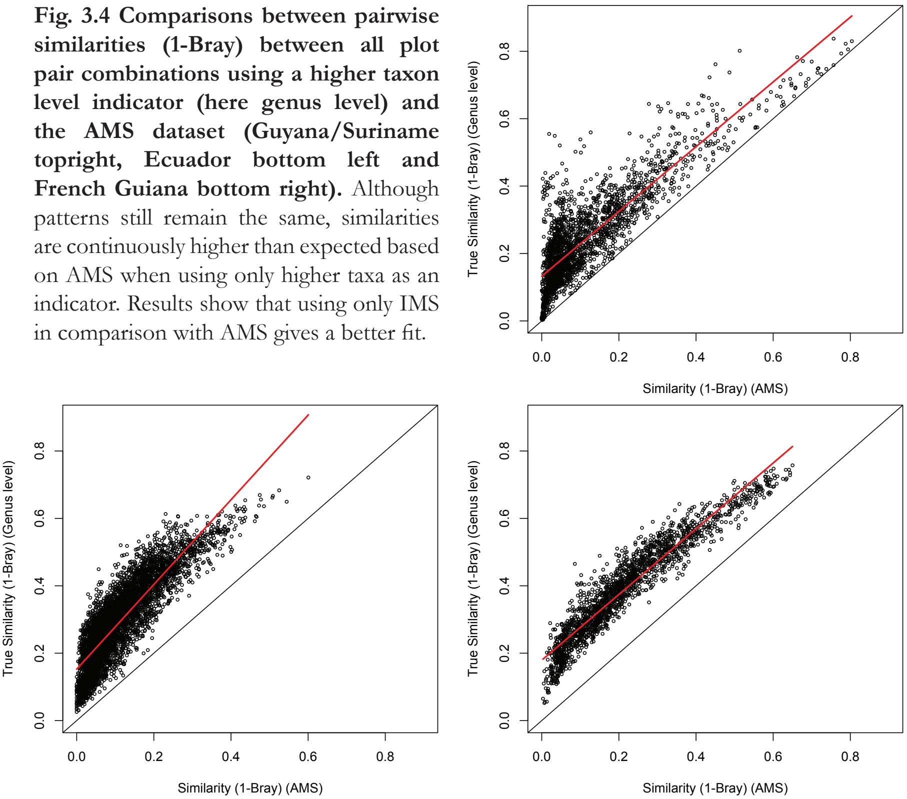
Fig. 3.3 Comparisons between the IMS and AMS dataset for species richness per plot (top left), Fisher's alpha (top right), pairwise similarities between all plot pair combinations (bottom left) and axis 1 scores of the Non Metric Multidimensional Scaling (bottom right). All analyses were performed on the three large subsets Guyana/Suriname (o; black), Ecuador (Δ; red) and French Guiana (+; blue). All analyses show extremely similar results and yield high R^2 values.

Results

Floristic composition and level of species identification

The proportion of IMS varied in the three datasets from 44-77%. In Guyana and Suriname (OSB), 67 plots yielded 37,446 individual trees, for a total of 1042 AMS and 458 IMS (44%). The mean number of UMS per plot was 27 with a median of 24. Mean fraction of IMS per plot for Guyana/Suriname was 70%. Ecuador (JEG) with a total of 72 plots yielded 34,544 individual trees, for a total of 2021 AMS and 1391 IMS (69%), with a mean number of 17 and a median of 16 UMS per plot. The mean proportion of IMS for each plot in Ecuador was 90%. In French Guiana

Fig. 3.4 Comparisons between pairwise similarities (1-Bray) between all plot pair combinations using a higher taxon level indicator (here genus level) and the AMS dataset (Guyana/Suriname topright, Ecuador bottom left and French Guiana bottom right). Although patterns still remain the same, similarities are continuously higher than expected based on AMS when using only higher taxa as an indicator. Results show that using only IMS in comparison with AMS gives a better fit.



(S-M), 63 plots yielded 35,075 individuals of trees, for a total of 1204 AMS and 925 IMS (77%). Mean number of UMS per plot was 15 with a median of 15. The mean proportion of IMS per plot in French Guiana was 91%. Linear regressions between the number of AMS and the number of IMS were high, with R^2 values of 0.938, 0.976 and 0.959 for Guyana/Suriname, Ecuador and French Guiana respectively (Table 3.2).

Predicted species diversity based on identified morpho-species

Linear regressions between Fisher's Alpha (FA) calculated using AMS and only the IMS were extremely high, yielding R^2 values of > 0.95 for all three datasets (Table 3.2). The slope of the linear model based on the Guyana/Suriname was 1.6. Using a 95% confidence interval for the slope showed that this was significantly different from the relation to y = x with slope 1 (i.e. when there is no difference between FA based on AMS or just IMS). This was the case for Ecuador and French Guiana as well, with deviation of the slopes of 1.12 and 1.10 respectively. As expected, FA showed an increase with an increasing number of species per plot for both IMS and AMS. FA calculated for just IMS ranged between 2.87-44.92 for Guyana/Suriname, 8.96-114.65 for Ecuador and 27.61-114.65 for French Guiana. When using AMS this was (in the same order) 4.65-78.17, 12.23-130.32 and 27.61-130.32. These differences were found to be significant after performing a paired sample t-test with significance levels for rejecting the H_0 of equal ranges with probabilities < 0.005 for all three datasets.

Patterns in morpho-species abundance. Because the slope between FA calculated for only the IMS and AMS deviated significantly from 1, we examined the rank abundance curves for both IMS and AMS for each dataset. The AMS datasets were consistently richer in species, especially the rare ones, when compared to the IMS datasets (Fig. 3.2). Moving from the AMS dataset to the IMS more species were lost than individuals, significantly affecting FA. For instance, the IMS dataset contains only approximately 21% of the number of singletons compared to the AMS dataset in Guyana/Suriname. For Ecuador and French Guiana this was 41% and 55% respectively. In terms of numbers there are a total of only 44 singletons in the IMS dataset of Guyana/Suriname against 210 in the AMS dataset (Ecuador = 212 vs. 518 and French Guiana = 114 vs. 208).

Similarity in species composition. Using IMS only, the similarity in species composition based on Bray Curtis was predicted very well for all three datasets (R^2 values of > 0.98) (Table 3.2) and the slope in all cases was almost identical to 1 (Fig. 3.3). Confidence intervals (c.i.) showed, however, that, despite high adjusted R^2 values, slopes from the linear regressions actually deviated significantly from 1 for all datasets when using the Bray Curtis index (Guyana/Suriname c.i. 0.917-0.927, Ecuador 0.958-0.961 and French Guiana 0.979-0.982). The difference between using either the Jaccard, Bray Curtis or Sørensen index for calculating similarities among plots appeared to be negligible, all resulted in adjusted R^2 values of > 0.96 (Table 3.2) with slopes from the linear regressions all still significantly deviating from 1 (for Jaccard: Guyana/Suriname c.i. 0.897-0.907, Ecuador 0.950-0.953 and French Guiana 0.973-0.976 and for Sørensen Guyana/Suriname c.i. 0.915-0.930, Ecuador 0.932-0.938 and French Guiana 0.969-0.974). Adjusted R^2 values using the Raup-Crick distance metric yielded values of > 0.91 for all three datasets. Examples of the patterns of distance decay with AMS and only IMS can be found for all three datasets in the Supporting Information Chapter three. The Mantel's r coefficient for Guyana/Suriname using only IMS was 0.4695; when using AMS this was slightly higher (0.5092). The differences in Mantel's r coefficient were smaller for Ecuador (0.4029 and 0.4039) and French Guiana (0.7944 and 0.7987).

Using higher-taxon level resolution in comparison with identified morpho-species. Using higher taxon level (genus level) data, similarities among communities are higher and much more deviant from the expected similarities based on AMS (Fig. 3.4) than with the IMS (Fig. 3.3). The latter shows a very strong linear regression, while regressions between similarities based on genus level appear to predict the pattern generated by AMS not as good (with R^2 values ranging from 0.74-0.90, Table 3.2) as using only the ISM.

Predictions of Multivariate analyses. Non Metric Multidimensional Scaling of all three subsets showed good segregation along the first two axes of the NMDS when using AMS as well as when using only IMS. Axis 1 scores derived from only the IMS and AMS were very similar (Fig. 3.4). All linear regressions of first axis scores for the AMS and IMS NMDS yielded adjusted R^2 values of > 0.97, for all three datasets. The same pattern emerged from using the second axis with (R^2 values of > 0.99) (Table 3.2). In all cases except French Guiana, deviation of the slopes from 1 was found not to be significant using a 95% confidence interval. Although for French Guiana the CI was between 0.984-0.993. Examples of NMDS results for all three datasets using either AMS or IMS can be found in the Supporting Information Chapter 3: Fig. S7.

Robustness of predictions: data stratification. IMS made up between 44-77% of all species encountered in the datasets (see above). After randomly selecting 50% and 25% of all IMS from each dataset and recalculating the distance decay in Similarity and Mantel's statistic using the Bray-Curtis, Sørensen and Jaccard index, regressions dropped slightly but they still yielded high linear regression coefficients (Table 3.2). For Guyana/Suriname 50 runs with 50% of IMS yielded adjusted R^2 values between 0.60 and 0.80 for the tree indices. Ecuador and French Guiana yielded even higher R^2 values for each index, ranging from 0.85-0.92. In the case of 25% of IMS drawn randomly from the total set of IMS this gave a mean linear regression coefficient R^2 between 0.51 and 0.59 for Guyana and in the ranges 0.75-0.79 and 0.71-0.82 for Ecuador and French Guiana respectively.

Discussion

We asked if omitting individuals that have no valid species name (UMS) from ecological datasets would change the overall result of several important ecological analyses. We showed that when using only the IMS of actual field data, major ecological patterns such as the differences in species richness among sites, floristic similarities among sites and ordination gradients in species composition were maintained. The linear regressions between analyses based on the IMS only or AMS (including all UMS) were extremely high for almost all analyses ($R^2 > 0.91$). This was the case even when simulating a larger fraction of UMS. And although FA underestimated species diversity, when using only IMS, linear regressions between FA from IMS and AMS still showed extremely high R^2 values, suggesting that spatial patterns in diversity will still be similar when using only IMS. However, if individuals can be assigned to morpho-species within plots this will also allow the comparison among plots from different resources (115), including the UMS.

Different methods have been proposed in the past to deal with unidentified morpho-species. By relaxing the taxonomic resolution (87), however, the prediction of similarity between our sites was lower than when omitting UMS (Figs. 3.3 and 3.4). Thus, although a genus level approach allows a larger number of individuals from the dataset to be used, its performance was not necessarily better. Cayuela et al. (2011) used a different method of trimming UMS from a dataset: instead of omitting individuals of UMS, they randomly reassigned them to species present in other plots (or to itself again, in which case it was considered a new species) (88). This resulted in plots becoming more similar then observed as all plots drew the names for the indets from a panmictic species pool. Omitting UMS results in lower similarities, rather than higher but with smaller deviation (cf. Fig. 1 from (88))

When UMS are omitted, a risk is introduced of underestimating the actual geographic range of the species, e.g. when these UMS are located at the range margins. It would then be expected that this would greatly influence the agreement in similarity of species composition between IMS only and AMS (85). However, this effect appears to be negligible in terms of determining patterns of tree species turnover, as shown by our extremely high regression coefficients between similarities among plots based on AMS and IMS alone (Fig. 3.3). For the sake of argument there is a slight decrease in the correlation (Mantel r) if only IMS are taken into account in the analysis but this effect arguably does not change the patterns of species turnover. Confidence intervals for the slope of the regression for the comparison of similarity values based on all three used indices showed that with an increasing amount of species identified (i.e. a lower proportion of UMS) as is the case with subsequent increased IMS when comparing the Guyana/Suriname, Ecuador and French Guiana datasets, the linear regression starts to approach a slope of 1. For example, with 77% identification of all species in French Guiana, a confidence interval of 0.979-0.982 shows that the slope of the regression between IMS and AMS similarity values calculated with the Bray Curtis index is extremely close to a slope of 1, indicating that the Bray-Curtis similarity values are nearly equal between the IMS and AMS dataset. This was also true when using the other indices

The similarity matrices are the input for the distance decay in similarity, Mantel test and NMDS. As a result it is obvious to expect that if the similarity matrices are very similar these will also generate very similar results when AMS and IMS are compared. We, however, did not know this a priori and had decided to show all three analyses as primary examples because they are all often used by ecologists. For almost all analyses (except NMDS first axis comparison for Guyana and Ecuador) there was a significant positive deviation from the relation y = x with slope 1, when comparing results of AMS and IMS. Hence, omitting species has a small but significant effect. However, this difference is apparently not enough to distort the actual pattern of species turnover. Results from the Raup-Crick analyses also showed that using both approaches to calculate the distance matrices, i.e. with and without permutation based on frequency dependent probabilities of selecting species to be used for Mantel's r, still yields similar results. There are some limitations to using this method. As it is a presence/absence based non-metric measure, identical samples can have dissimilarities above zero and samples sharing no species can have dissimilarities less than one. Samples sharing rare species in particular appear to be more similar as the probability of sharing these species is lower in comparison with samples sharing more common species and data is always treated as presence/absence. In addition, Lennon et al. (2001) showed that strong local differences (i.e. in adjacent plots) in species richness might have an influence on species similarities when using the Sørensen index (116). But even in the light of these limitations, the results from the similarity analyses indicate that, while leaving out unidentified species might compromise species ranges, it does not seem to affect overall similarity, thus remaining a useful approximation for similarity analyses. Results from the NMDS indeed supported the other analyses. Scores from the first axis of the NMDS were nearly identical between only the IMS and AMS. This was also true for the second axis scores. As regressions between NMDS scores of both the first and second axis showed extreme good regression coefficients (R^2 values all >0.97) it shows that it is in fact possible to omit UMS from datasets without losing large scale patterns as are analysed when using NMDS. If a strong underlying gradient, for instance due to different forest types, would be responsible for the robustness of patterns, they could be maintained if a large enough fraction of plots in each forest type is still present after omitting UMS. Table 3.1 shows a summary of the datasets used and the types of forest incorporated in the analyses and although five different types of forest Igapó (IG), Podzol (PZ), Swamp (SW), Terra Firme (TF) and Várzea (VA) were used, the far majority of plots is on Terra Firme soils suggesting forest types are not likely the reason for maintaining these patterns.

Common species dominate ecological patterns. Even when simulating a larger proportion of the complete dataset to be unknown, the majority of analyses still yielded very comparable results. Considering this simulated loss of information, this suggests that patterns of species diversity and composition are robust enough to emerge from (very) limited datasets. Most likely this is due to the fact that common species are common enough to even have a pattern, whereas rare species are often so restricted they do not affect the large-scale patterns much. Lennon et al. (2004) already showed that the more common species were mostly responsible for richness patterns in avian species (117). It would appear that in tropical tree species the common species also dominate major ecological patterns, such as species turnover. Even when using the Jaccard index for similarity, which is only based on presence or absence, results from the similarity analyses showed that omitting UMS made no difference in the overall result (although deviation from the relationship y = x was significant). If IMS consist mostly of common species, this common species-domination as explained above would explain why using only IMS results in the same patterns as when using AMS. To test this we plotted a rank abundance curve on a logarithmic scale. It becomes immediately apparent (Fig. 3.2) that the AMS dataset include many more rare species than did the IMS subset. In fact, omitting the UMS from the dataset results in the rank abundance curve showing a lognormal distribution instead of the logseries-distribution when AMS are plotted. In a sense, omitting UMS truncates the datasets from the right, cutting of the rare species. This also explains why our results for Fisher's alpha showed an underestimation when using only IMS and why similarities between plots using just IMS and AMS deviate with increasing similarity. UMS are not randomly distributed among the common and rare species but are mostly rare species. Hence, FA calculated with N and S for just the IMS will generally be an underestimate.

Conclusions

Finding near identical similarities of species composition and patterns from NMDS results suggest that patterns of similarity and thus composition are robust. Although Fishers alpha based on IMS or AMS showed nearly identical spatial patterns, using a dataset with AMS is still preferred, as FA is not based on comparison and will be underestimated when using only IMS. Overall, the results presented here suggest that irrespective of metrics used, analyses and their limitations; strong ecological patterns still arise using only IMS. In other words, having a large number of unidentified species in a dataset may not affect our conclusions as much as is often thought. However, this should not be interpreted as an argument to omit all UMS all the time. They remain important as they may represent important species (118) and are essential for the calculation of correct diversity measures.

Acknowledgements

We would like to thank Olaf Bánki for his help with collecting and harmonizing the Guyana/Suriname dataset and all others who have provided valuable suggestions, comments and feedback on the manuscript.

Supporting Information

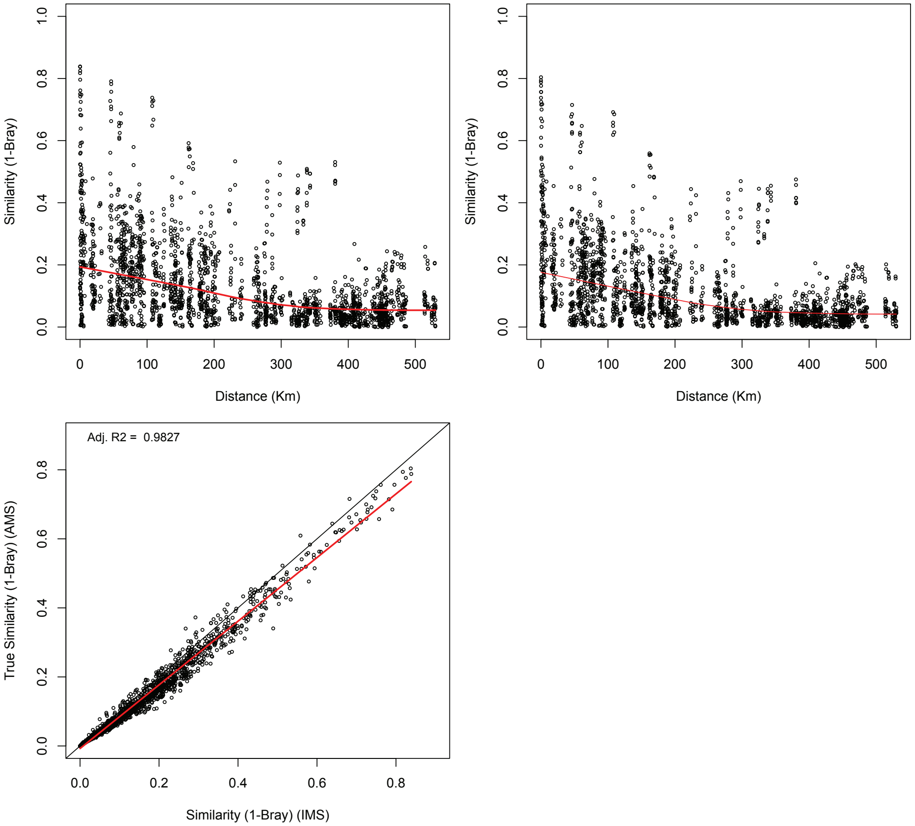


Figure S1. Example showing the distance decay in similarity (DDS) for the Guyana/Suriname dataset based on the distance matrices calculated with the Bray-Curtis index used for the Mantel statistic. Analysis of DDS are shown for only IMS (upper left), AMS (upper right) and the linear regression for Guyana/Suriname (lower left)

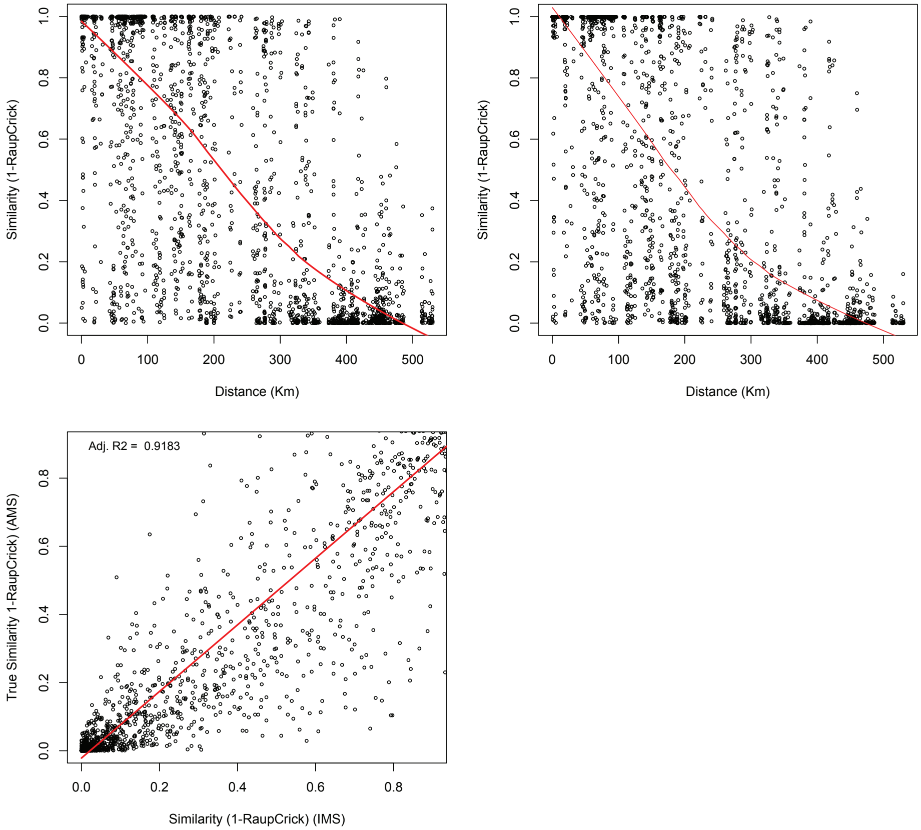


Figure S2. Example showing the distance decay in similarity (DDS) for the Guyana/Suriname dataset using the Raup-Crick analyses. Analysis of DDS are shown for only IMS (upper left), AMS (upper right) and the linear regression for Guyana/Suriname (lower left).

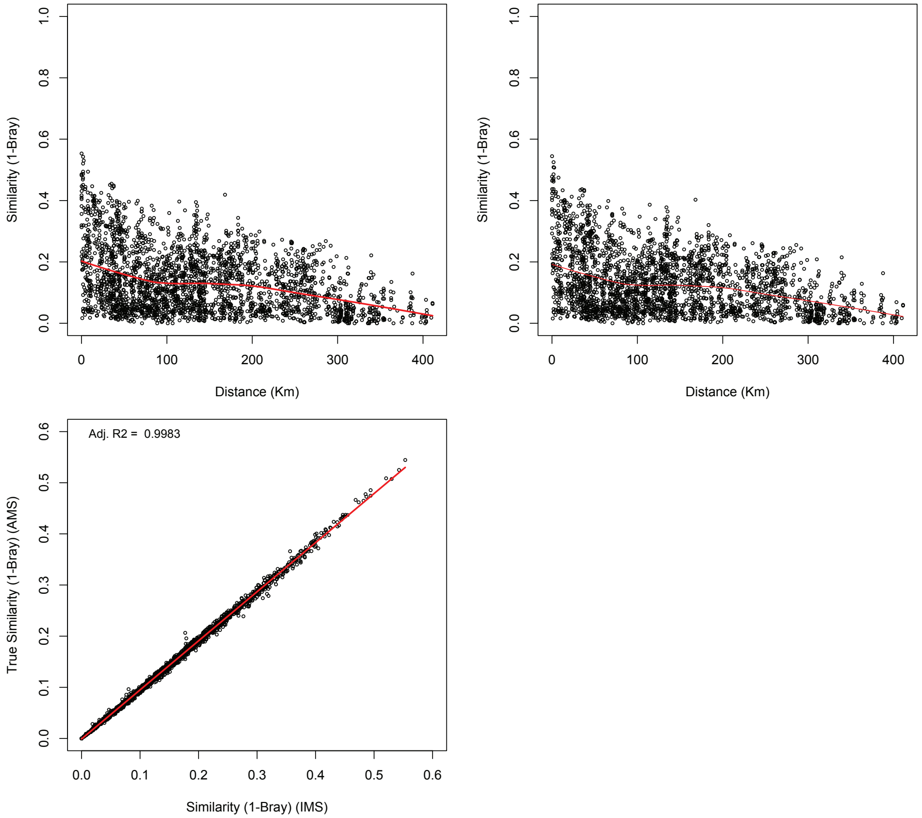


Figure S3. Example showing the distance decay in similarity (DDS) for the Ecuador dataset based on the distance matrices calculated with the Bray-Curtis index used for the Mantel statistic. Analysis of DDS are shown for only IMS (upper left), AMS (upper right) and the linear regression for Ecuador (lower left).

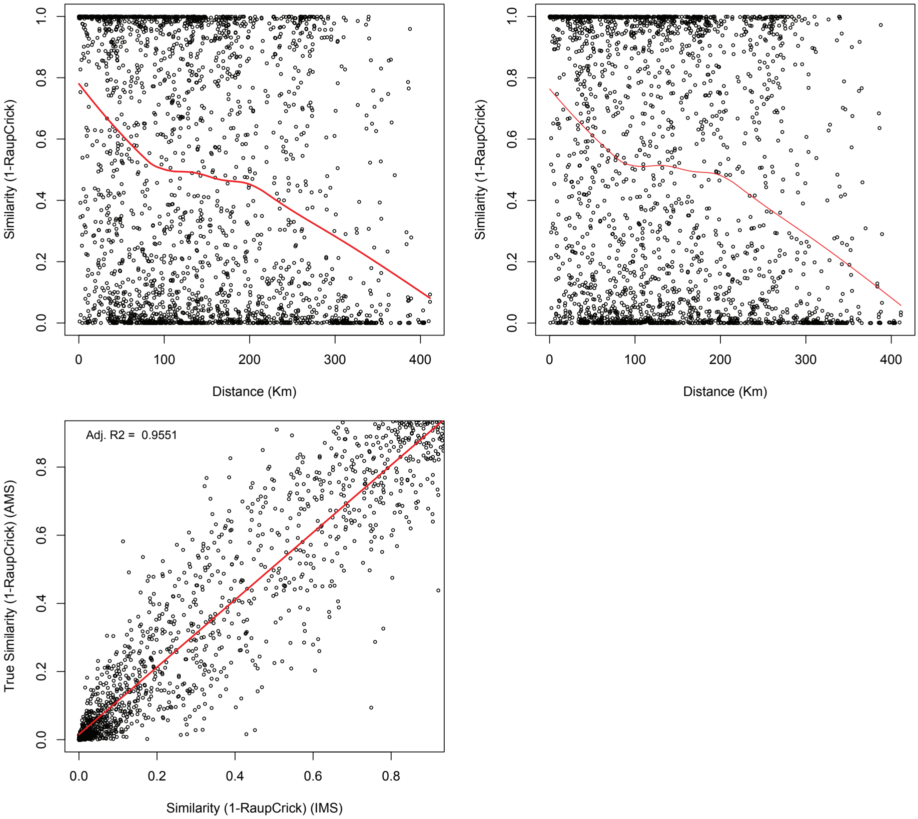


Figure S4. Example showing the distance decay in similarity (DDS) for the Ecuador dataset using the Raup-Crick analyses. Analysis of DDS are shown for only IMS (upper left), AMS (upper right) and the linear regression for Ecuador (lower left).

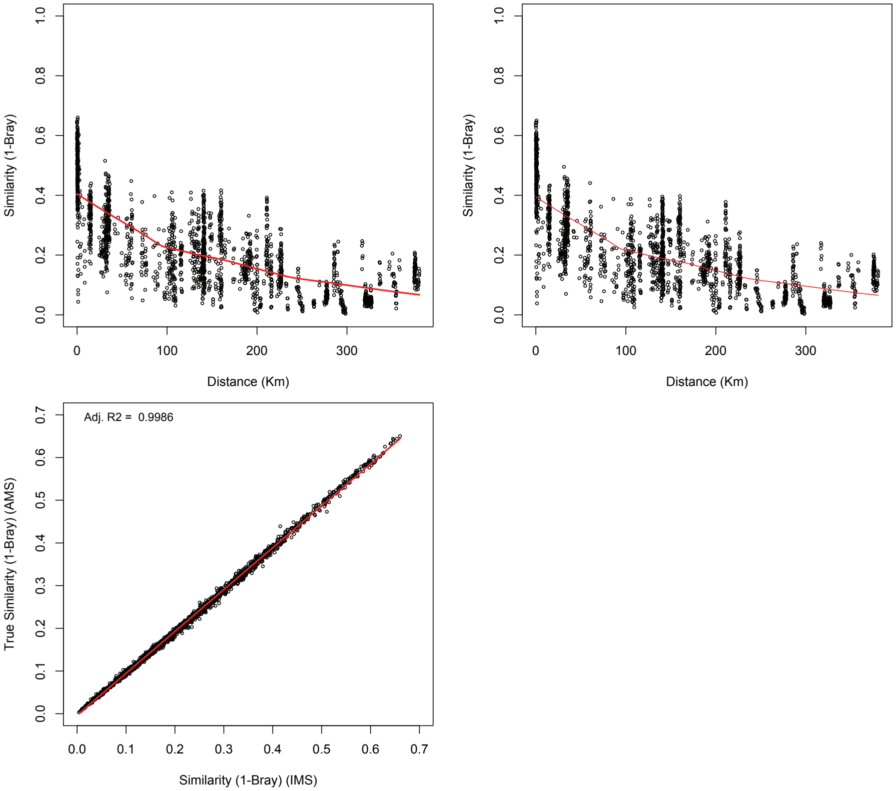


Figure S5. Example showing the distance decay in similarity (DDS) for the French Guiana dataset based on the distance matrices calculated with the Bray-Curtis index used for the Mantel statistic. Analysis of DDS are shown for only IMS (upper left), AMS (upper right) and the linear regression for French Guiana (lower left).

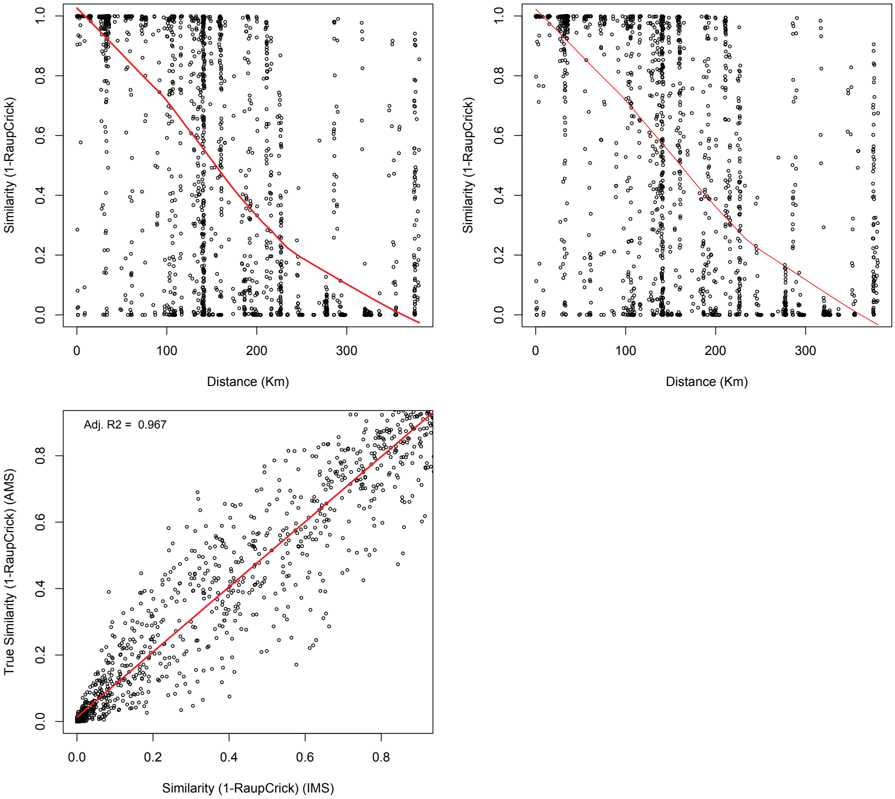


Figure S6. Example showing the distance decay in similarity (DDS) for the French Guiana dataset using the Raup-Crick analyses. Analysis of DDS are shown for only IMS (upper left), AMS (upper right) and the linear regression for French Guiana (lower left).

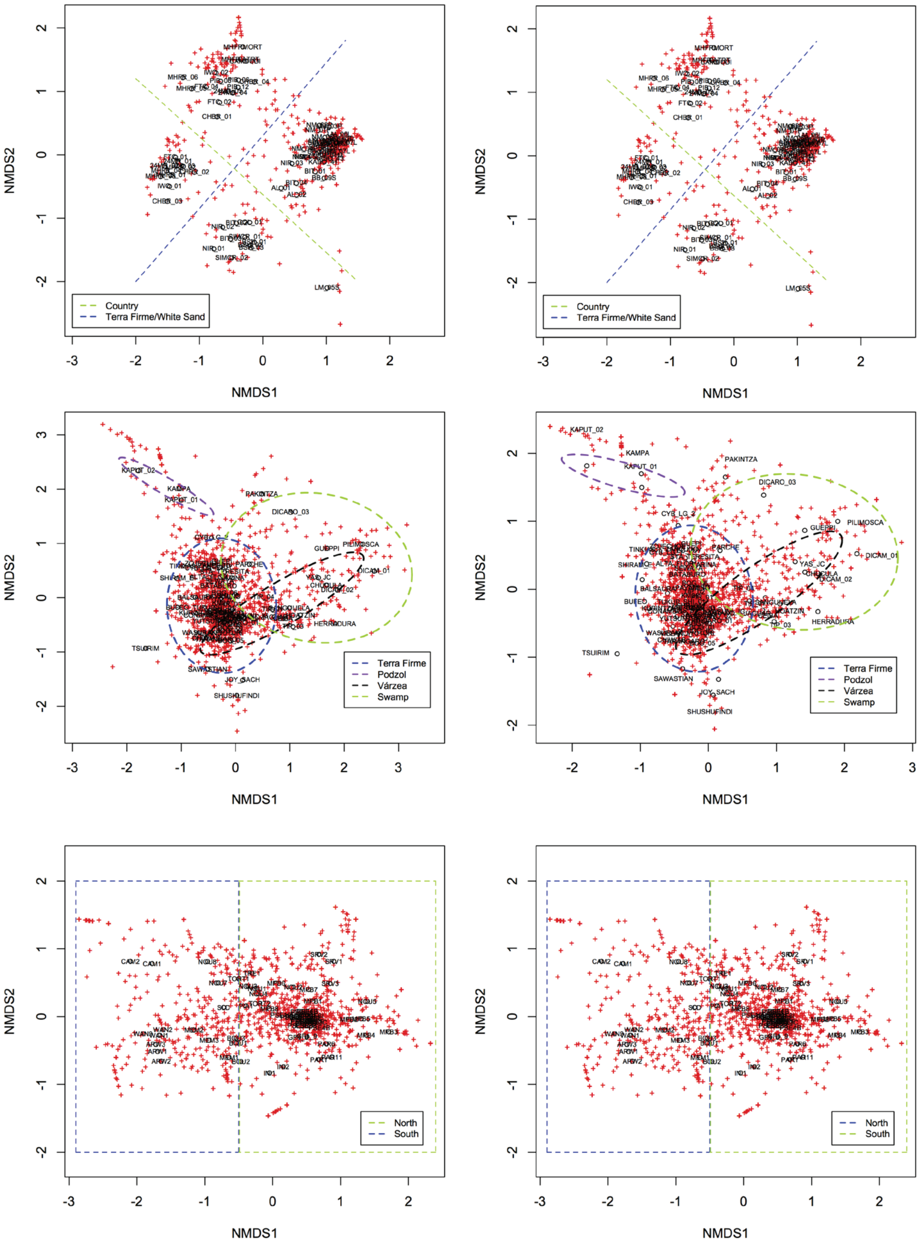
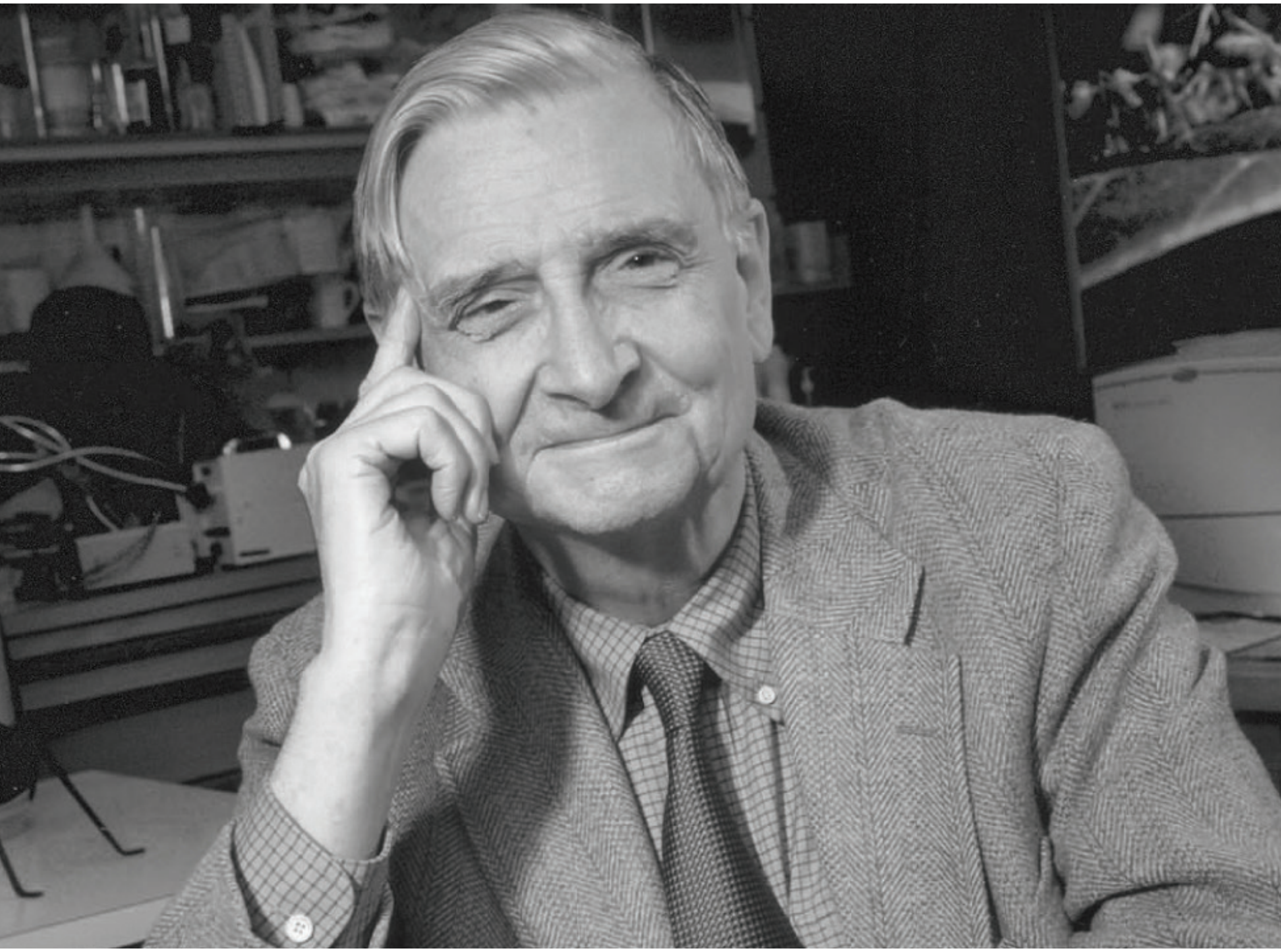


Figure S7. Example showing the Non Metric Muldimensional Scaling (NMDS) ordination procedure for Guyana/Suriname (upper), Ecuador (middle) and French Guiana (lower) using meta-MDS. Analyses are shown for only IMS (left) and AMS (right). Dashed lines indicate different grouping based on country (Guyana/Suriname), forest type or geographic subdivision (North/South).



No one knows the diversity in the world, not even to the nearest order of magnitude. We don't know for sure how many species there are, where they can be found or how fast they're disappearing. It's like having astronomy without knowing where the stars are.

Edward. O. Wilson, Time Magazine (13 Oct 1986)

Chapter Four

Estimating species richness in hyper-diverse large tree communities

(Published in Ecology 98.5 (2017): 1444-1454)

Hans ter Steege[1,2,3]*, Daniel Sabatier*[4]*, Sylvia Mota de Oliveira*[1]*, William E. Magnusson*[5]*, Jean-François Molino*[4]*, Vitor F. Gomes*[2]*, Edwin T. Pos*[1,6]*, Carol V. Castilho*[7]*, Rafael P. Salomão*[2]

[1]*Naturalis Biodiversity Center, Leiden, The Netherlands*
[2]*Museu Paraense Emílio Goeldi, Belém, PA, Brazil*
[3]*Systems Ecology, Free University Amsterdam, The Netherlands*
[4]*Institut de Recherche pour le Développement, UMR AMAP, Fonctionnement et Évolution des Plantes, Montpellier, France*
[5]*Instituto Nacional de Pesquisas da Amazônia, Manaus, AM, Brazil*
[6]*Ecology & Biodiversity group, Department of Biology, Utrecht University, The Netherlands*
[7]*Embrapa Roraima, Boa Vista, RR, Brazil*

Abstract

Species richness estimation is one of the most widely used analyses carried out by ecologists, and nonparametric estimators are probably the most used techniques to carry out such estimations. We tested the assumptions and results of nonparametric estimators and those of a logseries approach to species richness estimation for simulated tropical forests and five datasets from the field. We conclude that nonparametric estimators are not suitable to estimate species richness in tropical forests, where sampling intensity is usually low and richness is high, because the assumptions of the methods do not meet the sampling strategy used in most studies. The logseries, while also requiring substantial sampling, is much more effective in estimating species richness than commonly used nonparametric estimators, and its assumptions better match the way field data is being collected.

Key words: *species estimation, nonparametric-estimators, logseries, tropical forests*

Introduction

Species-richness estimation is one of the most widely used analyses carried out by ecologists, either to compare samples obtained with different efforts, or by extrapolation, to predict the number of species present in an area larger than the one sampled. Extrapolation methods are frequently used for geographically large areas, where coverage of the complete range is out of reach, too labor intensive, or too expensive.

Parametric species richness estimation is based on parameter inference for either one of the two main relationships describing assemblages: the number of individuals (*N*) in a community or the area (*A*) this community occupies. In these cases, the number of species (*S*) only depends on the relative or Rank Abundance Distribution (RAD) of the species (119) or the Species-Area Relationship (SAR) (120). As a general rule of thumb, in any number of random samples of an area, the number of species that remain undetected will increase with increased *S* and *A* (121), precluding any attempt to directly quantify the RAD or the SAR from samples. This clearly poses a problem in tropical forests that are generally both large and rich. There has been a long argument as to whether the logseries (98), the log-normal (122), or alternative distributions (123) give the best fit for RADs, how much the fit is dependent on scale or sampling completeness, and to which extent the best fitting model reflects the biological processes underlying the distribution. The use of nonparametric estimators of species richness such as Chao, ICE (Incidence-based Coverage Estimator of species richness), and Jackknifing, has been proposed as a way of dealing with this uncertainty, because they do not assume any underlying distribution. It would be wrong, however, to suppose that they are less sensitive to other assumptions than parametric methods or that they do not suffer from other drawbacks. Brose et al. (124) noted that sampling-theoretical methods of estimation require high sampling intensity to avoid what Wang and Linday (125) call the "*severe under-estimation observed from popular nonparametric estimators due to the interplay of inadequate sampling effort, large heterogeneity and skewness.*" Xu et al. (126) also reported that nonparametric methods severely underestimate richness and emphasized that these methods should not be used across heterogeneous landscapes. This is largely because nonparametric estimators based on a sampling estimate of the rare-tail of the SAR are very sensitive to the shape of the abundance-distribution. As underlined by Harte and Kitzes (127), "*The rare tail is emphasized because the shape of the species-area relationship is especially influenced by the numbers of rare species*". Although the performance of estimators has been frequently compared (124, 126, 128–130), much less of the ecological literature critically evaluates their assumptions and caveats.

Perhaps the most commonly used estimator for species richness is the Chao1 nonparametric estimator (131, 132), which estimates the number of species as: $S_{estimated} = S_{observed} + f_1^2/(2f_2)$, where f_1 is the number of species with 1 individual in the sample (singletons) and f_2 is the number of species with 2 individuals in the sample (doubletons). The Chao1 estimator and other nonparametric estimators make no assumptions about the underlying species-abundance distribution, but do assume that sampling is random with replacement across the whole area. When $f_1 = 0$, it is assumed that all species have been collected and $S_{estimated} = S_{observed}$ (132).
Chao Bunge (133), Chao Lee ACE, Chao Lee ACEI (134), and Jackknife (135) are variations on the original Chao 1 estimator. They are also dependent on the fractions of the rare or infrequent species, and require "*a sufficiently high overlap fraction [..] to produce a reliable estimate of the species*" (133), and, finally, are all based on the capture-recapture principle that requires sampling with replacement.

The logseries in contrast is not based on a capture-recapture principle and was among the first attempts to mathematically describe the relationship between the number of species and number of individuals in a biological context by Fisher (98) and is given by: $\Phi_n = \alpha x^n/n$, where: Φ_n is the number of species with n individuals; α is Fisher's α; $x = N/(N + \alpha)$ (N being the number of individuals in the total sample; x being asymptotically equal to 1 with large sample sizes). Hence, we expect α from samples to quickly approach α of the total landscape, after which it will be practically independent of sample size. Fisher's alpha can be calculated from the number of individuals (N) and species (S) in a sample by iteratively solving: $\alpha = S/ln(1 + N/\alpha)$. The logseries is essentially a geometric summation, which builds up from the first term (Φ_1), the singletons. The number of singletons is thus predictable in a logseries ($\Phi_1 = \alpha x$) and always the largest class. As x is very close to 1 for reasonably large samples, $\Phi_1 \approx \alpha$ in such samples. Similarly, the number of doubletons is: $\Phi_2 = \alpha\, x^2/2 \approx \alpha/2$. When we assume that RAD's of communities follow the logseries, this has implications for the nonparametric Chao1 estimator. For large samples, the Chao1 estimator (note that $f_1^2/[2f_2] = \Phi_1^2/[2\,\Phi_2]$) will simply become: $S_{estimated} = S_{observed} + \alpha^2/[2(\alpha/2)] = S_{observed} + \alpha$. Consequently, we predict that for reasonably large samples, for which α is constant, Chao1 always estimates the number of unseen species as α, regardless of the size of the samples.

Hubbell's neutral theory was the first ecological theory deriving the logseries from the basic biological processes of birth rate (b) and death rate (d) (29, 136). It can be shown that in this model $x\ (N/[N + \alpha]) = b/d$. NT derives a distribution, the Zero Sum Multinomial (ZSM), which for large communities with little drift approaches a logseries. For small local communities (limited immigration and drift), the ZSM approaches a lognormal (29).

Here we compare commonly used nonparametric estimators of species richness to one parametric estimator based on the logseries for the purpose of estimating species richness in large areas of tropical forest. We specifically chose the logseries as we are trying to estimate richness in very large areas where the ZSM approaches this distribution. We show by simulations and comparisons with empirical data that the assumptions of the parametric estimator are less sensitive to deviations than those of the nonparametric estimators.

Methods

Simulations. We modelled forest communities of 100 x 100 1-ha plots (a 100 km^2 square area), each plot with 500 individuals. We initially filled each of the 10,000 hectares with a random sample of 500 individuals from a metacommunity (MC). The MC was constructed using a logseries of 15 million individuals and a Fisher's α of 300, which is roughly equivalent to a rich central Amazonian rainforest (see Field data). We used a logseries as this conforms to the structure expected (29) and found in tropical forests (14, 136, 137). After filling the plots randomly from the MC, the mean Fisher's α of all plots and that of the virtual forest initially is, as expected, equivalent to that of the MC. During the simulations, trees were randomly selected to be removed (1 per plot per time step) and new recruitment could come from dispersal (m) from 4 sources:

1) Recruitment from dispersal inside the plot (m_{plot}), equivalent to local recruitment. Local recruitment is random within the plot, i.e. we assume no spatial structure inside the plots.

2) Recruitment from dispersal from the surrounding eight plots. Dispersal probability based on dispersal distance was based on the model of Chisholm and Lichstein (138), modified by Pos et al. (139). The dispersal probability from the adjacent plots ($\mathrm{m}_{\mathrm{adjacent}}$) is computed from dispersal distance as (139):
$m_{adjacent} = 0.3 * (A - (l - 2*d)^2)/A$. Where: A is the area of the plot (10,000 m^2), l = length of the plot (100 m), and d = the average dispersal distance. Assuming an average dispersal range of 10-40 meters $m_{adjacent}$ is in the range of 0.108-0.288.

3) Recruitment from dispersal from the surrounding forest (10,000 ha), comparable to long-distance dispersal. Individuals for replacement were drawn randomly from the 10,000 ha. This assumes that long-distance dispersal is not spatially driven. We used a probability of $m_{forest} = 0.1*m_{adjacent}$.

Table 4.1 Botanical inventories used for the analysis. Locations are Barro Colorado Island (BCI), Reserva Ducke (RD), Piste de St Elie (PSE), Monte Branco Plateau (MBP). Variables are number of plots sampled, plot area (ha), number of individuals sampled (N), number of species recorded (S), the target area for which estimates were made, number of individuals in the target area based on average density, and reference to the data source: 1) (35); 2) (140); 3) (141); 4) (142).

Locality	# plots	Plot area	N	S	target area	target individuals	Reference
BCI	50	1	21,457	225	50 ha	21,457	1
RD	72	0.5	25,066	1233	100 km^2	7,200,000	2
PSE	20	1	12,450	574	1500 ha	933,750	3
MBP	301	0.25	36,546	703	3750 ha	1,821,229	4

Table 4.2 Species estimates based on plot samples in BCI, RD, PSE, and PMB

	BCI	se	RD	se	PSE	se	PMB	se
Number of plots	50		72		20		301	
number of individuals	21,457		25,066		12,450		36,546	
number of species	225		1233		574		703	
target area	50 ha		100 km^2		1500 ha		3750 ha	
target individuals	21,457		6,960,000		933,750		1,821,229	
S$_{\text{estimated}}$ with								
Fisher's α	225		2759		1110		1185	
Chao 1984	239	8.3	1,408	32	724	36	821	31
Chao Bunge	243	9.6	1,423	32	715	34	823	31
Chao Lee ACE	238	6.1	1,375	20	669	18	738	16
Chao Lee ACEI	241	8	1,405	26	694	25	805	23
Jackknife	244	6.1	1,591	59	1066	124	920	40

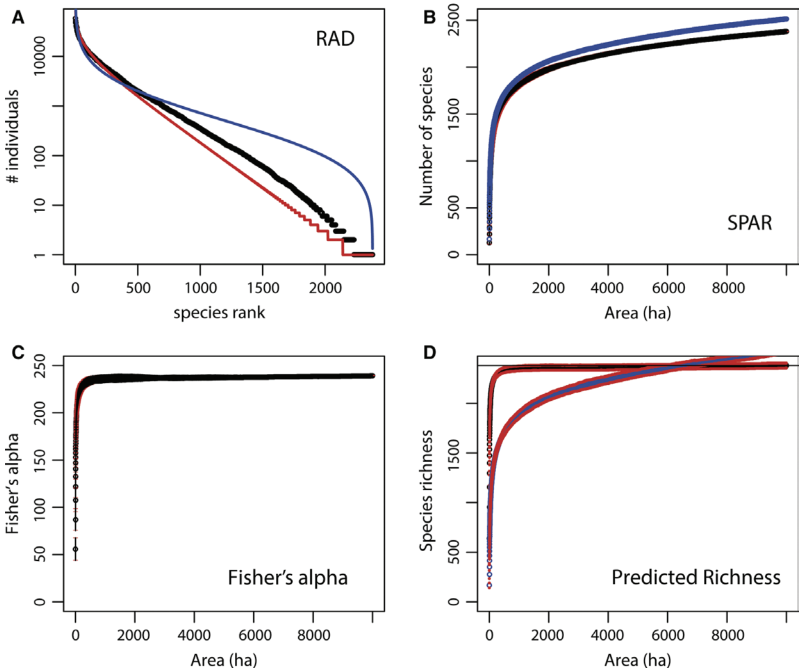


Fig. 4.1 Simulation of a 10,000-ha virtual forest with mean dispersal distance of 20 m. Parameters used are m_{plot} = 0.78688; $m_{adjacent}$ = 0.192; m_{forest} = 0.0192; m_{MC} = 0.00192; v = 10 4. (A) Rank abundance distribution (RAD) of the total virtual (black) with logseries fit (red) and lognormal fit (blue). (B) Species area (SPAR) curve for the total virtual forest and estimated richness (Sestimated) based on Chao1 (blue). (C) Fisher's α area curve for the virtual forest. (D) Species richness estimated with Fisher's α (black), Chao1 (blue), each with 95% CI (red), and actual species richness of the simulated community (horizontal line). m_{plot} = local recruitment; $m_{adjacent}$ = recruitment from adjacent plots; m_{forest} = recruitment from total forest; m_{MC} = recruitment from metacommunity; v = speciation.

4) Recruitment from dispersal from the MC, this is comparable to infrequent very long-distance dispersal, also termed vagrancy. The individuals were drawn randomly from the MC, assuming that very long-distance dispersal too is not spatially driven. We used a probability of $m_{MC} = 0.01*m_{adjacent}$.

5) Speciation (v) as defined in the Unified Neutral Theory of Biodiversity and Biogeography (29): $v = \theta/(2*J) = 250 /(2*10,000*500) = 2.5e-5$. Where θ is the biodiversity number, asymptotically equivalent to Fisher's alpha and J is the size of the community.

Parameters 2-4 were calculated first. Local recruitment (1) was then calculated as: $m_{plot} = 1 - m_{adjacent} - m_{forest} - m_{MC} – v$. We ran 30,000 time steps for each model with mean dispersal distances of 10, 15, 20, 25, 30, and 40 m.

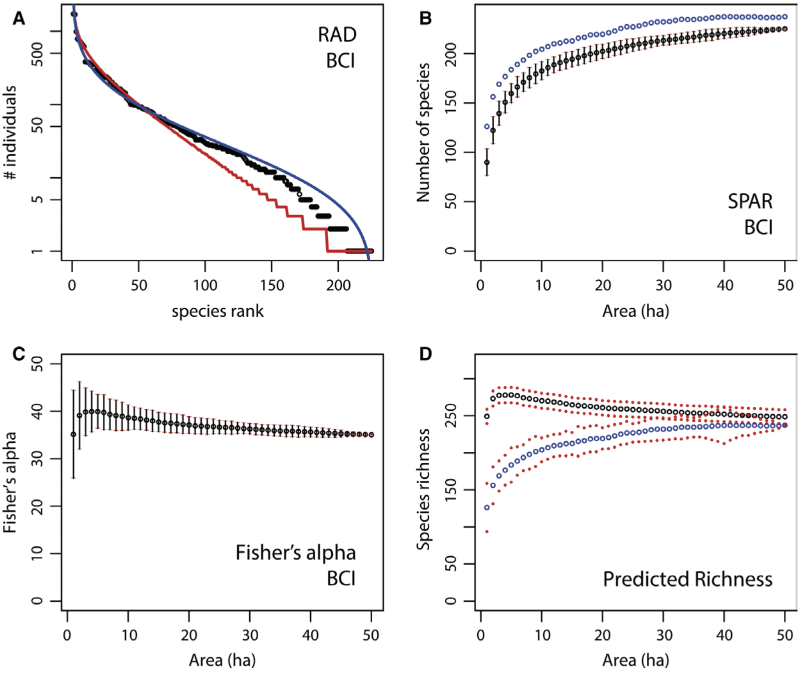


Fig. 4.2 Barro Colorado Island field data (BCI). (A) Rank abundance distribution (RAD) of BCI with logseries fit (red) and log- normal fit (blue). (B) Species area curve for BCI and estimated richness ($S_{estimated}$) based on Chao1 (blue). (C) Fisher's α area curve for BCI. (D) Species richness estimated for a 100-ha area on BCI with Fisher's α (black) and Chao1 (blue), each with 95% CI (red).

At each time step, 1 individual per plot was randomly selected to be replaced by another individual based on the 5 probabilities above. Thus, 10,000 individuals were replaced at each time step. After each simulation, we plotted the RAD with a fit of the logseries and lognormal, the Species Area Curve with Chao1 estimator, the Fisher's α to area curve, and the predicted richness based on Fisher's α and the Chao1 estimator. All curves were based on the average of 50 draws from 1 to all 10,000 plots. We also plotted the results for the average of 50 random draws of 100 plots from our virtual forest. In addition, we also ran the simulation model for a sample of 49 ha of forest (7x7 ha), using the field data of BCI (Table 4.1). We simulated a forest area of 49 plots, using a MC of 15 km2 (the size of BCI), an alpha of 50 and density of 429 ind ha-1, a dispersal distance of 40 m (138) for madj = 0.288, and ν = 0.00119. Simulations and calculations were carried out with custom-made scripts in R (51).

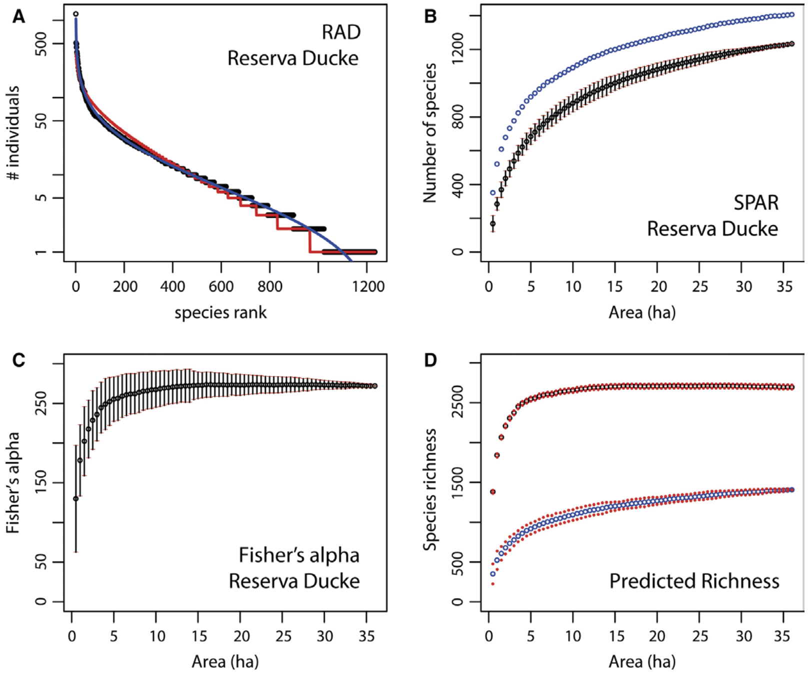


Fig. 4.3 Reserva Ducke field data (RD). (A) Rank abundance distribution (RAD) of RD with logseries fit (red) and lognormal fit (blue). (B) Species area curve for RD and estimated richness ($S_{estimated}$) based on Chao1 (blue). (C) Fisher's α area curve for RD. (D) Species richness estimated for the total 100-km2 RD area with Fisher's α (black) and Chao1 (blue), each with 95% CI (red).

Field data. We used field data from 4 sites. 1) Barro Colorado Island (BCI), a 50-ha plot in old growth forest (35). This well-known dataset was also used in Chao et al. (132); 2) Reserva Ducke (RD; Supporting Information (SI) Chapter 4 S1: Fig. S1), a forest reserve of 100 km^2 in central Amazonia, just north of Manaus (140); 3) Piste de St Elie (PSE, SI Chapter 4 S1: Fig. S2), mixed forest in northern French Guiana (141); 4) the Monte Branco Plateau (MBP, SI Chapter 4 S1: Fig. S3), a large bauxite plateau of 3750 ha in Para, Brazil (142). BCI tree data was extracted from vegan (143), tree data for RD and PSE are integrated in the ATDN database (14) and extracted from that source, MBP tree data (R.P. Salomão, unpublished data) was taxonomically harmonized with the ATDN database.

We extrapolated the species richness for an area in which the plots were located; for RD for 7.2 million individuals (the area of the full 100 km^2 reserve); for PSE an imaginary 1500 ha forest area encompassing the plots; for MBP the 3750 ha that comprises the complete plateau (Table 4.1). The plots are well spread across these areas. For BCI we estimated richness for the 50-ha plot. For each of the plot datasets we carried out the following analyses:

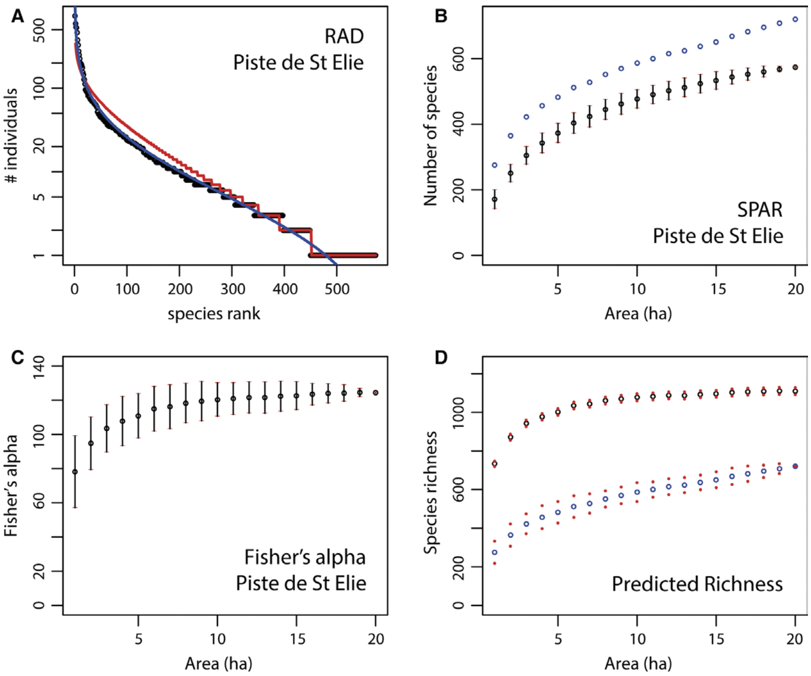


Fig. 4.4 Piste de Saint Elie field data (PSE). (A) Rank abundance distribution (RAD) of RD with logseries fit (red) and lognor- mal fit (blue). (B) Species area curve for RD and estimated richness ($S_{estimated}$) based on Chao1 (blue). (C) Fisher's α area curve for RD. (D) Species richness estimated for the total 15 km2 area surrounding the plots with Fisher's α (black) and Chao1 (blue), each with 95% CI (red).

1. Plotted the RAD of the dataset with the exact logseries and lognormal for the number of individuals (*N*) and species (*S*) in the field sample;

2. Constructed a curve of the mean species richness by area, based on 50 randomizations of the field data;

3. Constructed a curve of the mean of Fisher's α by area, based on the same 50 randomizations of the field data;

4. Estimated species richness in the target area for all sub-samples of the 50 randomizations based on Fisher's α of the sub-samples as follows: $S = \alpha * \ln(1 + N/\alpha)$ (98); where α = Fisher's α, and N is the number of trees in the subsample and the variance of S as (98): $var_S = \alpha \ln([2N + \alpha]/[N + \alpha]) – \alpha^2N/(N + \alpha)^2$;

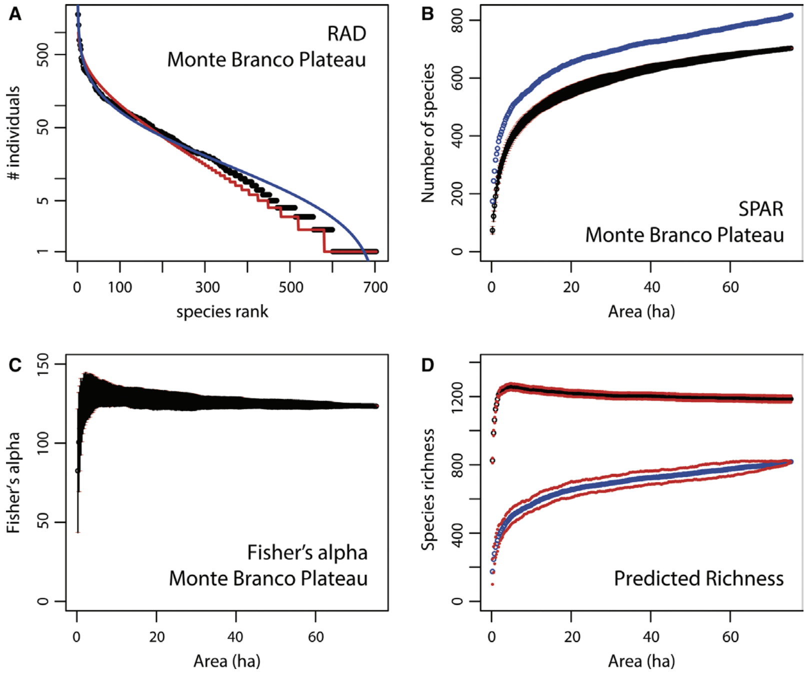


Fig. 4.5 Monte Branco Plateau field data (MBP). (A) Rank abundance distribution (RAD) of MBP with logseries fit (red) and lognormal fit (blue). (B) Species area curve for MBP and estimated richness ($S_{estimated}$) based on Chao1 (blue). (C) Fisher's α area curve for MBP. (D) Species richness estimated for the total 37.5 km2 MBP area with Fisher's α (black) and Chao1 (blue), each with 95% CI (red).

5. Estimated species richness in the target area for all sub-samples of the 50 randomizations, based on Chao1: $S_{est} = S_{obs} + f1^2/(2f_2)$;

6. Estimated the species richness for the field dataset for a number of nonparametric estimators (Chao 1989, Chao Bunge, Chao Lee ACE, Chao Lee ACEI, Jackknife), as provided in the R-package SPECIES (144).

The 50 randomizations of the plot data were produced without replacement from one plot to the number of plots in the field dataset.

Results

Simulations. The simulations of our virtual forest with mean dispersal distance of 20 m produced an RAD that is close to a logseries (but not fully identical) (Fig. 4.1A). Species richness calculated with the Chao1 estimator as predicted becomes $S_{observed}$ plus ~ Fisher's α for larger samples (Fig. 4.1B). While Fisher's α and species richness calculated with Fisher's α tend to asymptotically approach the community value, species richness calculated with the Chao1 estimator follows the shape of the species area curve and finally overestimates the richness of the total sample by approximately Fisher's α.

All simulations (*d* = 10 - 40 m) show similar results (SI Chapter 4 S1: Figs. S4, S6, S8, S10, S12, S14, S16; Data S1: SPAR samples.csv). With increasing mean dispersal distance and, hence, stronger input from the adjacent plots, Fisher's α tends to be overestimated slightly before it reaches the value of the total virtual forest and the number of species in the full virtual forest increases from 2071 to 2098. The calculations for 50 samples of 100 plots suggest that although Fisher's α predicts a richness closer to the known richness for the virtual forest, it is still an underestimate of 3-17% (SI Chapter 4 S1: Figs S5, S7, S9, S11, S13, S15; Data S1: sample by nr. of plots.csv). For a similar sample size, the Chao1 estimator provides an underestimate of 43-51%, depending on the dispersal distance chosen (SI Chapter 4: Data S1).

Simulations of 49 ha of BCI. Simulations of a 49 ha virtual plot based on the BCI data produced a RAD (SI Chapter 4 S1: Fig. S18) very similar to that of the forest in the real 50 ha BCI plot (Fig. 4.2). Fisher's α was very close to the final value for the simulated forest after 10 plots. Consequently, species richness was also close to its simulated richness after sampling 10 plots. Species richness calculated with Chao1 is, as predicted, the species area curve plus Fisher's α of the sample. Thus, even when all individuals have been sampled, Chao1 still predicts unobserved species with a magnitude of Fisher's α. This is because, as in real forests, the virtual forest of 49 ha still contains singletons.

Field data. In all cases: BCI (Fig. 4.2), RD (Fig. 4.3), PSE (Fig. 4.4), and MBP (Fig. 4.5), the RAD showed a hollow curve with few common and many rare species and, except for BCI, the logseries provided a reasonable fit. In all cases, Fisher's α was very close to that of the full sample with less than 20 plots sampled. For small samples, Chao1 provided a severe underestimate for the richness in the sample, and even for the final sample, $S_{estimated}$ was almost equivalent to $S_{observed}$ + Fisher's α.

Species estimates for the target area made with Fisher's α were much larger than those made with the asymptotic Chao1 estimator, which were close to $S_{observed}$ + Fisher's α of the measured data (Figs. 4.2-4.5). All other nonparametric estimators, too, predict much lower values for richness, comparable to the Chao1 estimator (Table 4.2). Only for BCI, where the area for which richness was to be estimated was similar to the actual sample, did the nonparametric estimators approach the estimate based on Fisher's α. For the BCI and MBP data, and simulations with higher mean dispersal distances, Fisher's α peaked before it levelled off to its final value similar to the simulations, i.e. it showed a hump (see Figs. 4.2, 4.5). Fisher's α, however, rose regularly for PSE, RD and for simulations with lower mean dispersal distances (Figs. 4.1, 4.3, 4.4).

Discussion

Based on our simulations with a spatially semi-explicit model, Fisher's α provides a more accurate prediction of species richness in the virtual forest communities than does the nonparametric Chao1 and other nonparametric methods, especially if sample intensity is low. We believe that the failure of nonparametric methods to estimate diversity is mainly due to the resampling approach with its need of high sampling effort and its expected loss of singletons, and the lack of definition of the target area. We elaborate on this below.

Based on resampling the BCI plot data, Chao et al. (132) found that, to detect 90% of the species, a median sample size of 80% of the area is necessary. Also Chiarucci et al. (128), using modeled vegetation, found that nonparametric estimators need at least 15-30% of the area to be sampled for reasonable estimates of the species richness of the whole area. Using these methods with low sampling effort leads to serious underestimation as Brose et al. (124) and our models clearly show. In real life, even though trees are not removed by our sampling (and resampling is thus statistically possible), the chances of resampling the same plot are negligible. In the Amazon with a sample of 1170 1-ha plots in an area of over 5 million km^2 (14), that chance would be just 2·10-9. At the intensities at which tropical forests are sampled (0.0002% for the Amazon) nonparametric methods simply cannot accurately estimate the number of species in the whole area. On top of that, when plot locations are known researchers are unlikely to resample a known area. Also, when locations of previous studies are known, researchers are unlikely to resample a plot.

With capture and recapture techniques and the nonparametric estimators tested, sampling is considered complete when no singletons exist anymore in the data (132).

In tree plots, the disappearance of singletons would be the result of sampling the data many times over with replacement (132). This resampling results in the estimated richness asymptotically approaching true richness when the number of singletons is zero, as the total number of species cannot be larger than those observed in the total dataset (132). We argued above that in the case of research in tropical forests, plots are probably never sampled with replacement.

Thus, the number of species is expected to increase with sample size as predicted by the '*First Law of Biodiversity*' ((120), "*larger samples yield more species*") and many other theories of Biodiversity (29, 53, 64, 80, 145). In addition, singletons will remain (often close in number to Fisher's alpha). In the above theories singletons are the representatives of the biological processes of immigration, extinction or speciation. Singletons might be species on their proverbial way out driven by extinction or new species coming in by speciation or migration. The latter are hence necessary to maintain richness. Without these processes fixation will occur due to ecological drift, analogous to genetic drift from population genetics. Thus, when sampling without replacement: the lack of singletons in these systems would suggest incomplete rather than complete sampling. This inconsistency can be extracted from the description of the method itself, where authors mention that "*given adequate sampling, lack of singletons indicates adequate sampling*" (132).

Finally, as most tropical tree field data conforms to the logseries (see references in Introduction), the Chao1 index becomes scale invariant, always estimating the same number of missing species, in the case of Chao1, to exactly the amount of Fisher's α. This was shown mathematically in the introduction for Chao1 and is supported by our simulations. While we did not show this mathematically for the other nonparametric estimators, they are derived from the same theoretical framework of capture-recapture and estimate similar richness (SI Chapter 4 S1: Fig. S19; Table 4.2) and thus also provide severe underestimates with low sampling intensities.

For the full Amazon area (~5.5 million km^2), ter Steege et al. estimated ~ 16,000 tree species based on a sample of 1170 plots of 1-ha (14). They applied at least 18 different extrapolation methods from software packages SPECIES (144), and CatchAll (146) to their plot data (14). Almost all were rejected, as they predicted the total number of Amazonian tree species to fall in the range 4015-6412, a demonstrably severe underestimation of the true species richness (147). A new estimator, implemented in CatchAll (WLRM_UnTransf) (146, 148) gave an estimated total richness above 11,000, closer to that calculated with their logseries extrapolation, but was not selected by the program as the best estimator. The ACE1_Max Tau estimator gave a result greatly exceeding the estimate with the log-series but its Tau was much higher (9048) than the recommended value (tau < 10). The failure of these models to fit the Amazonian data was not surprising. These estimators performed poorly because at least one of their assumptions, high sampling intensity, was not met - a condition unlikely to be met in any large forested area. Based on an extensive search in several data providers and herbaria, (149) found that nearly 12,000 tree species have actually been collected in Amazonia, with a collecting density as low as 10 collections per 100km^2. They conclude that the estimate of 16,000 is entirely plausible. Importantly, the number of species found is almost twice that estimated with most nonparametric methods.

Using different methods to estimate or extrapolate the SAR, like Maximum Entropy inference (80, 127) or a power law based fitting from multi-scales sampling (150, 151), also showed that regional scale diversity of trees was estimated acceptably from small plots samples. Interestingly, the abundance distribution model arising from the Maximum Entropy approach is most often a logseries (127). Using the logseries is, however, not without assumptions either. Our virtual forest is neutral with regard to the environment, i.e. demographic probabilities for each individual, regardless of species identity, are equal. Hence, in addition, the only cause of aggregation is limited dispersal of individuals but given enough time, even ranges of very dispersal limited species can become large. In real life, species will segregate the environment based on ecological preferences as well. Hence, beta-diversity in real forests is higher than in our virtual-forest stand and a peak of Fisher's α is expected when a large heterogeneous area is sampled over a range of sampling intensities.

BCI is known to have clear segregation of species based on soil moisture (152) and the relationship Fisher's α to area peaks at relatively low number of plots. We also expect the species on MBP to be similarly clumped because of the clear peak in Fisher's α at low sample sizes. At MBP plot size may also influence the peaking of Fisher's α. As the plots are smaller (0.25 ha), the recruitment to the plots will be more affected by the adjacent plots as madjacent is very much dependent on the ratio between the plot boundary and mean dispersal distance (138). The peak modeled and observed can be explained by a relationship between beta- and alpha-diversity. At low migration rates, recruits mostly come from within plots, hence beta-diversity is maximized but alpha-diversity is not because each plot is practically isolated and losing species due to ecological drift. This means that, for just sampling one plot, Fisher's α will be much lower than the average of the whole forest. Continuous sampling, however, will gradually result in the average Fisher's α. There will be no peak because the probability for each plot bringing new species to the whole is the same and thus the increase will be gradual until Fisher's α is equal to that of the virtual forest. When migration increases, however, plots close by exchange more species and beta and local alpha diversities increase simultaneously. In this case, sampling a few plots randomly will likely initially overestimate Fisher's α, because each sample includes new species for the total sample due to the combined higher beta and alpha-diversity, creating a fast rise in Fisher's α. However, continuing the sampling at some point does add more individuals to the total sample, though species will be resampled, lowering Fisher's α again. When dispersal is so high as to be similar across the complete virtual forest, composition would essentially be very similar for all plots with very high local alpha- and low beta- diversities and Fisher's α would not peak but increase fast to its virtual-forest value (as in the virtual 49 ha BCI, Fig. 4.3).

Is estimating species richness still a long way off? Chiarucci (153) suggested that '*estimating species richness is still a long way off!*' Nonparametric estimators underestimate richness (see above and (126)), while area-based estimators tended to overestimate richness (126). Xu et al. (126) concluded that Maximum Entropy greatly overestimated richness. However, their perceived overestimate is based on the richness they expected, which was based on a list of species found in their area. We believe that many of us do not fully comprehend the consequences of the logseries model. One of us was also surprised when we estimated the expected species for RD, which was much more than was expected based on extensive fieldwork for the Flora of the area (154) and ecological fieldwork. However, with an Fisher's α of 271 for the plots of RD, assuming that this is close to the correct Fisher's α for the area, we expect 271 species with only 1 individual, 135 with two individuals, 62 with 3 individuals, 31 with 4 individuals, etc. RD covers 100 km^2, with an average tree density of 696 trees ha-1 (14). That indicates a total of 6.96 million individuals. The chance of finding a singleton species there with feasible sampling intensity is thus very, very small. This is the consequence of using this theoretical framework - see also (136). Because many researchers using nonparametric estimators assume that sampling is complete when the samples contain no singletons, an assumption that does not agree with ecological theory or with most ecological sampling, they are likely to severely underestimate richness when sampling level is low. Therefore, we suggest that the use of nonparametric estimators should be discouraged in studies with low sampling intensity in large remote areas. If the data can reasonably be assumed to follow a logseries, species estimation by means of Fisher's α is likely a better option. Other methods that produce abundance distributions with many singletons, matching most observational data, such as various parametric methods (155) or phenomenological theories, such as Maximum Entropy (80) are probably also good alternatives.

Fisher's paradox. The term Fisher's paradox was coined by Hubbell (136):

> "*The logseries is an infinite series that mathematically goes on forever. But the world's forests are finite in size. So what happens to estimates of species abundance when the entire world is your sample? […] The paradox would seem to run even deeper, because Fisher's logseries predicts that many more of the world's tropical tree species are hyper-rare. […] The truth is, we still have inadequate data to definitively answer the "how many tropical tree species?" question. Ecologists at present are forced to make huge extrapolations from existing inventory plot data to the entire world.*"

Hubbell (136) believes hyper-rare species do exist, as do we and in the case of areas smaller than the world, so do singletons. What are then those singletons? For an area like the Amazon, a huge and open system, singletons are most likely the result of species (locally) going extinct or new immigrants. ter Steege et al. (149) (SI Chapter 4 S1: Fig. S7) showed that several singleton species are in fact species found only once in the Amazon but common in the Cerrado, Andes and even Atlantic forest, 'vagrants' in the viewpoint of Magurran and Henderson (156). However, this may suggest that singletons or other hyper rare species are found mainly on the edges of an area. In the Amazon they were not and include such iconic species as *Asteranthos braziliensis* Desf. (endemic to the middle and upper Rio Negro) and *Duckeodendron cestroides* Kuhlm. (endemic to an area around Manaus, central Amazon). We believe that even if all individuals of the Amazon forest could be measured and identified, the biological processes of extinction and immigration would lead to the presence of at least ~750 singleton species, based on the Fisher's α found for the area (14) and a huge amount of hyper-rare species, some of which may have small contracted ranges, some of which may even be spread over large areas (157). One of the most important merits of NT is to emphasize the role of migration in building and maintaining community structures. However, the underlying mathematical model is based on a discretization down to the individual level, where a random process is supposed to play and can be expressed as per capita probabilities. In a complex system such as tropical forests, clearly not only chance acts upon birth, death, dispersal and migration. This could result from acquiring a new competitive advantage, losing a competitor because a pest, losing a pest because a super-pest develops. Myriad combinations are possible. The processes involved at local scale are not exclusively random but from local to global their combined effects on species abundances may appear to be.

Conclusion

To evaluate diversity of a rich, complex, large, open system, a parametric approach based on a probabilistic model such as Fisher's logseries, seems to be more applicable than a non-parametric one, because such a system is driven by the random walk resulting from an infinity of processes that vary among scales, and where chance affects many biological processes, and not just the random sampling context considered by nonparametric models.

Supporting Information

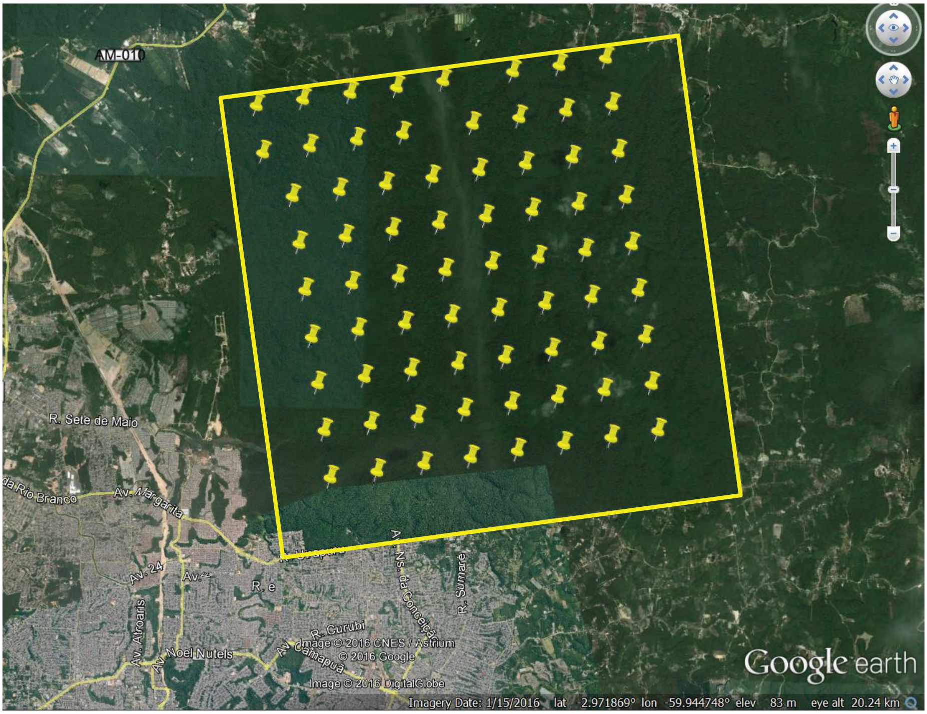


Fig. S1. Reserva Ducke, located just north of Manaus (AM), Brazil. The 72 0.5-ha plots are situated on a trail grid with trails 1 km apart. The reserve (yellow square) is ~10x10 km.

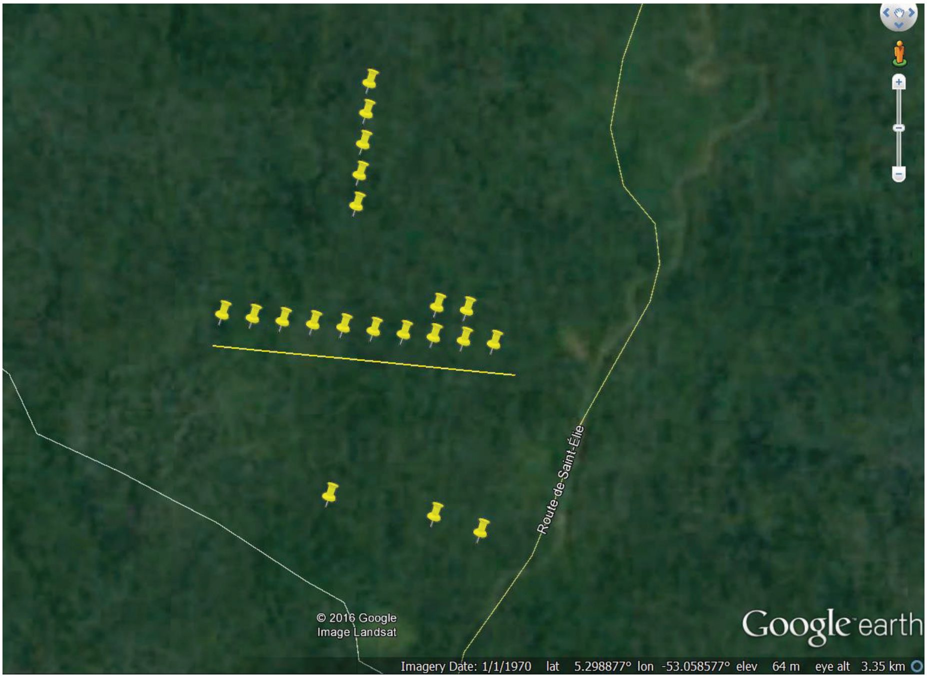

Route-de-Saint-Elie
© 2016 Google
Image Landsat
Google earth
Imagery Date: 1/1/1970 lat 5.298877° lon -53.058577° elev 64 m eye alt 3.35 km

Fig. S2. Map of 20 plots on the Piste de St. Elie location, French Guiana.

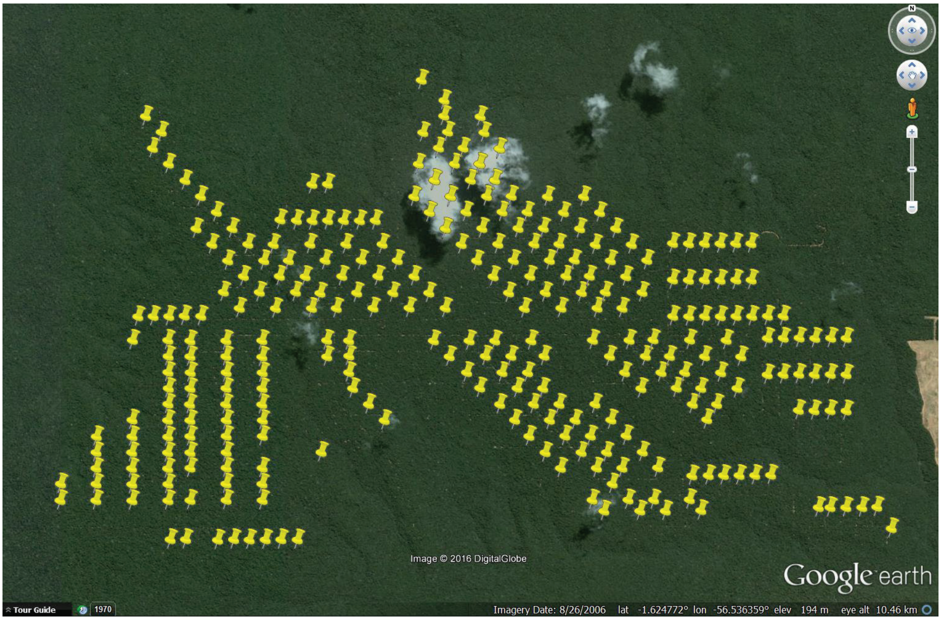


Image © 2016 DigitalGlobe
Google earth
Tour Guide
1970
Imagery Date: 8/26/2006 lat -1.624772° lon -56.536359° elev 194 m eye alt 10.46 km

Fig. S3. Map of 301 0.25-ha plots on the Monte Branco Plateau, Trombetas, Pará, Brazil. The plots were established along exploration lines of a geological survey.

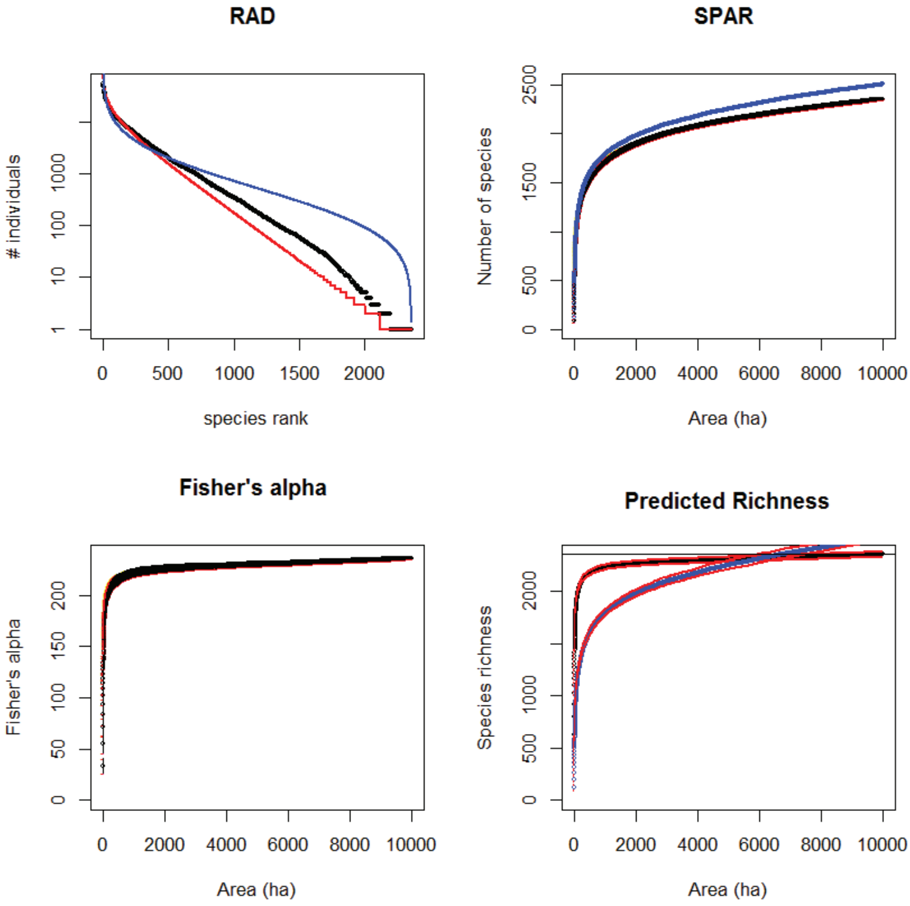


Fig. S4. Simulation of a 10,000 ha virtual forest with mean dispersal distance of 10 meters. Parameters used: m_{plot} = 0.88012; $m_{adjacent}$ = 0.108; m_{forest} = 0.0108; m_{MC} = 0.00108; ν = 10-4. A. Rank abundance distribution (RAD) of the total virtual (black) with logseries fit (red) and lognormal fit (blue). B. Species area curve for the total virtual forest and estimated richness ($S_{estimated}$) based on Chao1 (blue). C. Fisher's α area curve for the virtual forest. D. Species richness estimated with Fisher's α (black) and Chao1 (blue), each with 95% CI, and actual species richness of the simulated community (horizontal line).

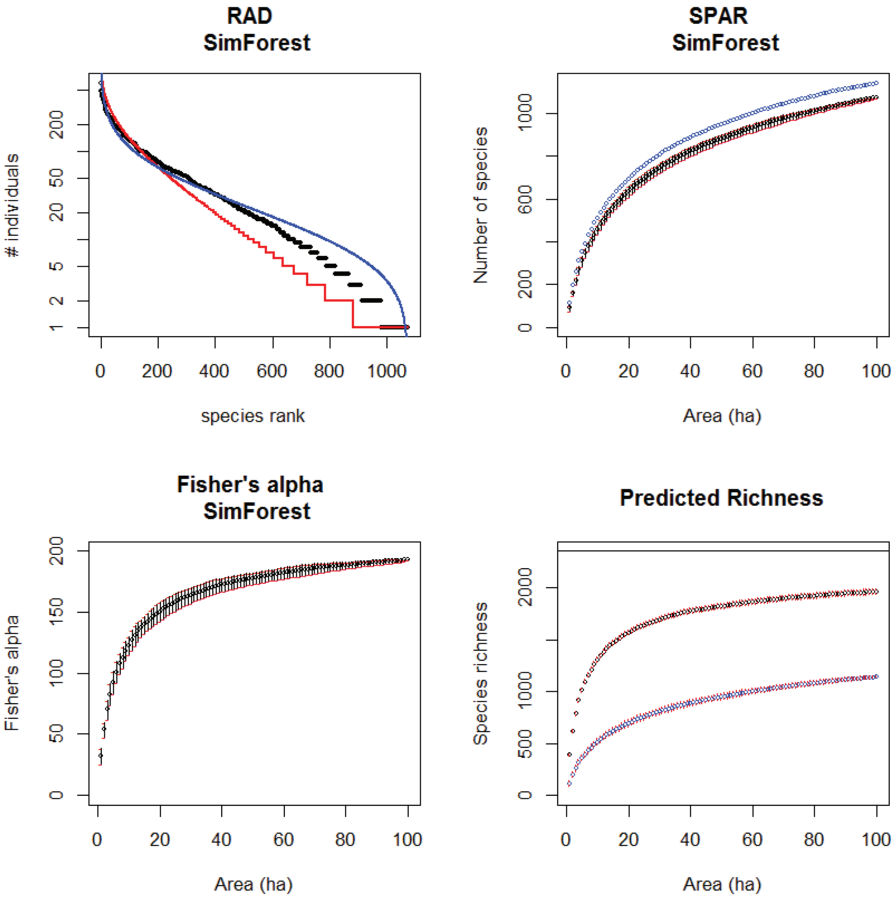


Fig. S5. Subset of 100 plots from a simulation of a 10,000 ha virtual forest with mean dispersal distance of 10 meters. Parameters see Fig. S4. A. Rank abundance distribution (RAD) of the 100 plots (black) with logseries fit (red) and lognormal fit (blue). B. Species area curve for the 100 plots and estimated richness ($S_{estimated}$) based on Chao1 (blue). C. Fisher's α area curve for the 100 plots. D. Species richness estimated with Fisher's α (black) and Chao1 (blue), each with 95% CI, and actual species richness of the simulated community (horizontal line).

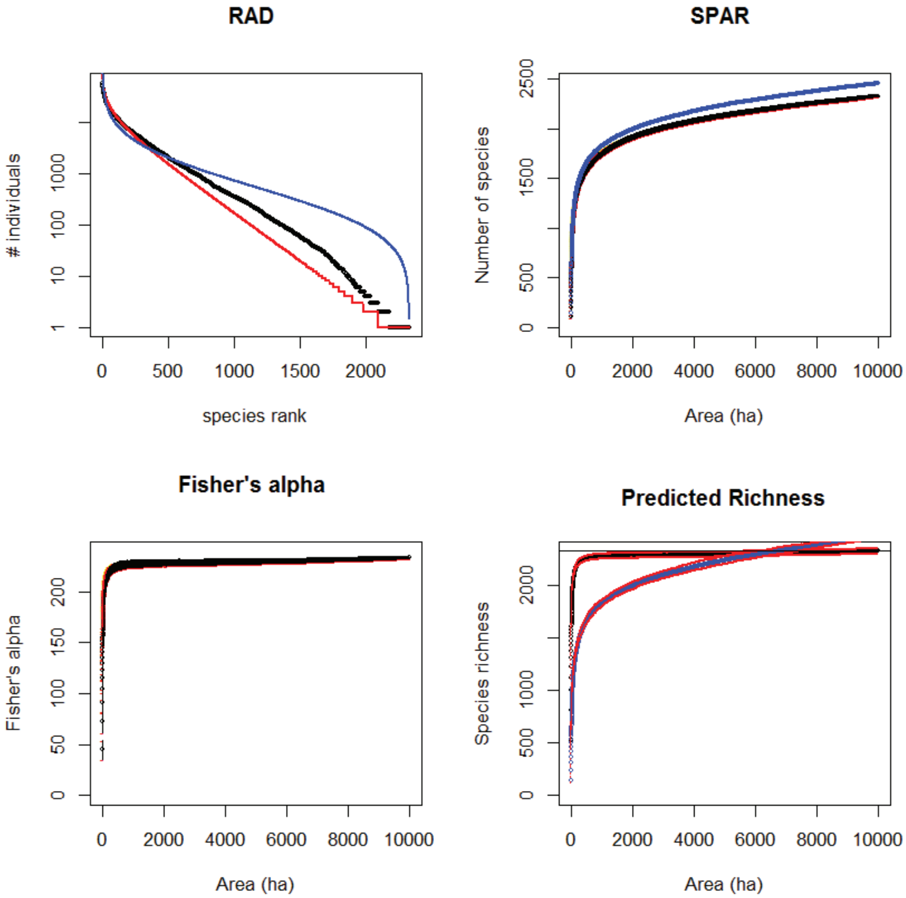


Fig. S6. Simulation of a 10,000 ha virtual forest with mean dispersal distance of 15 meters. Parameters used: m_{plot} = 0.83017; $m_{adjacent}$ = 0.153; m_{forest} = 0.0153; m_{MC} = 0.00153; ν = 10-4. A. Rank abundance distribution (RAD) of the total virtual (black) with logseries fit (red) and lognormal fit (blue). B. Species area curve for the total virtual forest and estimated richness ($S_{estimated}$) based on Chao1 (blue). C. Fisher's α area curve for the virtual forest. D. Species richness estimated with Fisher's α (black) and Chao1 (blue), each with 95% CI, and actual species richness of the simulated community (horizontal line).

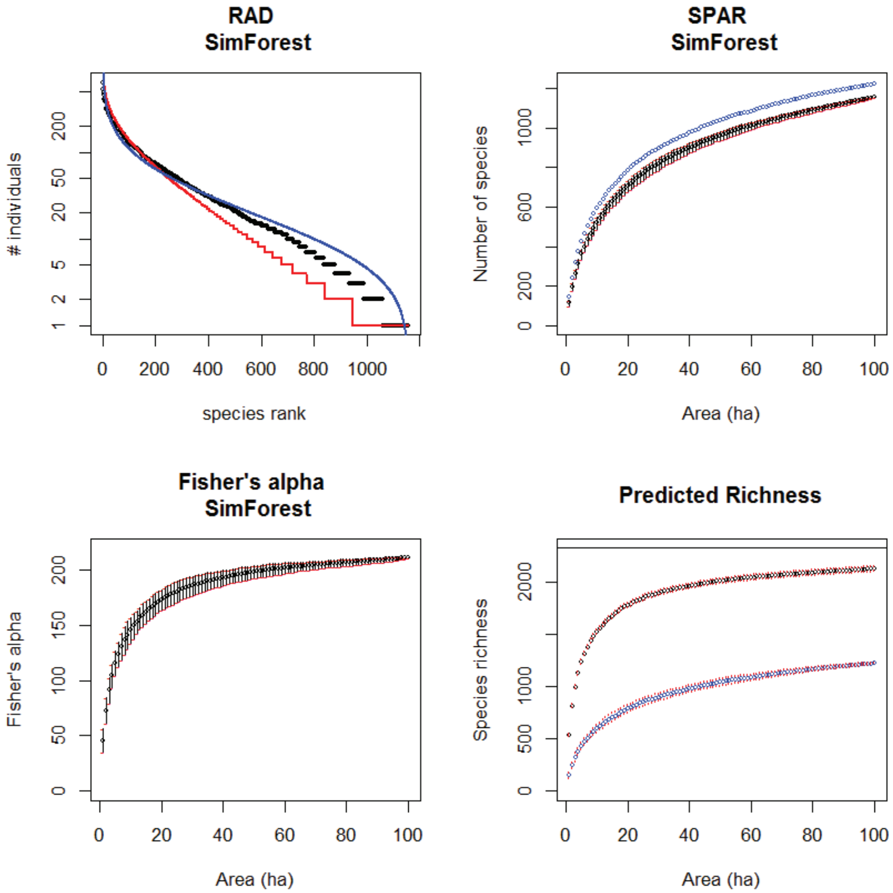


Fig. S7. Subset of 100 plots from a simulation of a 10,000 ha virtual forest with mean dispersal distance of 15 meters. Parameters see Fig. S6. A. Rank abundance distribution (RAD) of the 100 plots (black) with logseries fit (red) and lognormal fit (blue). B. Species area curve for the 100 plots and estimated richness ($S_{estimated}$) based on Chao1 (blue). C. Fisher's α area curve for the 100 plots. D. Species richness estimated with Fisher's α (black) and Chao1 (blue), each with 95% CI, and actual species richness of the simulated community (horizontal line). This figure is equal to Fig. 1 in the main text.

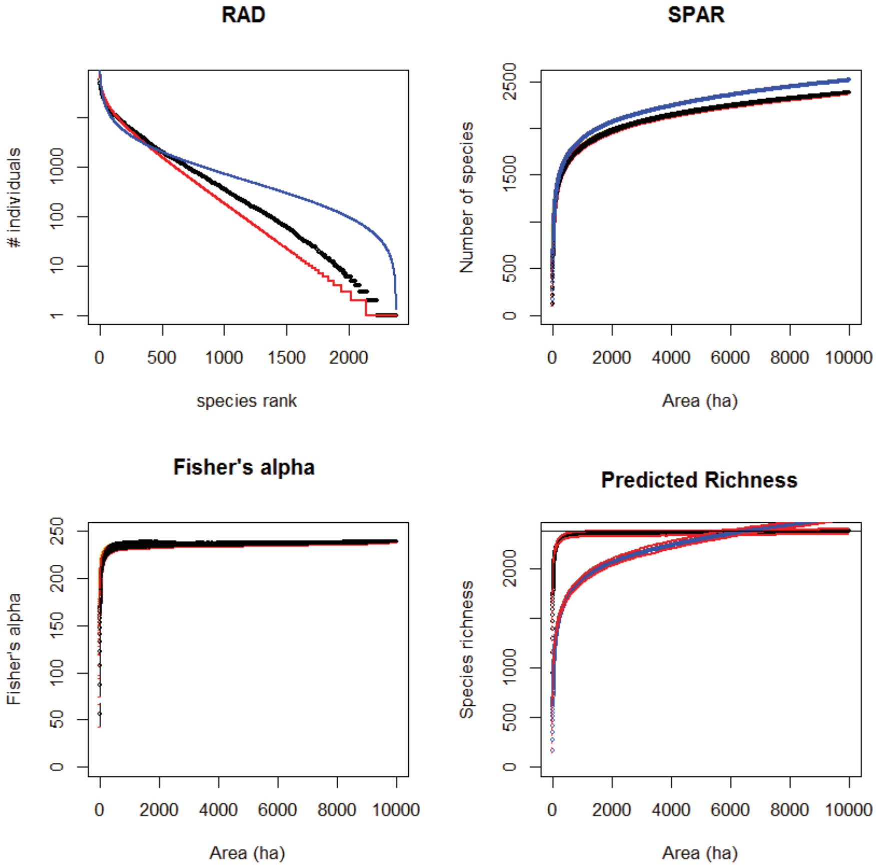


Fig. S8. Simulation of a 10,000 ha virtual forest with mean dispersal distance of 20 meters. Parameters used: m_{plot} = 0.78688; $m_{adjacent}$ = 0.192; m_{forest} = 0.0192; m_{MC} = 0.00192; ν = 10-4. A. Rank abundance distribution (RAD) of the total virtual (black) with logseries fit (red) and lognormal fit (blue). B. Species area curve for the total virtual forest and estimated richness ($S_{estimated}$) based on Chao1 (blue). C. Fisher's α area curve for the virtual forest. D. Species richness estimated with Fisher's α (black) and Chao1 (blue), each with 95% CI, and actual species richness of the simulated community (horizontal line).

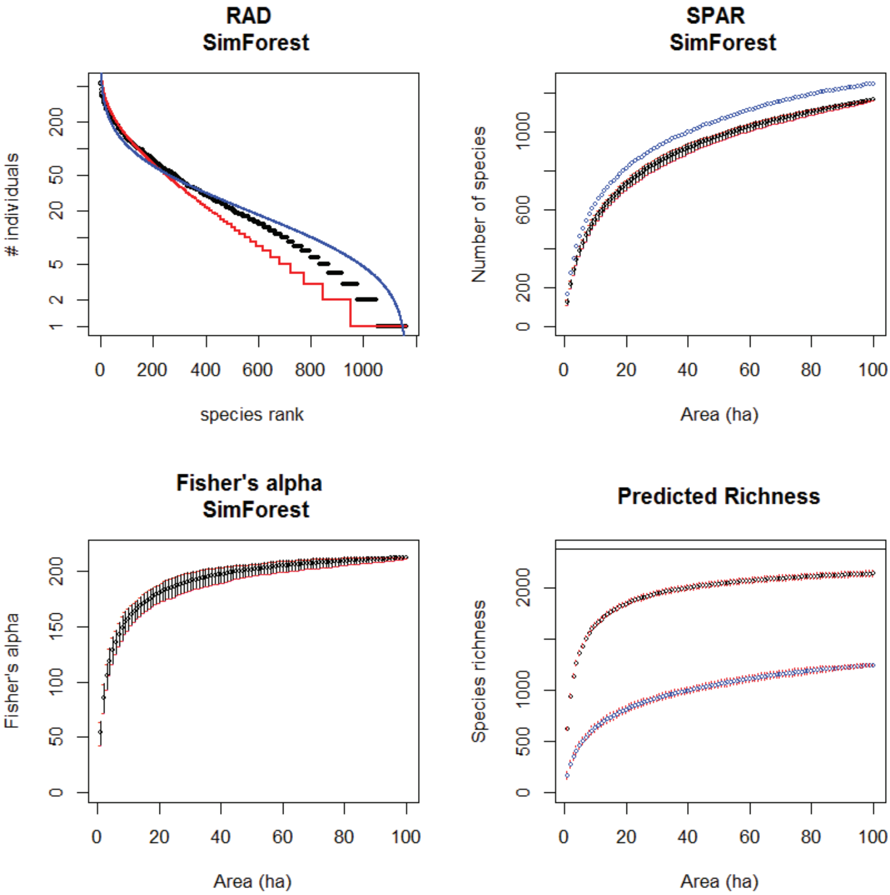


Fig. S9. Subset of 100 plots from a simulation of a 10,000 ha virtual forest with mean dispersal distance of 20 meters. Parameters see Fig. S8. A. Rank abundance distribution (RAD) of the 100 plots (black) with logseries fit (red) and lognormal fit (blue). B. Species area curve for the 100 plots and estimated richness ($S_{estimated}$) based on Chao1 (blue). C. Fisher's α area curve for the 100 plots. D. Species richness estimated with Fisher's α (black) and Chao1 (blue), each with 95% CI, and actual species richness of the simulated community (horizontal line).

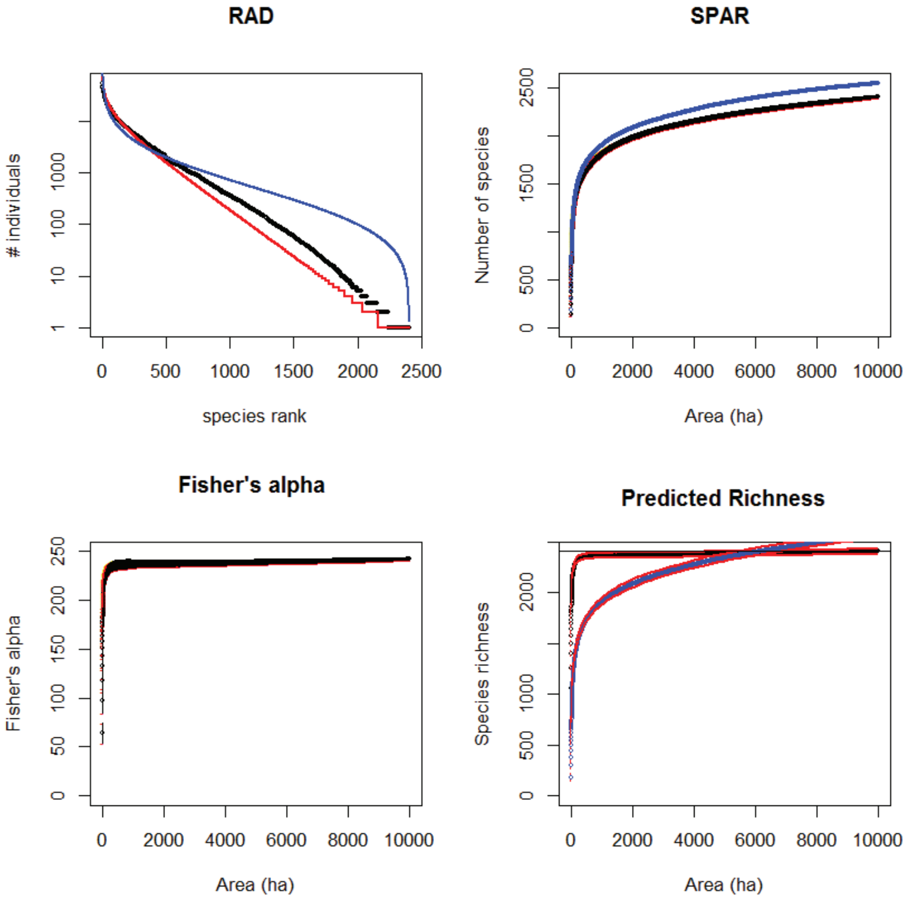


Fig. S10. Simulation of a 10,000 ha virtual forest with mean dispersal distance of 25 meters. Parameters used: m_{plot} = 0.75025; $m_{adjacent}$ = 0.225; m_{forest} = 0.0225; m_{MC} = 0.00225; ν = 10-4. A. Rank abundance distribution (RAD) of the total virtual (black) with logseries fit (red) and lognormal fit (blue). B. Species area curve for the total virtual forest and estimated richness ($S_{estimated}$) based on Chao1 (blue). C. Fisher's α area curve for the virtual forest. D. Species richness estimated with Fisher's α (black) and Chao1 (blue), each with 95% CI, and actual species richness of the simulated community (horizontal line).

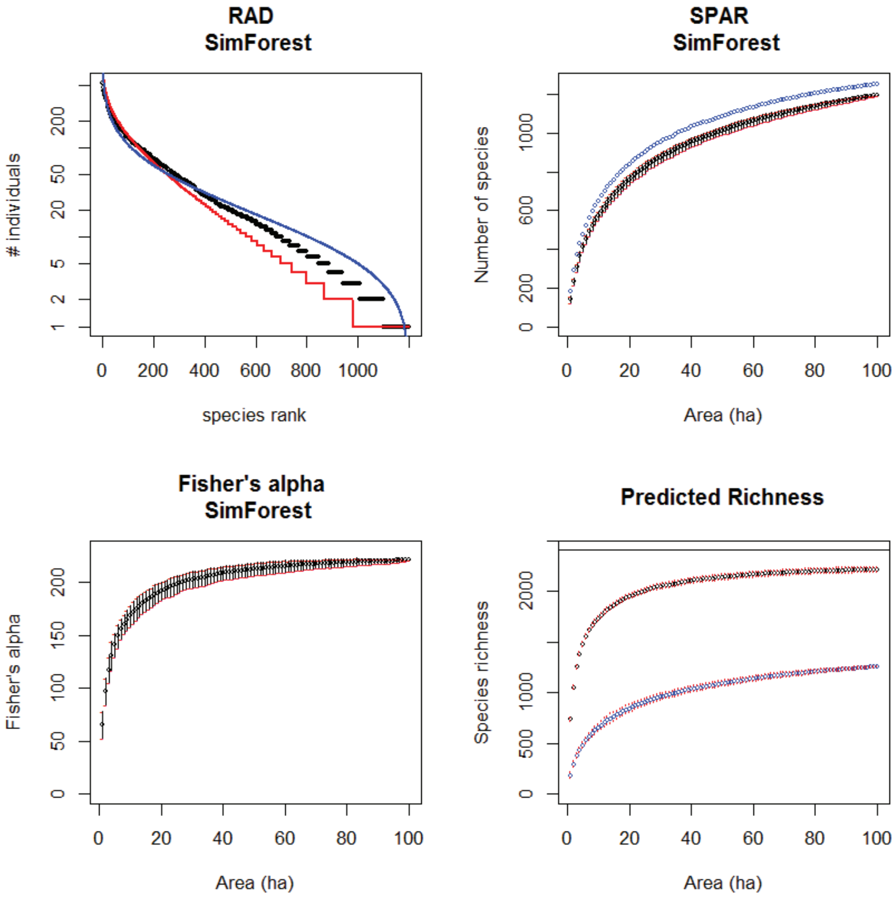


Fig. S11. Subset of 100 plots from a simulation of a 10,000 ha virtual forest with mean dispersal distance of 25 meters. Parameters see Fig. S10. A. Rank abundance distribution (RAD) of the 100 plots (black) with logseries fit (red) and lognormal fit (blue). B. Species area curve for the 100 plots and estimated richness ($S_{estimated}$) based on Chao1 (blue). C. Fisher's α area curve for the 100 plots. D. Species richness estimated with Fisher's α (black) and Chao1 (blue), each with 95% CI, and actual species richness of the simulated community (horizontal line).

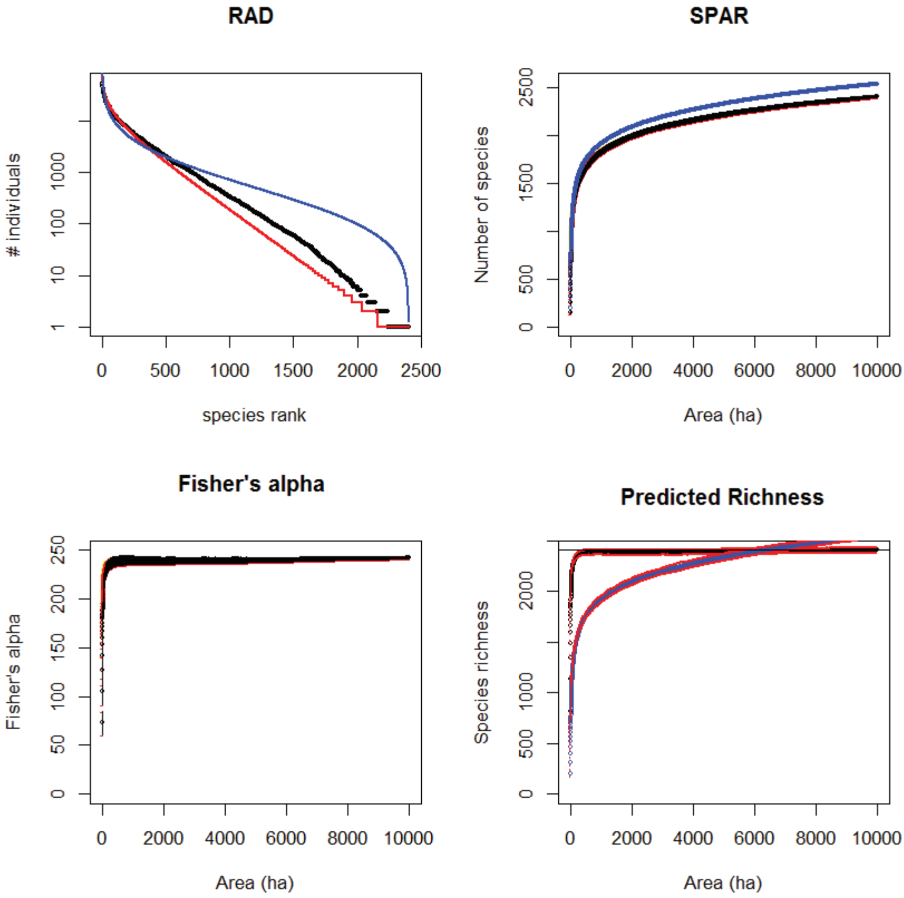


Fig. S12. Simulation of a 10,000 ha virtual forest with mean dispersal distance of 30 meters. Parameters used: m_{plot} = 0.72028; $m_{adjacent}$ = 0.252; m_{forest} = 0.0252; m_{MC} = 0.00252; ν = 10-4. A. Rank abundance distribution (RAD) of the total virtual (black) with logseries fit (red) and lognormal fit (blue). B. Species area curve for the total virtual forest and estimated richness ($S_{estimated}$) based on Chao1 (blue). C. Fisher's α area curve for the virtual forest. D. Species richness estimated with Fisher's α (black) and Chao1 (blue), each with 95% CI, and actual species richness of the simulated community (horizontal line).

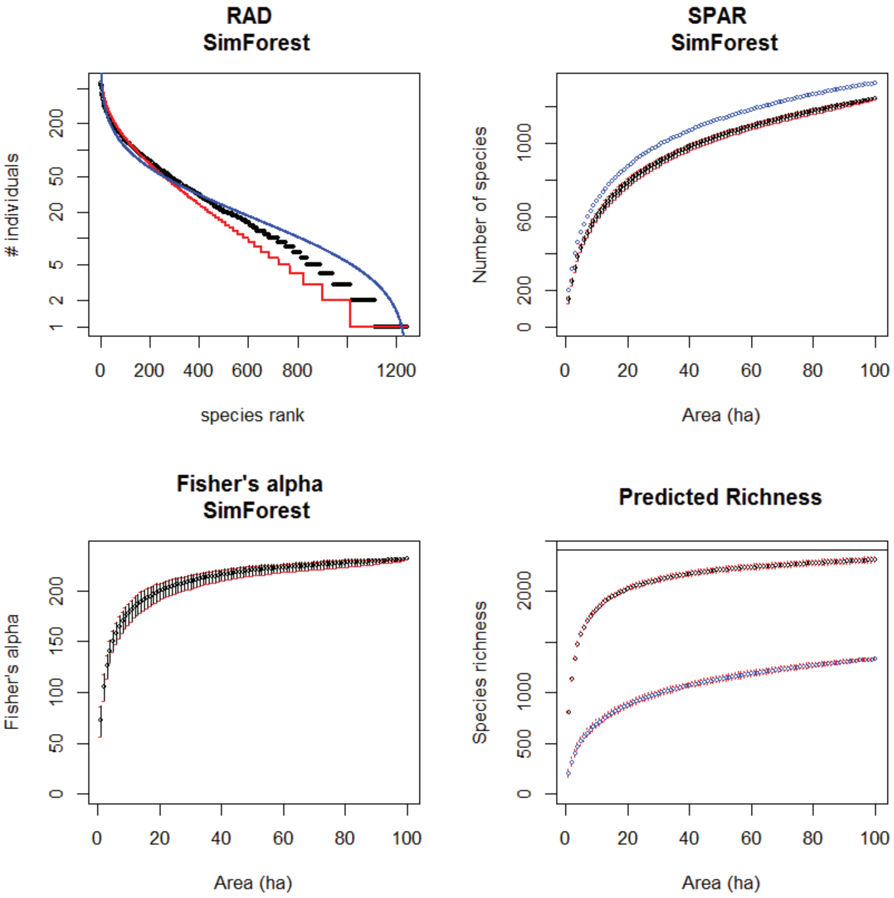


Fig. S13. Subset of 100 plots from a simulation of a 10,000 ha virtual forest with mean dispersal distance of 30 meters. Parameters see Fig. S12. A. Rank abundance distribution (RAD) of the 100 plots (black) with logseries fit (red) and lognormal fit (blue). B. Species area curve for the 100 plots and estimated richness ($S_{estimated}$) based on Chao1 (blue). C. Fisher's α area curve for the 100 plots. D. Species richness estimated with Fisher's α (black) and Chao1 (blue), each with 95% CI, and actual species richness of the simulated community (horizontal line).

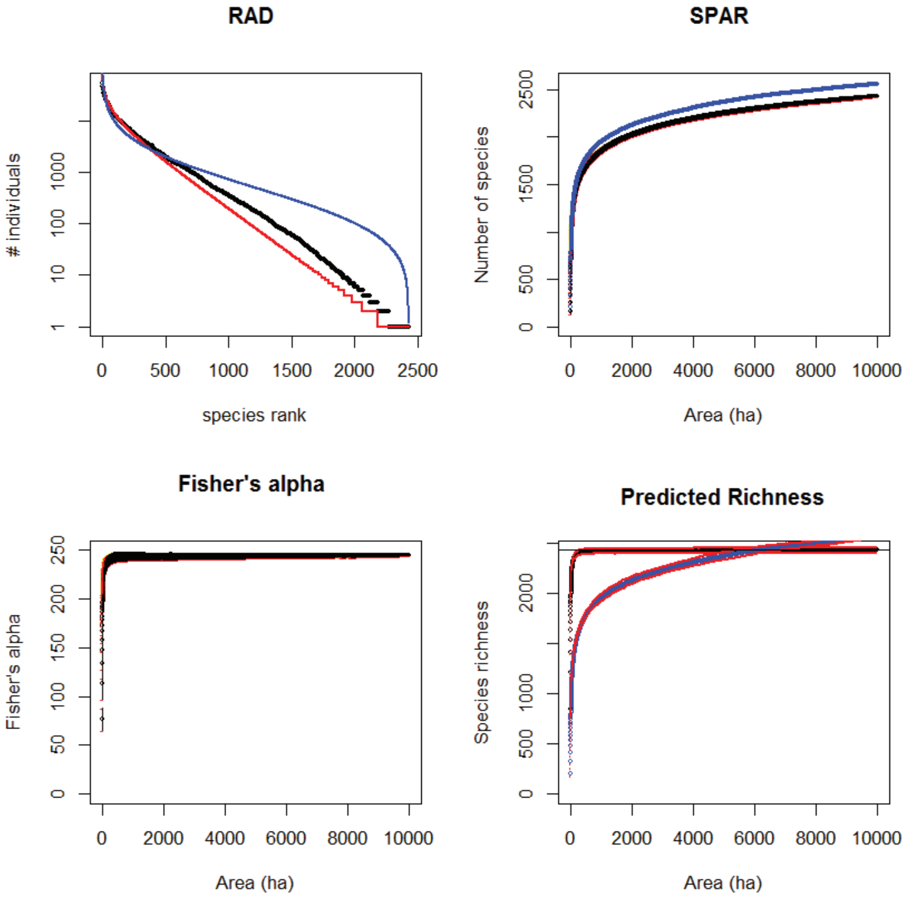


Fig. S14. Simulation of a 10,000 ha virtual forest with mean dispersal distance of 35 meters. Parameters used: m_{plot} = 0.69697; $m_{adjacent}$ = 0.273; m_{forest} = 0.0273; m_{MC} = 0.00273; ν = 10-4. A. Rank abundance distribution (RAD) of the total virtual (black) with logseries fit (red) and lognormal fit (blue). B. Species area curve for the total virtual forest and estimated richness ($S_{estimated}$) based on Chao1 (blue). C. Fisher's α area curve for the virtual forest. D. Species richness estimated with Fisher's α (black) and Chao1 (blue), each with 95% CI, and actual species richness of the simulated community (horizontal line).

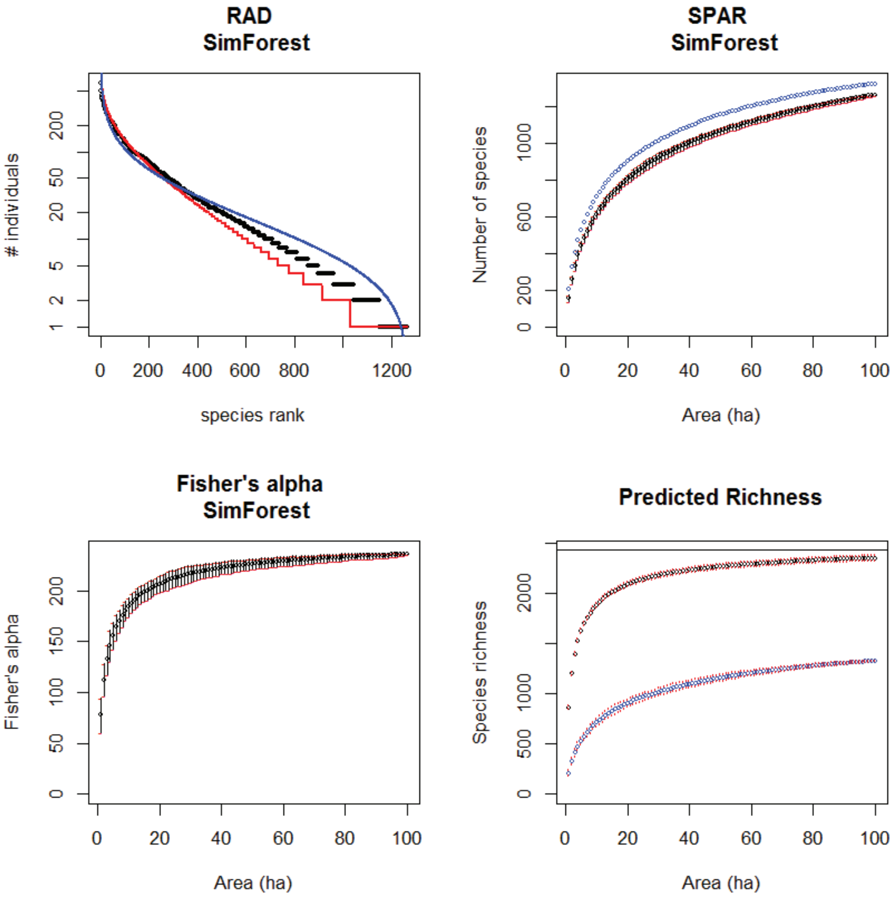


Fig. S15. Subset of 100 plots from a simulation of a 10,000 ha virtual forest with mean dispersal distance of 35 meters. Parameters see Fig. S14. A. Rank abundance distribution (RAD) of the 100 plots (black) with logseries fit (red) and lognormal fit (blue). B. Species area curve for the 100 plots and estimated richness ($S_{estimated}$) based on Chao1 (blue). C. Fisher's α area curve for the 100 plots. D. Species richness estimated with Fisher's α (black) and Chao1 (blue), each with 95% CI, and actual species richness of the simulated community (horizontal line).

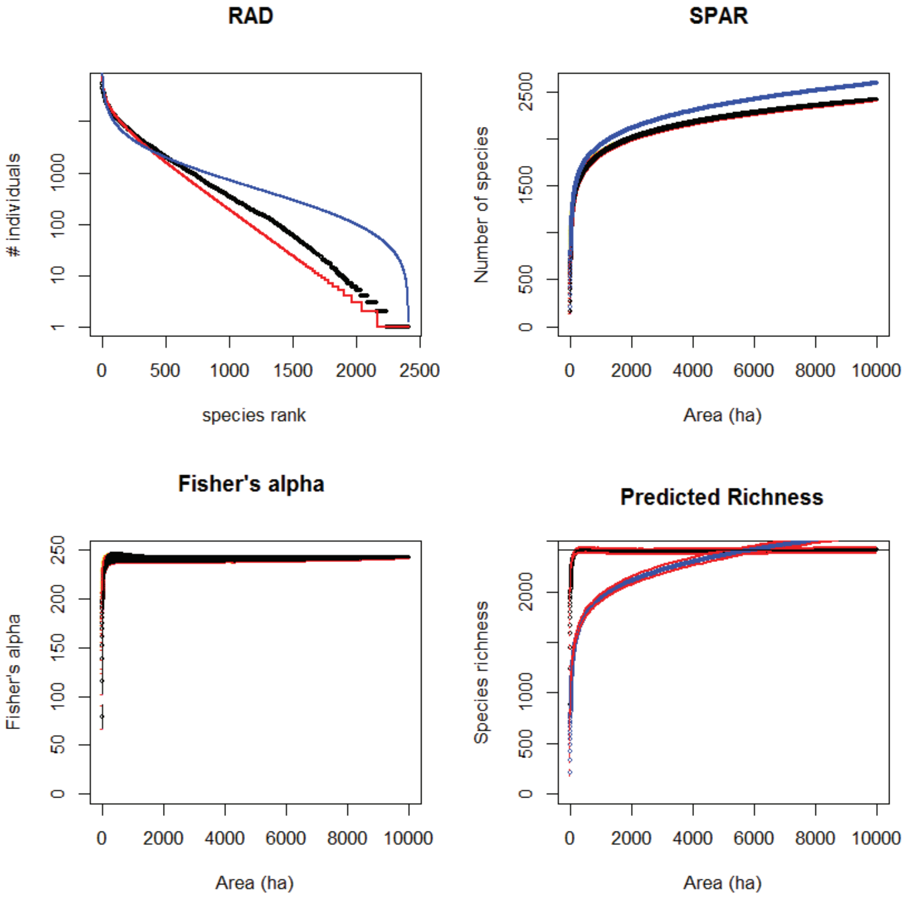


Fig. S16. Simulation of a 10,000 ha virtual forest with mean dispersal distance of 40 meters. Parameters used: $m_{plot} = 0.68032$; $m_{adjacent} = 0.288$; $m_{forest} = 0.0288$; $m_{MC} = 0.00288$; $\nu = 10-4$. A. Rank abundance distribution (RAD) of the total virtual (black) with logseries fit (red) and lognormal fit (blue). B. Species area curve for the total virtual forest and estimated richness ($S_{estimated}$) based on Chao1 (blue). C. Fisher's α area curve for the virtual forest. D. Species richness estimated with Fisher's α (black) and Chao1 (blue), each with 95% CI, and actual species richness of the simulated community (horizontal line).

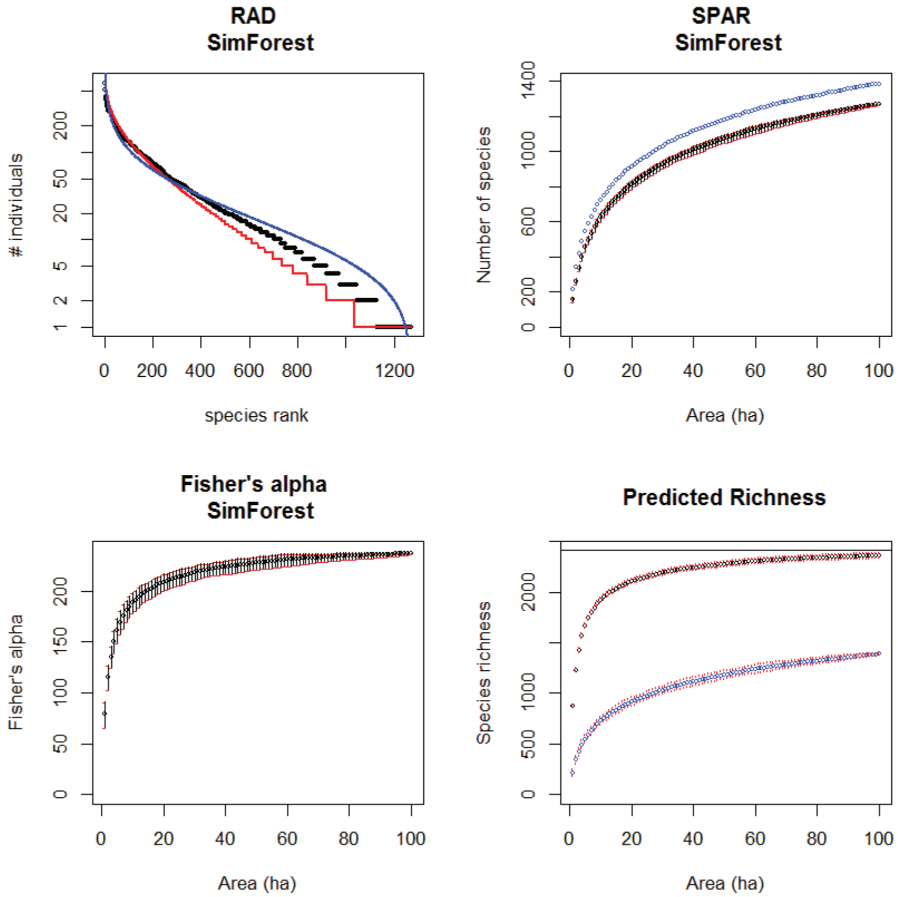


Fig. S17. Subset of 100 plots from a simulation of a 10,000 ha virtual forest with mean dispersal distance of 40 meters. Parameters see Fig. S16. A. Rank abundance distribution (RAD) of the 100 plots (black) with logseries fit (red) and lognormal fit (blue). B. Species area curve for the 100 plots and estimated richness ($S_{estimated}$) based on Chao1 (blue). C. Fisher's α area curve for the 100 plots. D. Species richness estimated with Fisher's α (black) and Chao1 (blue), each with 95% CI, and actual species richness of the simulated community (horizontal line).

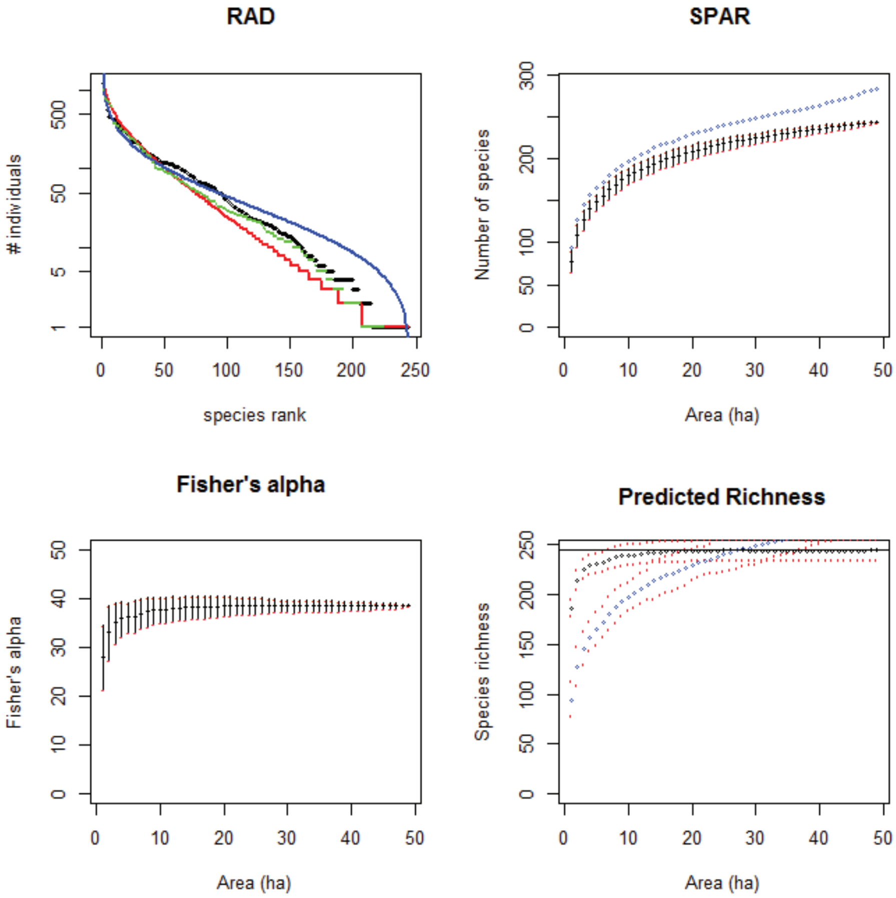


Fig. S18. Simulation of a 49 ha virtual forest from BCI data, with mean dispersal distance of 40 meters. Parameters used: m_{plot} = 0.68032; $m_{adjacent}$ = 0.288; m_{forest} = 0.0288; m_{MC} = 0.00288; ν = 0.00119. A. Rank abundance distribution (RAD) of the total virtual (black) with logseries fit (red), lognormal fit (blue), and RAD from plot data of BCI (green). B. Species area curve for the total virtual forest and estimated richness ($S_{estimated}$) based on Chao1 (blue). C. Fisher's α area curve for the virtual forest. D. Species richness estimated with Fisher's α (black) and Chao1 (blue), each with 95% CI (red), actual species richness of the simulated community (horizontal line).

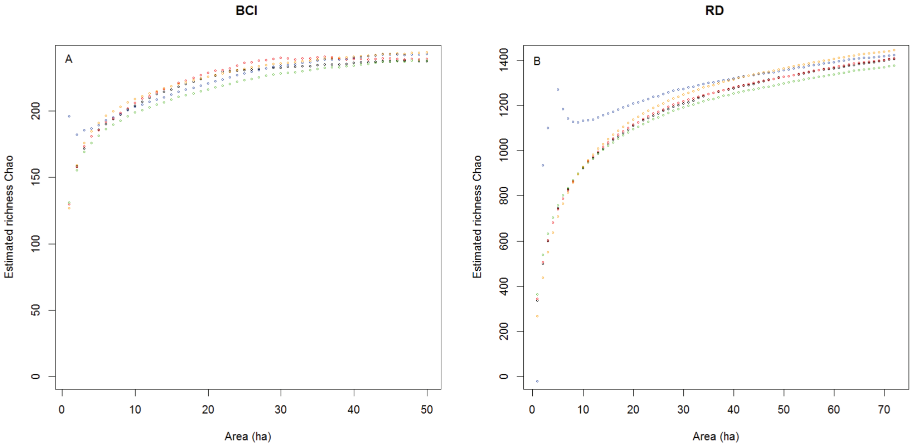
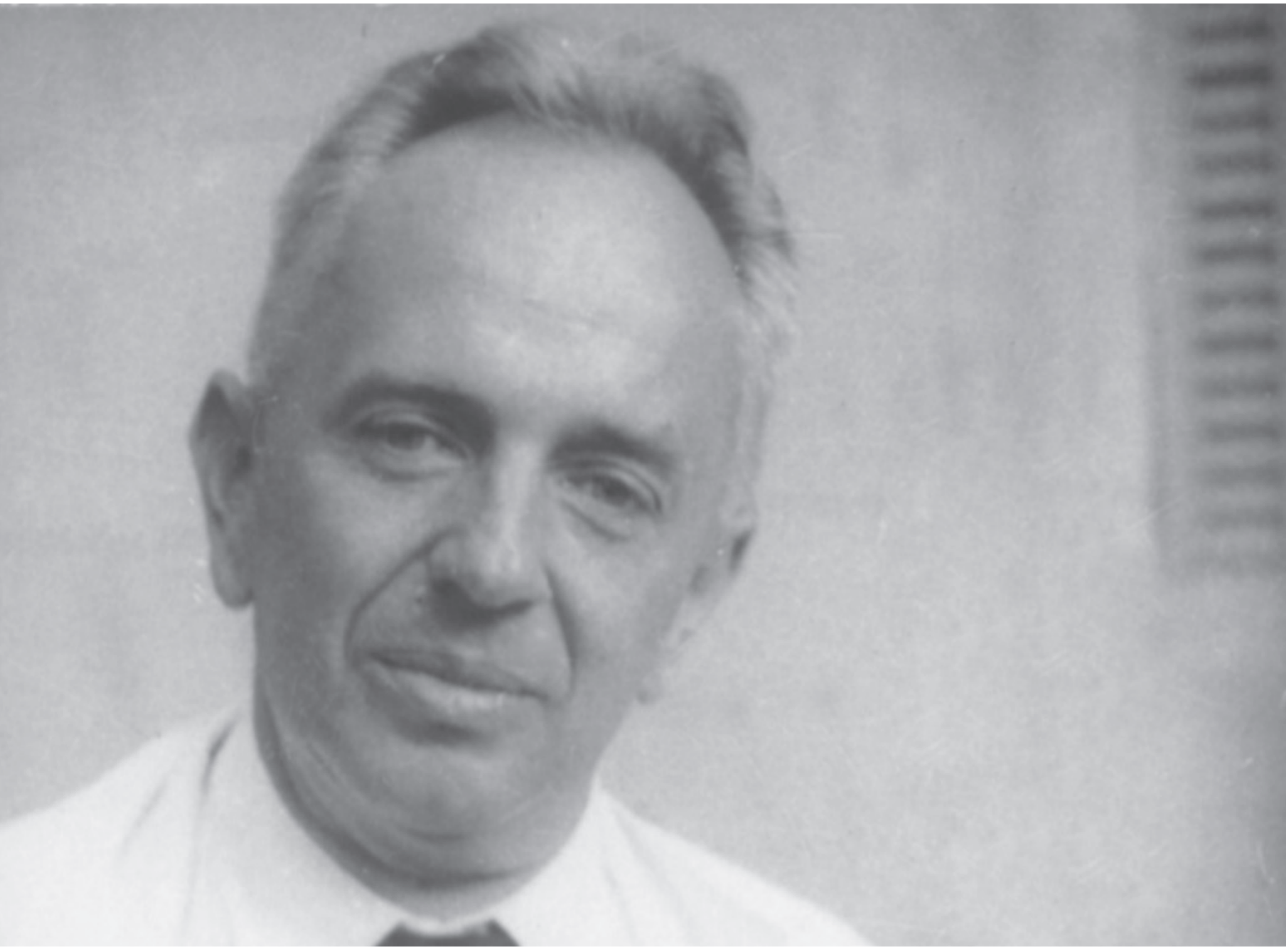


Fig. S19. Richness estimated with nonparametric species richness estimators as a function of the number of plots sampled. Nonparametric richness estimators for field data of BCI (A) and RD (B). In BCI 50 ha are sampled without replacement. In RD 72 plots are sampled without replacement. Black: Mean Chao1; Red: Chao1984; Blue: Chao Bunge; Green: Chao Lee; Orange: Jackknife. Estimators from the R package *Species* (Wang 2011)(144).



Seen in the light of evolution, biology is, perhaps, intellectually the most satisfying and inspiring science. Without that light it becomes a pile of sundry facts, some of them interesting or curious but making no meaningful picture as a whole.

Theodosius Dobzhansky, The American Biology Teacher (1973)

Chapter Five

Estimating and interpreting migration of Amazonian forests using spatially implicit and semi-explicit neutral models

(Published in Ecology and evolution 7.12 (2017): 4254-4265)

Edwin Pos[1,2], Juan Ernesto Guevara Andino[3], Daniel Sabatier[4], Jean-François Molino[4], Nigel Pitman[5,6], Hugo Mogollón[7], David Neill[8], Carlos Cerón[9], Gonzalo Rivas-Torres[10], Anthony Di Fiore[11], Raquel Thomas[12], Milton Tirado[13], Kenneth R. Young[14], Ophelia Wang[15], Rodrigo Sierra[13], Roosevelt García-Villacorta[16,17], Roderick Zagt[18], Walter Palacios Cuenca [19], Milton Aulestia[20], Hans ter Steege[2,1]

[1]Ecology and Biodiversity Group, Utrecht University, The Netherlands
[2]Naturalis Biodiversity Center, Group of Dynamic Biodiversity, Leiden, The Netherlands
[3]Department of Integrative Biology, University of California, Berkeley, CA 94720-3140, USA
[4]IRD, UMR AMAP, Montpellier, France
[5]The Field Museum, 1400 S. Lake Shore Drive, Chicago, IL 60605-2496, USA
[6]Center for Tropical Conservation, Nicholas School of the Environment, Duke University, Durham, NC 27708, USA
[7]Endangered Species Coalition, 8530 Geren Rd., Silver Spring, MD 20901, USA
[8]Universidad Estatal Amazónica, Puyo, Ecuador
[9]Universidad Central Herbario Alfredo Paredes, Escuela de Biología Herbario Alfredo Paredes, Ap. Postal 17.01.2177, Quito, Ecuador
[10]Colegio de Ciencias Biológicas y Ambientales and Galápagos Academic Institute for the Arts and Sciences, Universidad San Francisco de Quito, Diego de Robles S/N e Interoceánica, Quito, Ecuador.
[11]Univ. of Texas at Austin, Department of Anthropology, SAC 5.150, 2201 Speedway Stop C3200 Austin, Texas 78712, USA
[12]Iwokrama International Programme for Rainforest Conservation, Georgetown, Guyana
[13]GeoIS, El Día 369 y El Telégrafo, 3° Piso, Quito, Ecuador
[14]University of Texas, Geography and the Environment, Austin, Texas, 78712, USA
[15]Northern Arizona University, Flagstaff, Arizona, 86011, USA
[16]University of Edinburgh, Institute of Molecular Plant Sciences, Mayfield Rd, Edinburgh, EH3 5LR, United Kingdom
[17]Royal Botanic Garden of Edinburgh, 20a Inverleith Row, Edinburgh EH3 5LR, UK
[18]Tropenbos International, Lawickse Allee 11, PO Box 232, Wageningen, 6700 AE, the Netherlands
[19]Universidad Técnica del Norte, Herbario Nacional del Euador, Quito, Ecuador
[20]Herbario Nacional del Ecuador, Casilla 17-21-1787, Avenida Río Coca E6-115, Quito, Ecuador

Abstract

Aim. With many sophisticated methods available for estimating migration, ecologists face the difficult decision of choosing for their specific line of work. Here we test and compare several methods, performing sanity and robustness tests, applying to large-scale data and discussing the results and interpretation.

Location. South America (Guyana, Suriname, French Guiana and Ecuador).

Methods. Five methods were selected to compare for their ability to estimate migration from spatially implicit and semi-explicit simulations based on three large-scale field datasets. Space was incorporated semi-explicitly by a discrete probability mass function for local recruitment, migration from adjacent plots or from a metacommunity.

Results. Most methods were able to accurately estimate migration from spatially implicit simulations. For spatially semi-explicit simulations, estimation was shown to be the additive effect of migration from adjacent plots and the metacommunity, only accurate when migration from the metacommunity outweighed that of adjacent plots, discrimination proved to be impossible. We show that migration should be considered more an approximation of the resemblance between communities and the summed regional species pool. Application of migration estimates to simulate field datasets did show reasonably good fits and indicated consistent differences between sets in comparison with earlier studies.

Main conclusions. Estimated migration is more an approximation of the homogenization among local communities over time rather than a direct measurement of migration and hence has a direct relationship with betadiversity. As betadiversity is the result of many (non)-neutral processes, we have to admit that migration as estimated in a spatial explicit world encompasses not only direct migration but is an ecological aggregate of these processes. The parameter *m* of neutral models then appears more as an emerging property revealed by neutral theory instead of being an effective mechanistic parameter and spatially implicit models should be rejected as an approximation of forest dynamics.

Keywords: neutral theory, parameter estimation, species composition, migration, species diversity, betadiversity.

Introduction

Whether stochastic or deterministic processes govern species distribution has been a long-standing debate, starting with the equilibrium vs. non-equilibrium theories more than 25 years ago (158). The Unified Neutral Theory of Biodiversity and Biogeography, or UNTB (29) refuelled this discussion (32, 34, 39, 159–161). Prior to this debate, the main accepted view of population dynamics was of a niche-based origin, i.e. species being specifically adapted to certain environments where they could thrive, while outcompeted elsewhere. Processes as competitive exclusion (47, 162) and niche partitioning (44, 163, 164) were believed to be the main drivers of differences in species composition. Actual niches occupied by species were thought to be determined by specific suits of adaptations for certain environments and biotic interactions among species (165). This combination of interspecific differences and environmental heterogeneity allowed for coexistence. In contrast, the UNTB is neither based on such interspecific differences nor environmental heterogeneity. It assumes that all individuals are ecologically equivalent in terms of demographic events such as birth and death, but also in rates of migration and their probability of speciation. As a result, the main differences in species composition are simply based on stochastic processes, resulting from ecological equivalence. It was not a fully novel approach, however, as the model of Island Biogeography by MacArthur and Wilson was also truly neutral in its mathematical foundations treating species equivalent in demographics, even though the authors still regarded species as having distinct niches in real life (53). Much work on neutral theory had already been developed in population genetics, some implicit, such as the Island Model (52), others explicit such as the Stepping Stone model (63). The UNTB relies heavily on these models of genetic differentiation between communities, with the neutral theory of molecular evolution (64) obviously being one of its pillars (29). Many criticized the UNTB (156, 166–171) and many supported it (35, 36, 74, 172, 173). Today, many ecologists agree that both deterministic and neutral processes play a role in determining species composition (174–177). To study their relative importance, models are often used to investigate whether communities behave neutrally or not. An important question still remaining is how to parameterize neutral models. Suggestions for estimating two of the core parameters of Hubbell's neutral model, speciation and migration, have been proposed over the years and the importance of parameter estimation has been discussed previously (178). These studies concentrated, however, specifically on the difference between estimating from a single (large) sample or multiple samples in a spatially continuous landscape. They did not focus on the role of spatial relationships, i.e. the effect of distance between plots when estimating migration. We feel this effect of distance is important because space and migration can be incorporated in two different ways; either spatially implicit (29, 60) or spatially explicit (35, 70, 179, 180). Models of the first kind work on the assumption of a panmictic system. They disregard the spatial position of individuals within each community as there is only one migration parameter *m*, determining whether a recruit is from the regional or local species pool but there is no within community dispersal limitation. Even though such models show good fits, the existence of such a panmictic community is unlikely, due to the physical dispersal ability of individuals versus the size of many communities (63). In contrast, spatially explicit models consider the metacommunity rather as the sum of a number of local communities, between which there exists an explicit spatial relationship. The first models, where the spatial position of each individual was explicitly modelled, were based on a discrete grid-like structure, each cell containing an individual which could disperse to either neighbouring cells (73, 181) or to other regions by implementing different dispersal kernels (35, 70). However, while there are quite some analytical solutions for the implicit models, only few exist for the explicit versions such as developed by O'Dwyer and Green (2010) by applying principles from physics.

Comparisons show that, although spatially explicit models should approximate the real world better, spatially implicit models provide better fits to empirical data (182, 183). Hence, the latter are more often used when estimating migration, even though field data comes from a spatially explicit reality. In this study, we therefore extend the comparison of estimation methods towards the practical ability of these methods to estimate migration from simulated datasets based on both spatially implicit and spatially semi-explicit models. We focus on five different parameter estimation methods: 1) a sampling formula by Etienne (184), 2) the Inference method by Jabot et al. (185), 3) The Gst statistic adopted from population genetics by Munoz et al. (186), 4) the two-stage sampling formula by Etienne (187), which is an extension on the two-stage-estimation method by Munoz et al. from 2007 (188) and 5) a method by Chisholm and Lichstein (138) based on plot geometry and absolute dispersal distances. A summary of the different estimation methods can be found in the Supporting Information Chapter 5: S1. For the interested and more mathematically oriented reader we refer to the original papers, as here we are focusing on the use of the methods rather than their exact mathematical derivation. Our first goal is to perform a sanity check on each method. They should at least be able to recover parameter estimates from models on which they are based. Our second and main goal is to establish whether these methods are also robust, i.e. if they are able to accurately recover parameters when performed on models a bit different from the models on which they are based. For this, we apply them to a spatially semi-explicit model in which migration can either be from a hypothetical metacommunity or from adjacent plots. Our last and third goal is to apply each method to field data. For this we use three different independent field datasets; Guyana/Suriname (GS), French Guiana (FG) and Ecuador (EC), which are highly distinct in their forest dynamics (189). Using both spatially implicit-, semi-explicit- and field data we hope to reach a broad public of ecologists working on similar problems.

Methodology

Comparison of model parameter estimation

Each parameter estimation method, as described above, was used to generate an estimation of migration for a number of situations using spatially implicit, (semi) explicit simulated and field datasets. Results were compared from the simulated datasets in terms of their ability to accurately describe migration as parameterized to construct the datasets. After using the simulated datasets we turned to the actual field data, having multiple local communities assumed to be a sample from the larger metacommunity for which migration was also estimated using the same estimation methods. Etienne's sampling formula (2005) and the Inference method of Jabot et al (2008) were both tested using the TeTame freeware version 2.1 (http://chave.ups-tlse.fr/projects/tetame.htm). Etienne's two-stage sampling method was tested using the PARI/GP environment (190). Chisholm & Lichsteins's method was tested using MATLAB (191) and the Gst statistic was computed using the package *untb* (192) in the R environment (51). Other R-packages used were *Quantreg*, *Vegan*, *Labdsv* and *FasianOptions* (99, 193–195).

Metacommunity simulation

For both spatially implicit and explicit simulations, the first step was to create the larger metacommunity. The relative abundance distribution of tree species in the Amazonian forests shows a nearly exact fit with Fisher's Logseries (14, 137). We therefore used this relationship and the related number of species for a given abundance (98) to derive the relative abundance distribution from the expected number of species (*S*) and individuals (*N*) in the metacommunity, given by $\Phi n = \alpha xn/n$. Here, Φn is the number of species with *n* individuals; α is Fisher's α and x is

Table 5.1 Summary of table S1, with the mean difference between given and estimated migration (Δm), using spatially implicit simulations. Results from the corrected plot geometry method by Chisholm & Lichstein are not shown as they yield a single value with a confidence interval shown in Table S1

Summary difference m.given vs m.est and range SD of estimations

	One-stage est.		Inference method		G_{st}-statistic		2-stage (Etienne)	
Dataset	Δm	SD range	Δm	SD range	Δm	SD range	Δm	SD range
Guyana/Suriname	.044	.032-.06	.0075	.009-.016	.0200	.043-.382	-	-
French Guiana	.071	.033-.060	.0078	.009-.016	.0240	.044-.325	-	-
Ecuador	.132	.022-.061	.0070	.008-.018	.0160	.043-.418	.004	.017-.046

given by *N/(N + α)* (*N* being the number of individuals in the total sample and x being asymptotically equal to 1 with large sample sizes). We created three different metacommunities: two for the simulated spatially implicit datasets and one for the spatially semi-explicit dataset. Because of the observed difference between the Guianas and Ecuador in terms of diversity and composition (Fig. S10 from (14)) and the regions being separated by a large geographical distance, we created two different metacommunities for the spatially implicit simulations related to these two regions rather than one large metacommunity. They are hereafter referred to as MC-high and MC-low respectively (MetaCommunity high and low diversity). Ter Steege et al. (2013) estimated mean tree densities for all species per degree grid cell and by fitting the mean rank abundance curve to Fisher's Logseries distribution estimated the total amount of species to be expected by country (Fig. S10 from (14)). We adopted these figures to construct MC-low (20,191,600,511 individuals and 4582 species) and MC-high (5,611,001,426 and 6834), for details on both see the Supporting Information Chapter 5. For the simulated spatially explicit dataset a separate metacommunity was constructed using the same methods based on the Reserva Ducke forest, with 5,5 million trees and a Fishers alpha of 272 (196); hereafter referred to as MC_spatial. The logseries for each community was constructed starting from the left tail (most dominant species). The fixed parameters *α* and x were first calculated from the number of individuals (*N*) and species (*S*), after which the maximum dominance according to Fisher's logseries for all species is calculated, which is then given the first rank. For each subsequent rank the predicted number of species is then calculated until all species are given a rank and all individuals are distributed.

Spatial semi-explicit									
Simulation parameters and yielded variables						Estimated migration			
						Inference method		G_{st}-statistic	
dataset	Nr. sp.	Nr. sing	m.local	m.adj	m.meta	m2	SD	m3	SD
1	1777	244	0	0	1.00	.990	.028	1.011	.0057
2	1088	37	.79	.20	.01	.140	.015	.156	.0012
3	1529	142	.79	.01	.20	.209	.021	.210	.0015
4	1542	147	.75	.05	.20	.244	.024	.247	.0017
5	1282	73	.75	.20	.05	.200	.019	.205	.0014
6	1093	48	.69	.30	.01	.197	.019	.215	.0013
7	1609	169	.69	.01	.30	.310	.027	.312	.0020
8	1277	74	.65	.30	.05	.260	.023	.270	.0017
9	1077	50	.59	.40	.01	.254	.021	.277	.0016
10	1666	182	.59	.01	.40	.416	.034	.419	.0024
11	1315	97	.55	.40	.05	.325	.027	.341	.0019
12	1056	36	.49	.50	.01	.310	.028	.330	.0019
13	1690	186	.49	.01	.50	.512	.040	.517	.0028
14	1301	96	.45	.50	.05	.380	.032	.400	.0023
15	1706	187	.39	.01	.60	.615	.046	.621	.0034
16	1727	220	.29	.01	.70	.716	.050	.721	.0039
17	1748	224	.19	.01	.80	.819	.050	.822	.0042

Table 5.2 Estimates of migration based on a semi-spatially explicit neutral model. Probability of migration was determined either from adjacent plots (*m.adj*), the metacommunity (i.e. all other plots except the local and adjacent plots; *m.meta*) or the local plot. Number of plots was 400 with a runtime of 1e^{8} for all datasets.

Spatially implicitly simulated data. For the spatially implicit datasets we used the exact same sampling procedure as proposed by Hubbell in the original UNTB. Each time step, one individual dies, which is replaced by an individual having an ancestor either in the local community (with probability *1-m*) or from the metacommunity (with probability *m*). The identity of the recruit is then only dependent on its relative abundance in the respective community. Datasets based on GS and FG (67 and 63 plots) were sampled from the MC-low assuming they share the same metacommunity and the dataset based on EC (72 plots) from MC-high. Sampling of the local communities was repeated for a range of migration parameters (Supporting Information Chapter 5: Table S1). For details on the number of time steps used see the Supporting Information Chapter 5 (S2). After the construction of the simulated datasets, migration was estimated using the above-mentioned estimation methods.

Spatially semi-explicitly simulated data. Spatially semi-explicit simulations were done by modeling a lattice of 20x20 plots, each with 500 individuals. We assume no spatial explicit arrangement of individuals within a plot. Taking a random sample from the metacommunity creates the forest at time t_0. Each time step (t+1) one individual from each plot to be replaced was chosen at random from the MC_spatial metacommunity, this was repeated for 10,000 time steps. Recruitment was generated from either of three sources: 1) migration from adjacent plots (*m.adj*), 2) migration from the MC_spatial metacommunity (*m.meta*) or 3) local recruitment (*1-(m.adj+m.meta) = 1-m*). According to studies on long distance dispersal of seeds (LDD) the majority of seeds (>99%) often fall within ca. 100 meters of their origin (197), depending on among others, seasonal conditions, wind speed, turbulence initiated by the canopy and particle fall velocity, which is obviously also affected by seed mass and shape (198, 199). As the plots from the field data used in this study are 1 hectare in size, it is reasonable to assume that migration either does not occur but there is local recruitment, or there is migration mostly from adjacent plots when the tree of origin would be on the edge of a plot, with occasionally seeds ending up further away. Hence, this subdivision in dispersal categories using a discrete probability mass function seems a likely approximation of the actual dispersal of individuals and allows for much faster calculations by the computer. Values for both *m.adj* and *m.meta* were based on an arbitrary division of the range of migration used for the spatially implicit simulations (see also Table 5.2).

Table 5.3 Parameter estimation for the three field datasets. For the corrected Plot Geometry method by Chisholm & Lichstein (2009) (138) the following parameters were used: Guyana/Suriname w = 100, d = 15-25 m, French Guiana, w = 100 m, d = 25-35 m, Ecuador, w = 100 m and d = 40-50 m

	Inference method		G_{st}-statistic		2-stage (Etienne)		Cor. Plot Geometry	
Dataset	m2	SD	m3	SD	m4	SD	m5	CI
Guyana/Suriname	.075	.050	.046	.044	.084	.074	.071	.055-.088
French Guiana	.22	.085	.11	.058	.170	.062	.103	.088-.119
Ecuador	.26	.153	.17	.152	.246	.114	.147	.133-.160

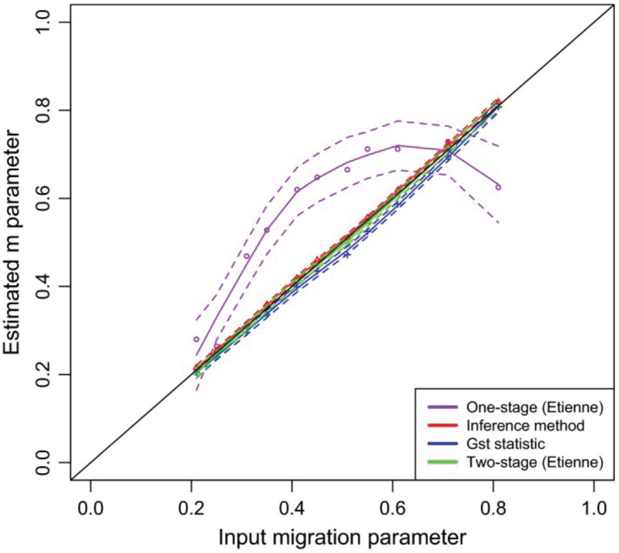


Fig. 5.1 LOESS regressions of the migration parameter used for input versus the estimated migration from the spatially implicit simulations. Results from each method indicated by color with broken lines indicating the 95.5% confident interval, polynomial degree and span used for the LOESS regression was 2 and .75, respectively.

Species composition of field data. Three different sets of field data from the Amazon Tree Diversity Network (14) were used for analysis. Two sets belong to the Guiana Shield: Guyana/Suriname (GS) combined and French Guiana (FG), the third set contains data from forests in Ecuador (EC). All three sets are completely independent and non-overlapping (200). Datasets are composed of 63-72 1 ha plots with all trees ≥10 cm DBH inventoried. Species names of all datasets were standardized with the W3 Tropicos database within each dataset, using TNRS (94), as described in more detail in ter Steege et al. 2013 (14). The EC dataset has 72 plots of one hectare, yielding 34,544 individuals and 2021 morphospecies. The GS and FG datasets are composed of 67 and 63 one-hectare plots respectively. In GS, 37,446 individual trees were distributed among 1042 morphospecies, and FG had 35,075 individuals belonging to 1204 morphospecies.

Table 5.4 Results from the spatially implicit models based estimates of *m* using the three separate field datasets. Fisher's alpha was averaged over all plots; first row of each set shows actual field data.

Dataset	method	migration	metacommunity	plots	species	singletons	Fisher's alpha
Guyana/Suriname	-	-	-	67	1042	210	198
Guyana/Suriname	Inference method	.075	MC-low	67	885	69	158
Guyana/Suriname	G_{st} statistic	.046	MC-low	67	826	83	146
Guyana/Suriname	Two-stage Etienne	.084	MC-low	67	896	78	164
Guyana/Suriname	Cor. Plot Geometry	.071	MC-low	67	801	97	151
French Guiana	-	-	-	63	1204	208	177
French Guiana	Inference method	.220	MC-low	63	1045	113	197
French Guiana	G_{st} statistic	.110	MC-low	63	964	105	179
French Guiana	Two-stage Etienne	.170	MC-low	63	975	116	188
French Guiana	Cor. Plot Geometry	.103	MC-low	63	910	95	169
Ecuador	-	-	-	72	2021	518	468
Ecuador	Inference method	.260	MC-high	72	1667	243	126
Ecuador	G_{st} statistic	.170	MC-high	72	1333	167	289
Ecuador	Two-stage Etienne	.246	MC-high	72	1489	196	324
Ecuador	Cor. Plot Geometry	.147	MC-high	72	1373	196	292

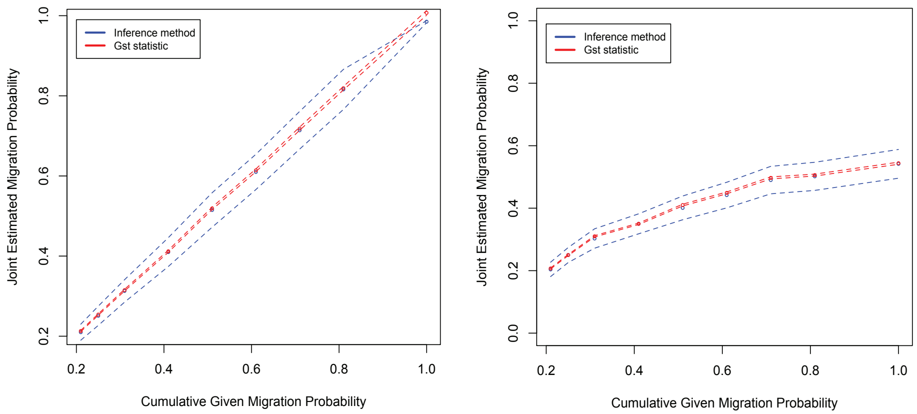


Fig. 5.2 Given joint migration probability with either migration predominantly coming from the metacommunity (left) or from adjacent plots (right) plotted against the estimated joint migration by both the Inference method (blue) and Gst statistic (red). Broken lines indicate the estimation plus or minus the standard deviation of the average over all plots used in the simulation. It is clear that when migration mostly comes from the metacommunity, both estimation methods are very accurate, and when migration from adjacent plots is dominant, both estimation methods are underestimations.

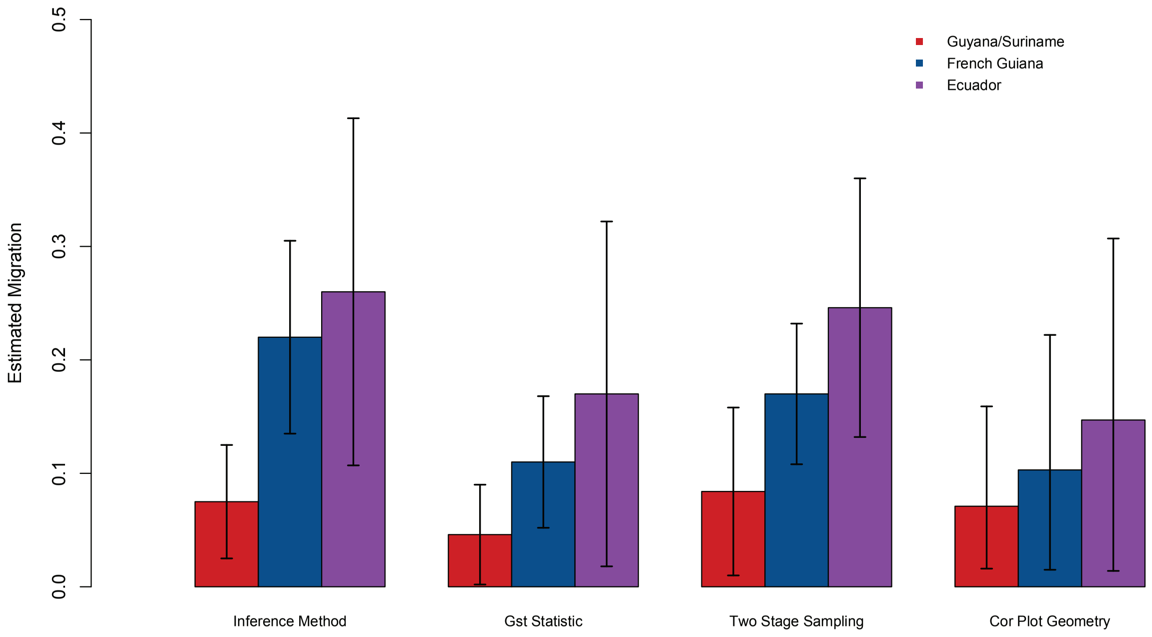


Fig. 5.3 Estimated migration probability from each empirical dataset (Guyana/Suriname, French Guiana, and Ecuador) for the Inference method, Gst statistic, Two-stage Sampling estimation and the Corrected Plot Geometry method. For the first three methods, whiskers indicate standard deviation of the estimation. For the corrected plot geometry method, they are representative of the confidence interval for the estimation.

Results

Comparing parameter estimation methods: spatially implicit and explicit

Sanity checks on each method showed that the Inference method and Gst statistic were able to approximate the complete range of migration parameters based on each different field dataset accurately. Etienne's one stage sampling method showed larger deviations. The two-stage sampling by Etienne was only used for the spatially implicit dataset based on EC due to extreme long computation time (see details in the Supporting Information Chapter 5: S1) but also generated accurate estimations. Average difference between given and estimated migration was .08, .007, .02 and .004 for the Etienne's one stage sampling, Gst statistic, Inference method, and Etienne's two-stage sampling respectively (see Table 5.1 for a summary and Supporting Information Chapter 5 Table S1 for details). All methods except the one stage sampling by Etienne thus showed very good accuracy when given migration parameters were plotted against the estimated migration (Fig. 5.1). Next we tested the robustness of each of the methods when applied to slightly different models. Etienne's one stage sampling formula was not used for estimating data from spatially explicit simulation because of the larger deviations found with the spatially implicit simulations. The corrected Plot Geometry method was also excluded because estimation of migration would be constant over the range of parameters used. The two-stage estimation method by Etienne was also not used due to practical limitations as explained earlier. Hence, we were only able to use the Inference method and Gst statistic. The migration estimates from the spatially semi-explicit simulations were the additive effect between migration from the adjacent plots and the metacommunity (Table 5.2). As both methods generate a single migration value, they were only able to estimate the joint migration probability. As example, in one of the simulated sets the parameters were set such that 1% of replacements were drawn from the 8 cells surrounding the cell in which an individual died and 20% of replacements are drawn from the metacommunity surrounding these adjacent cells (*m.adj* of .01 and *m.meta* of .20, Table 5.2, dataset 3). Both the Gst statistic and the Inference method estimated a migration probability of .21, indicating that these probabilities are additive in the estimation and it is still unknown whether migration is from close by or far away. Estimation of the joint migration probability was only accurate when migration from the metacommunity was higher than from the adjacent plots. In the contrasting situation (*m.adj* > *m.meta*) estimations were generally an underestimation of the joint migration probability (Table 5.2 and Fig. 5.2).

Parameter estimation from field data. The Gst statistic, Inference and Etienne's two-stage sampling formula were used to estimate migration from the three field datasets. Calculation of migration using the Corrected Plot Geometry method was based on the following parameters: edge length of plot (w) 100 meters for all three sets (as each plot is 1 ha) and mean absolute dispersal distances in the ranges 15-25 meters for GS, 25-35 meters for FG as it has more pioneer species in comparison to the first, and 40-50 meters for EC as it is relatively comparable to the BCI plot in Panama having rich soils sustaining rapid dynamics (i.e. fast growth). This yielded migration parameters of .237 with a confidence interval (CI) of .182-.293, .344 (CI .293-.396) and .489 (CI .444-.533) for GS, FG and EC respectively (see also Table 5.3). After applying the correction as explained in the Supporting Information Chapter 5 this was .071 (CI .055-.088), .103 (CI .088-.119) and .147 (CI .133-.160). Here CI is given instead of SD as the corrected Plot Geometry method gives a single estimate depending on plot geometry and mean dispersal distance, the CI is then related to the lower and upper limit of the dispersal range. In the same order (GS, FG and EC), the Gst statistic yielded estimates of .046, .11 and .17 (SD: .044, .058 and .152). Using the Inference method, this was .075, .22 and .26 (SD .050, .085 and .153), for Etienne's two-stage sampling this was .084, .170 and .246 (SD .074, .062 and .114); see also Table 5.4 and Fig. 5.3.

Comparing parameter estimation from field and simulated datasets. We implemented all migration parameters from the spatially implicit simulations in the spatially implicit model and compared the results of the relative Rank Abundance Distribution (RAD), number of species and singletons (i.e. species with one individual) and Fisher's alpha generated from the simulations to the actual field data (Supporting Information Chapter 5: Figure S5, S6 and Table 5.4). As example, for GS (having 1042 morphospecies and 210 singletons), the simulated dataset based on the MC_low metacommunity using a migration of 0.046 (Gst statistic estimation from field data) yielded a total number of 826 species belonging to 41,875 individuals (67 plots times 625 individuals) with 83 singletons and an average Fisher's alpha of 146. When using .075 as the probability of migration (as estimated by the Inference method) this was 885 species, 69 singletons and a Fisher's alpha of 158. For a migration of .084 (Etienne's two-stage sampling method) this was 896 species, 78 singletons and a Fisher's alpha of 164 and finally with a migration parameter of .071 (corrected Plot Geometry method) this was 801 species, 97 singletons and 151 as Fisher's alpha. Using the spatially implicit simulations with the estimation migration probabilities hence tended to show less species and a smaller amount of singletons than the actual field data, which was the case for FG and EC as well (see Table 5.4). For the comparison of RAD's from field data and simulations see Supporting Information Chapter 5, Figs. S5 and S6.

Discussion

Most methods used for estimating migration rates of neutral models are based on Hubbell's original spatially implicit model or its derivations (178, 184–188, 201, 202). This implicit approach contrasts strongly with reality for tropical trees, as the morphology of for example fruits, seeds and also different strategies play an important role in defining the average dispersal distance of plants (203–205). In addition, in real life, dispersal limitation is also not neutrally distributed among species. Although this disagreement is quite apparent, the inference of migration using such estimation methods is often done to study forest dynamics and the relative importance of niche versus neutral processes shaping communities. Here we show that although the estimation methods we compared were able to correctly estimate migration from models of which they were derived, they fail to do so for models in which there is a spatially explicit relationship. For the spatially implicit simulations, the Inference method (185) and Gst statistic (186) yielded comparable results and were able to estimate migration very precisely (Table 5.1 and Supporting Information Chapter 5: S1). The two-stage sampling method by Etienne was only used for the spatially implicit datasets based on EC due to long computation time, but showed comparable results. The only exception was the one stage estimation method by Etienne, which in particular for higher probabilities of migration showed a larger deviation (see also Fig. 5.1). This method is based on the likelihood calculation of P[D|θ, m, J], the multivariate probability of observing a current specific species abundance distribution given the constraints of the parameters (see also Supporting Information Chapter 5: S1). This in essence is the sum of all possible species-ancestry-abundance distributions. The problem that could occur here, although we did not test this explicitly, is that this may be a result of the way m is related to I by $m = I/(I+J-1)$ with J the size of the community. Hence, I is equal to $m(j-1)/(1-m)$ and when m approaches unity, I reaches infinity. Thus as migration approaches one and I becomes increasingly large, the expression (4) from Etienne (2005) is reduced to become only dependent on one term, namely $A=J$. Intuitively this means that all individuals in the community are a potential ancestor, thus coming from the metacommunity and likelihood estimates of migration could potentially deviate substantially from what is given. Other problems might be caused by the way this method is implemented in the software as used in this study (Etienne, personal communication). Perhaps further study into this phenomenon could shed more light on these results.

When we turn to the semi-spatial explicit simulations we see a different result. Each method yields only a single estimation for migration per sample. As such, it was obvious they would only estimate a joint migration probability instead of those from separate sources of migration. This total migration rate, however, could still be the correct total migration, if it would in fact measure actual migration or at least approximate it. Given that there is no spatial relationship in the model from which the methods are derived, however, we expected that estimation methods based on a spatial implicit reality would struggle to infer migration when this is larger from nearest neighbour communities than that from the larger metacommunity. Although intuitively this makes sense, as far as we know this has not been tested with actual large-scale field data before nor has it been shown to what extent it would deviate using a quantitative modelling approach. Our results supported our expectation and showed that this joint estimation was accurate only when migration from the metacommunity was higher than from the adjacent plots. In contrast, if *m.adj > m.meta* which would be the normal situation in reality for tropical trees, estimations were consistently found to be an underestimation of the joint migration probability (see Table 5.2 and Fig. 5.2). Although only the Inference method and Gst statistic were used for the latter, we assume given the earlier results on the spatially implicit simulations that the two-stage sampling by Etienne would generate similar results.

Here we show the consequences of using estimation methods based on a spatially implicit model to estimate migration from a spatially explicit reality. When the majority of migration is coming from the metacommunity, even spatially semi-explicit simulations approach a spatially implicit reality. One could ask if we would ever expect estimations of migration to be accurate when we are using spatially implicit models. Given the model's assumptions and rules we think this would only be the case if the actual system approaches a spatially implicit system, i.e. when there is no true spatial relationship between composition and geographical distance. In this case these methods would estimate migration correctly (i.e. *m = m.adj + m.meta*). In bryophytes this may be the case, or at least the data were consistent with the predictions of the spatially implicit neutral model (206). When spores get in the upper wind layers they are capable of travelling almost across the entire Amazon, although the majority of replacements will still be local recruits. In such a spatially implicit reality, each local community is considered a sample from the metacommunity, how much it actually resembles the metacommunity depends on the migration parameter (estimated to .2 for the bryophytes). In Hubbell's original UNTB, species abundances deviate from the expected abundance (its proportional abundance in the metacommunity) because migration determines the time that ecological drift operates within the local community. In other words, the migration parameter determines differences in species diversity between the plot under consideration and the diversity of the total sample used for analysis and hence has a direct relationship with betadiversity found in the total sample. This is meaningful when estimating from neutral spatially implicit simulations, where the only relationship is that of migration between each plot and the metacommunity. When it comes to the real world it is a different matter as betadiversity can be the result of many neutral and non-neutral processes (Supporting Information Chapter 5: Fig. S9). As such, it also becomes apparent why the neutral model shows such good fits when estimating migration and implementing it in a neutral model, even though we know the world is not neutral. Migration as estimated from a spatially implicit model encompasses not only dispersal but is in fact an ecological aggregate of all processes determining betadiversity; dispersal, time, competition, habitat selectivity, predation, frequency dependent mortality etc. It is the link between the (summed) regional species pool and each local community.

As example, different forest dynamics can play an important role in determining forest diversity and hence the estimation of migration. Wood density, relative growth rate and seed mass are related to dispersal, shade-tolerance and are considered indicative for difference between pioneering or non-pioneering species (207–211). High wood density, slow growth and large seed mass are reflected in slower forest dynamics (31, 212). In contrast, low wood density, low seed mass and faster turnover of individuals are reflecting faster forest dynamics. Marzluff and Dial showed that turnover and seed mass influence the ability to colonize new resources, leading to a potential higher diversity for forest having higher turnover and smaller seed mass (213). On the other hand, strong selective pressures or a very homogeneous environment in combination with fast turnover might cause plots to look more similar to each other due to natural selection, hence decreasing differences in species composition or even decreasing total species richness. In both cases, estimation of migration would potentially be relatively high as similarity between plots is also fairly high (low betadiversity), but again, neither neutral processes nor dispersal had little to do with it. Strong natural selection and a very heterogeneous habitat can also cause high betadiversity, decreasing estimates of migration. The above-mentioned processes shape species composition and have an influence on the connection between the regional species pool and the local species pools, but have no neutral fundament. To be fair, the stochastic (neutral) counterpart of selection, ecological drift, can obviously also cause differences in species composition. Similar to population genetics, if drift is very pronounced, rare species will disappear and systems will lose diversity. But we know that this is by far not the only mechanism responsible for differences in community composition and that estimates of migration do not tell us specifically how much influence this stochastic mechanism has in shaping diversity. Regarding this mechanism, we did observe an interesting pattern in the ratio between observed and expected singletons according to Fishers Logseries. As communities are structured according to Fisher's Logseries, we can calculate the expected number of singletons based on the total number of species and individuals and compare this with the observed number of singletons in each sample. When forests are well mixed in the case of little dispersal limitation, the observed number of singletons should approach the expected number of singletons dependent on sample size. When this ratio deviates from one, this indicates that local plots are less connected to each other over larger distances resulting in a clumpier distribution. This eventually means fewer singletons than expected according to Fisher's Logseries based on the number of individuals and species. We showed that there indeed was a strong relationship between the amount of migration from the metacommunity and this ratio of expected versus observed singletons. This idea is further explained and studied in the Supporting Information (S4: Further analysis of migration using Fishers Logseries). A last note on interpreting estimates of migration focuses on the aspect of time. Given enough time at an ecological time-scale, a collection of local communities will potentially have shared much of their species overall. Even when having low direct migration between each local community. This is the result of each local community acting as a stepping-stone, if individual species travel short distances each generation, they can still travel great distances. This inevitably increases the theoretical value of migration. Small differences in species composition (and thus high estimates of migration) can thus be the result either of low migration over a long period of time or high migration in a short period of time.

Reinterpreting estimation of migration from field data. We showed that estimates of migration from all three regions differed markedly (see Table 5.4). Although there were small differences between estimations when using different methods, relative differences between each dataset within one method were comparable. Guyana and Suriname showed the lowest migration probabilities, followed by French Guiana and finally Ecuador. Knowing that these estimates of migration are actually ecological aggregates, what differences in these forests can we attribute to these differences in migration probability? The relationship between community dynamics and alpha diversity was already shown for forests within Guyana (208). Ter Steege et al. furthermore showed that on average, Western Amazonian forests are 150 individuals ha-1 denser than Eastern Amazonian forests and also have a higher alpha diversity (115). Forests of the Guiana Shield also experience a less heterogeneous environment and a more climax species composition having a higher seed mass and higher wood density in comparison with forests of Ecuador (209). This all suggests that forests of the Guiana Shield probably experience an overall stronger selection pressure, slower dynamics and potentially also a higher impact of ecological drift due to smaller population sizes and less dispersal ability. All of these potentially lead to a stronger distance decay of similarity and a higher betadiversity, both also shown earlier (214). This would also explain a lower estimate of migration of forests of the Guiana Shield in comparison with Western Amazonian forests, such as those of Ecuador.

Conclusions

We have shown that estimation of migration using methods based on species composition fails when estimating from spatially (semi-)explicit simulations. Estimation was only correct when our spatially semi-explicit model approached a spatially implicit world. We summarize that there are three major problems when using estimation methods based on spatially implicit models on a spatially explicit reality: 1) Estimations of migration relate to the differences in species diversity between plots and the diversity of the total sample used for analysis as it is based on a spatially implicit model, not an actual mechanism of dispersal; 2) As differences in species diversity can be the result of a number of potential causes, the migration parameter does not solely reflect neutral dynamics as it is assumed to do so in neutral models. It is an aggregated ecological parameter, capturing a myriad of different processes. And 3) even if the migration parameter could actually be considered being reflective of the migration of individuals and not including any other mechanisms, these methods still only look at the end result of the homogenization. Hence, it does not shed any light on actual current forest dynamics, as it can be the effect of much migration in a short period or little migration over a long period.

The only method used in this paper not based on species composition and hence not influenced by the problems mentioned above is the (corrected) Plot Geometry method (138). This uses plot geometry and absolute dispersal distances of individuals. It therefore attempts to estimate the actual amount of migration per time index as migration, although the original authors still implemented this into a spatially implicit model. For spatially (semi-) explicit models, it offers a much more intuitive implementation of migration and shows promising results (196). We propose that the next steps would be to study the real importance of migration by implementing such a mechanistic estimate of dispersal into semi-spatially explicit models (215). By doing so, we not only investigate the influence of dispersal directly but also have a more objective way to study the influence of neutral processes and to distinguish between sources of betadiversity. If dispersal would be the only mechanism driving diversity, such models should be able to predict community composition to a good degree. If not, then other mechanisms must be invoked. The interesting question is how this differs between different regions, e.g. between more dynamic and slow forests such as Ecuador versus the Guyana Shield (215). A different interesting question is regarding the influence of species richness and the ratio between species richness of the metacommunity and the local communities. Here we focused on tropical forest systems as we have access to large-scale datasets to test these models. But asking similar questions across multiple scales of diversity would most likely yield even more questions on the importance of regional diversity and the size of the species pool, which may prove a significant challenge.

Our main conclusion here is that spatially implicit models mimic the real world in a very good way simply because they make us of an aggregated ecological parameter, incorporating not only dispersal but also everything determining the connection between a regional species pool and a local species pool. But the world simply is not spatially implicit; at least not for tropical tree species and we should reject all inference from such models on whether communities behave neutrally or not. Knowing this contains all possible filters that have been proposed, it does not further our knowledge of forest dynamics as we can only infer whether there is strong or weak filtering, it being either dispersal or establishment or both. Obviously, if we feed non-neutral (assuming the real world is non-neutral) data into a neutral model, models will still create output and methods for estimation of parameters will still generate parameter values. The importance, however, lies in the interpretation of these estimates. In neutral models, the emphasis lies on limited migration of individuals for explaining differences in composition. Many biologists thus interpret migration from such models as a mechanistic explanation for said differences. What we have tested here is whether this is reasonable or not and show is that it is not and that we should be careful with these interpretations. As such, either assuming neutral dynamics or not, we can not be sure what we are actually estimating from our spatially explicit world using methods based on species composition: low migration, high selective pressures, slow dynamics or fast dynamics, stronger drift, weak or strong natural selection, effects over short or long periods? The only thing we know is that we are estimating how much difference there is between the plots and the overall pool of diversity and it is unlikely this is based solely in implicit neutral dynamics.

Acknowledgements

We would like to thank Rampal Etienne and Franck Jabot for taking the time to review the mathematical approaches used in this study. Furthermore, Daniel Sabbatier and Jean-François Molino are supported by an "Investissement d'Avenir" grant managed by Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-0025).

Supporting Information

S1 Summary of used estimation methods

We focus on five different parameter estimation methods 1) Etienne's Sampling formula (184, 187), 2) the Inference method by Jabot et al (185), 3) The Gst statistic adopted from population genetics (186), 4) Etienne's two-stage sampling formula (202), which is an extension on the two-stage-estimation method by Munoz et al. (2007) (188) and 5) a method by Chisholm & Lichstein (2009) based on plot geometry and absolute dispersal distances (138). Each method is summarized briefly below.

Etienne's sampling formulae. Etienne (2005) presented a sampling formula for calculating the joint probability of species abundances in multiple local samples and its use to estimate migration. He based his work on the original spatially implicit version of the UNTB (29, 74, 184), where immigrants are drawn from a regional species pool according to one aggregated migration parameter *m* (see also figure S1). This migration parameter was transformed into the dispersal number *I*, which is related to sample size J and the migration probability *m* by *I = (m(J-1))/(1-m)* showing that *m* is also related to sample size by *m = I/((I+J-1))* (216). Considering that each local community is a sample from the larger regional species pool we can imagine species entering the local community, establishing themselves and from that point forward potentially increasing in abundance by propagating (taking up the role as ancestors for all individuals of that same species in the future time). From the moment these ancestors arrive in the local community and going forward in time, each individual belonging to a specific species gives rise to a number of descendants being of the same species (as there is only speciation in the metacommunity), ultimately resulting in the species abundance distribution in the present community (184). If we would know all these intermediate steps (i.e. the number of individuals each ancestor of a specific species would have produced) it would be possible to derive what Etienne calls the "species-ancestry–abundance" distribution (D+) (184). As this is not possible in many cases, the current observed species abundance needs to be calculated as the sum of all possible species-ancestry-abundance distributions, which in turn is given by the multivariate probability P[D|θ, m, J] of observing a specific species abundance distribution (217). This multivariate probability is the likelihood P of observing a specific species abundance distribution D (for S species, D= (n1,n2, …, nS)) under the constraint of the suit of parameters used, thus P[D|θ, m, J], for the exact derivation of this expression, see Etienne & Olff (2004)(217).

Although still rather complex, Etienne presented a simplification of this multivariate probability in 2005, which was incorporated in the freeware program Tetame <current version 2.1: http://chave.ups-tlse.fr/projects/tetame.htm>. Etienne's estimation method of 2005, however, still has the assumption that all samples share a similar dispersal limitation. In 2007, Munoz et al. proposed a two-stage estimation method developed to circumvent problems encountered by the sampling formula of Etienne when dealing with small samples (188). They first start by resampling a single individual from each separate sample (first stage) from which theta ($\theta meta$) is then estimated using Ewens' sampling formula (218). This procedure is repeated numerous times and the results averaged. The second stage of the estimation procedure is simply calculating the migration parameters I and m for each separate sample in the dataset using Etienne's sampling formula as described above with the use of $\theta meta$ as θ. However, this approach has difficulties when estimating θ in the case of having either few samples or many samples that are very different from each other (187). Later, Etienne therefore provided a renewed sampling formula as an improved two-stage estimation method, better suited for dealing with multiple samples having potentially very differing degrees of dispersal limitation (202). This "two-stage Etienne estimation method" will also be used in the comparison; the coding for estimation was made available in the PARI/GP environment (190) as appendix (202). There were, however, a few practical issues while attempting to use the coding. Only older versions of PARI/GP could be used (pers.com Rampal Etienne) and calculation time was lengthy (taking approximately 4 days per subset to compute on a standard desktop computer with an Intel Core i5-4670 3.40 GHz processor and 8Gb RAM, which in total would mean roughly 200 days for all spatially implicit and the semi-explicit datasets). This in combination with the apparent small difference between the methods after performing the calculation on part of the data lead to the choice to use this method only for estimation of the field datasets and one spatially implicit dataset based on the empirical set of Ecuador.

Inference method. Jabot et al. (2008) used a different approach by using the pooled abundances of all species over all samples as regional species abundances to infer parameters for their model, rather than a summary based on the estimation of $\theta meta$ (185). They assumed that plots are randomly placed and not stratified to particular habitats. If this would be the case, then the estimate of migration would be biased, resulting from a non-random sampling of the regional species pool. Jabot et al. consider the regional relative abundances to be fixed, however, and by doing so ignore spatial aggregation of species. Using the constructed regional species pool, a maximization of the likelihood for the parameters to be estimated is achieved by optimization of their sampling formula (185). This method is also available in the Tetame freeware program (version 2.1).

Gst statistic. Munoz et al. (2008) proposed an estimation method based on the Fst statistic from population genetics (186). Originally, the Gst statistic as proposed by Nei in 1973 infers the extent of genetic differentiation between populations by comparing the variation in alleles among populations (i.e. betadiversity) and overall allelic diversity at the specific locus (gamma diversity). Nei's genetic distance measure (219, 220) has assumptions comparable to Hubbell's NT: all loci have the same rate of neutral mutation, there is a stable effective population size and this population is in a mutation versus drift equilibrium. Munoz et al. (2008) made a similar approach as Hubbell (29) using theory from population genetics to estimate parameters for the neutral model. A very large set of alleles from a single locus is comparable to the number of species in large area. As such, the Gst statistic can be viewed as a measure of the variation among samples not in allelic diversity, but in species diversity. To estimate migration, three estimators of similarity are used: Fintra as the probability that two individuals in a sample (k) are conspecific, Finter that two individuals one from sample k and one from sample l are conspecific and Fglobal that two individuals samples from the larger species pool are conspecific (186). The average intrasample similarity ($\tilde{F}_{intra}$) over all samples is related to Simpsons alpha diversity (221) as Div_{α} = 1- $\tilde{F}_{intra}$, F_{global} is related to gamma diversity in a similar fashion as $\tilde{F}_{intra}$ is to Div_{α} by Div_{γ} = 1-F_{global} (186) and lastly, betadiversity as Div_{β} = $\tilde{F}_{intra}$ - Fglobal. These similarity statistics together form the Gst statistic for each separate sample k by:

$$G_{st}(k) = \frac{F_{intra}(k) - F_{global}(k)}{1 - F_{global}(k)}$$

This relationship measures the relative dissimilarity between a sample and the regional species pool, which is considered to be the sum of all samples. Munoz et al. (2008) show that after derivation, the Gst(k) is dependent only on the migration number I(k) when the sum of all individuals is much larger than that of a single sample. With this it is possible to estimate I(k) from each sample, under the assumption, however, that these samples have enough distance between them to actually be separate distinct local communities. For a detailed analysis of the exact and approximate estimators (sampling without and with replacement) and the relationship of the average betadiversity to similarity relative to a specific sample k (186).

Chisholm & Lichstein' 2009 plot geometry method. All of the above described estimation methods lack a biological interpretation of the migration parameter. Chisholm and Lichstein (2009) presented expressions related to actual dispersal kernels and plot geometry for approximating the amount of migration, making it a more "biological intuitive" method (138). The main result of their efforts is their expression (2): $m \approx Pd/\pi A$, with P as the perimeter of the plot in meters, d the actual mean absolute dispersal distance in meters and A the total surface area of the plot in square meters. This expression holds only for large plots, as long as the size of the plot is relatively large in comparison to the mean dispersal distance it can be applied. They also show this for the case of a square plot and bivariate Gaussian kernel, yielding their expression (3):

$$m = 1 - \left\{ erf\left(\frac{w\sqrt{\pi}}{2d}\right) - \left(1 - exp\left[-\frac{\pi w^2}{4d^2}\right]\right)\frac{2d}{\pi w}\right\}^2$$

Where erf is an error function and w is the edge of the plot. As the size of the plot increases relative to the dispersal distance d the erf and exp function tend to 1 and 0 respectively, resulting in the same expression as (2). As example, for a square plot with an edge length of 100 meters and mean dispersal distance of 30 meters, the error function becomes .999 and the exponent becomes 1.6e^{-4}, thus becoming negligible and resulting in the general solution of expression (2) (Chisholm & Lichstein 2009). In order for the erf function to be >.99 and the exp function to be <.025 the ratio [d/w] should be at least < .45 meaning that the edge of the plot size should roughly be at least twice the mean dispersal distance. For this study, when this was the case expression (2) was used to approximate m, otherwise expression (3) was used for approximating the migration parameter. This method does suffer from a serious drawback: it only takes into account the number of immigrants arriving in the local community from outside of a plot as potential replacements. Recruits, however, can come from inside the local community as well. When for instance the replacement falls on the edge of a plot, assuming a Gaussian dispersal kernel, the probability distribution is symmetric. To solve this, we used a corrected version of the Plot Geometry method; see S3.

S2. Stabilization time and plot size of simulations

Two major components determine whether simulations stabilize in terms of their species abundance distribution: size of the plot and runtime of the simulation. To determine both we made use of Fisher's Logseries distribution (98). To test whether time and plot size was adequate, a simulation was run with and without dispersal limitation. As the local communities start out as a random sample of the metacommunity, in the latter case they reflect the composition of the metacommunity. The number of singletons and Fisher's alpha is then expected to be almost identical to that of the metacommunity, given a large enough sample due to random variation of the sampling procedure. According to Fisher's logseries the total number of individuals (NT) found in any community can be calculated by $NT = S_1 + 2S_2 + 3S_3 + 4S_4 + iS_i$ with S_i being the abundance class a species belongs to (i.e. S_1 are all singletons, S_2 all doubletons etc.). In terms of Fisher's parameters α and x this is equal to $Nt = \alpha x + \alpha x^2 + \alpha x^3 + \alpha x^4 + \alpha x^i$, where αx is the number of species predicted to have only 1 individual (singletons), $\alpha x^2/2$ the number of doubleton species etc. Assuming x is a value between 0 and 1, Fisher's logseries approaches a geometric series given by *1/(1-x)*. Nt then becomes $Nt = \alpha x(1/(1-x)) = \alpha x/(1-x)$. From this we can derive the value of x using $NT\ (1-x) = \alpha x$, which leads to $NT = x(\alpha+NT)$ and finally to $x = NT/(\ NT+\alpha)$. By plugging this into the first term of the logseries we come up with the equation for the expected number of singletons in the community, $\Phi 1_{exp} = \alpha x = \alpha NT/(NT+\alpha)$. By calculating the expected number of singletons ($\Phi 1_{\mathrm{exp}}$) for any given community given the total community size and comparing this with the actual number of singletons observed of the sample ($\Phi 1_{\mathrm{obs}}$) in the case without dispersal limitation we can test whether sample size was adequate. To test generation time for scenario's with dispersal limitation we simulated a set of 25 local communities, each having 625 individuals starting out as a random sample of the metacommunity. Migration was set to an arbitrary value of 0.5, meaning half of replacements is coming out of the metacommunity and half are local recruits (as this is a spatially implicit model). After the initial sampling, at some point local communities running neutral dynamics should stabilize in terms of the species abundance distribution and the difference between Fisher's alpha of the previous generation and the next generation should become negligible (Figure S3).

Simulated metacommunities and stabilization of the sampling procedure.
Construction of the MC-Low yielded 20,191,600,511 individuals distributed over 4,582 species with a Fisher's alpha of 251. For the MC-High this yielded 5,611,001,426 individuals belonging to 6,834 species, with a Fisher's alpha of 416 (Figure S2). Analysis of the stabilization for sampling each of the simulated spatially implicit datasets showed that approximately 625 individuals and $1e^{5}$ sampling rounds was sufficient to minimize the difference between $\Phi 1_{\mathrm{exp}}$ and $\Phi 1_{\mathrm{obs}}$ (Figure S3).

S3 A comment on the plot geometry multi subplot calculation

According to Chisholm & Lichstein (2009), their Plot Geometry method also allows estimation of migration for disconnected local communities (138). As the direct dispersal between distant subplots becomes negligible, approximately at twice the mean dispersal distance according to their simulations, they state the migration parameter can be calculated by

$$m = \sum_{i=1}^{K} P_i \widehat{m}_i$$

with K the number of subplots, p_i the probability of death occurring in subplot i and $\widehat{m}_i$ the probability that the parent of the replacement has an origin outside this or any subplot in the total community. Because direct dispersal between subplots at a certain distance becomes negligible, migration for each subplot can be approximated by assuming each plot is completely isolated and therefore estimation can be performed by using expression 3 of Chisholm & Lichstein (2009) which states:

$$m = 1 - \left\{ erf\left(\frac{w\sqrt{\pi}}{2d}\right) - \left(1 - exp\left[-\frac{\pi w^2}{4d^2}\right]\right)\frac{2d}{\pi w}\right\}^2$$

In addition, because under the assumptions of the UNTB (29) the probability of a death of a particular individual is simply proportional to the relative abundance of the species it belongs to in that subplot, the approximation for migration becomes:

$$m = \frac{1}{J}\sum_{i=1}^{K} J_i m_i$$

They continue by stating that when there are many disconnected subplots, the assumption of panmixis of the original UNTB might be a potential violation but that despite this shortcoming, their method still remains useful as instead of simulating a single migration parameter you could calculate it for each separate subplot. However, we fail to see the added value of this last statement, as

$$\sum J_i = J \ \text{ and } \ \frac{\sum J_i}{J} = 1$$

m reduces to $m=m_i$, meaning that assuming the plot geometry and mean dispersal distance for each plot is constant, one could simply calculate it for one plot and it would be the same for all plots. Only in the case where each subplot has different geometry or mean dispersal distance this has merit.

A correction for the plot geometry method. Chisholm & Lichstein (2009) proposed an estimation method based on simple plot geometry and the mean dispersal distance of individuals. They assume an infinite two-dimensional landscape on which a quadrat of area A is "thrown down", being the local community (LC). The LC then consists of J individuals, calculated as the density of individuals in the infinite landscape (ρ) captured within the area A. Similar to Hubbell's UNTB (29), individuals within the LC die and are being replaced at random from either parents within the LC (local recruitment) or from outside of the LC (migrants). Each individual, either within or outside the local community is thus capable of producing offspring. In contrast with the original UNTB, Chisholm & Lichstein incorporate space by assuming the offspring of each parent has the ability to disperse according to a radially symmetric dispersal kernel, assuming there is no difference between individuals. Imagine the LC is divided into *xn* by *yn* gridcells of 1 m^2 with *n* the edge length of the area *A*. Each cell is occupied by an individual and each timestep an individual dies, vacating one of the grid cells with coordinate (*xi, yi*). Chisholm & Lichstein consider the location of the parent individual for the replacement randomly selected from a dispersal kernel, which is centered at the dying individual. It then is the mean dispersal distance of individuals, in combination with the location (*xi, yi*) that determines the overall probability of a replacement coming from either inside or outside the local community.

For replacements near the edge of the plot area A, this latter probability is the largest considering half of the kernel is situated outside of the plot, rather then inside (figure S4A). They justify using the replacement location as the center considering the kernels are symmetrical from either perspective and the result will be the same when you consider it from the parent perspective. Because they calculate it from the perspective of selecting the parent from the dispersal kernel centered at the replacement location, they only take into account a small fraction of potential recruits from both within and outside of the plot. In this case indeed, swopping from either the parent or the replacement perspective does not matter because they are symmetrical (figure S4B,C). The consequence however also is that the probability kernel only is calculated with a maximum mean μ as the edge of the plot (L in the figures S4A-D), because a replacement location cannot be outside of the surface area A. If we, however, look from the perspective of the potential parents, considering a dispersal kernel with a mean dispersal distance (σ) of 20 meters the entire probability density curve has a range of approximately +-σ4 from the edge of the plot (figure S4D, in this case with a mean dispersal distance of 20 meters this would be the edge +- 80 meters). This in turn means that parents within the range of L+- 4σ can potentially supply the replacement (figure S4D). We therefore simulated all dispersal kernels for each meter in the range L+-4σ and calculated the entire surface area's of the probability density kernels coming from both inside the plot (figure S4D, red) and outside (blue) and divided the surface area of the blue density kernels by the total, yielding the ratio of replacements being a migrant from outside the plot. Disregarding the mean dispersal distance σ, this is always approximately 30%, meaning that from assuming an individual every meter, the ratio of migrants of the total estimated migration by the Plot Geometry method should be multiplied with a correction of .3. As is shown in the main text, using this corrected migration parameter to simulate a spatially implicit forest shows this has extremely good fits with the actual field data.

S4 Further analysis of migration using Fishers Log series

Although estimation methods were only able to estimate joint migration rate and only accurately for the limitation of *m.meta>m.adj*, we did observe an important pattern in the amount of singletons per plot and in the total simulated forest. The amount of singletons was strongly dependent on the amount of migration from either source, adjacent or the metacommunity. An increase in the probability of migration from the metacommunity was being reflected in a relative high number of singletons. In contrast, this was relatively low when migration was mostly coming from adjacent plots. Interestingly, this pattern is comparable to the amount of rare alleles found in population genetics in relation to genetic drift, which is also dependent on the amount of migration. Disregarding selective pressures, when populations experience little to none migration of individuals, they will eventually become fixated for specific alleles due to sampling error. Following a similar train of thought, due to the probabilistic nature of sampling in neutral theory models, singletons are lost and are not being replaced from the metacommunity but by the more common species from adjacent plots. In this case, the more common species increase in abundance as they are shared more and more among adjacent plots, hence the amount of singletons will decrease. We can use this pattern to shed light on whether much migration is coming from adjacent localities or from the main source pool of species. As communities are structured according to Fisher's Logseries (98), we can calculate the expected number of singletons ($\Phi1_{exp}$) based on the total number of species and individuals and compare this with the observed number of singletons in the samples ($\Phi1_{obs}$) if forests follow true neutral dynamics. A low singleton ratio of $\Phi1_{obs}/\Phi1_{exp}$ is expected when most migrants are from adjacent plots (leading to clumped patterns of species composition). When the ratio approaches 1 this indicates that most migrants are originating from the metacommunity, ultimately representing a homogeneous random sample of the metacommunity when all migrants actually do originate from the metacommunity. We tested this for all combinations of migration used in the spatially semi-explicit simulations and indeed found that a higher ratio was indicative of a higher migration probability from the metacommunity (Figure S7). We also applied this to the field data. For each dataset, $\Phi1_{exp}$ was calculated for each plot and compared with $\Phi1_{obs}$. Results showed that for all datasets there was on average a higher than 1 ratio of observed versus expected singletons (Figure S8 and Table S2). This means that on average, plots have more singletons than would be expected purely on the logseries and neutral dynamics. As is explained in the main text, this might be explained by the fact that these forests indeed experience more than just neutral dynamics. Perhaps there is strong selection forcing species on their way out more than we would expect on the basis of stochastic processes alone or differential selection might allow specialists to persevere in low numbers.

Table S1. Estimates of migration from spatially implicit simulated datasets based on the MC-low and MC-high metacommunities. For each dataset created, the runtime and number of plots are indicated. Size was set on 625 individuals per plot and runtime was 1e^5 generations, based on stabilization time of the sampling procedure as explained in the main text. Species richness depends on both runtime and migration. Estimated migration parameters are shown with standard deviation of the mean calculated for all plots.

GS based (MC-Low, 67 plots)

Simulation				Estimated Migration					
				One-stage est.		Inference method		G_{st}-statistic	
dataset	Nr. sp.	Nr. sing	m	m1	SD	m2	SD	m3	SD
1	1062	131	.21	.231	.032	.219	.015	.211	.043
2	1100	151	.25	.250	.032	.255	.015	.247	.059
3	1130	163	.31	.302	.036	.314	.015	.297	.077
4	1155	157	.35	.294	.034	.359	.016	.341	.095
5	1166	164	.41	.433	.048	.415	.015	.394	.120
6	1199	172	.45	.443	.051	.458	.015	.430	.141
7	1213	185	.51	.522	.060	.515	.015	.487	.170
8	1220	196	.55	.553	.054	.553	.014	.521	.198
9	1246	211	.61	.543	.047	.628	.013	.574	.216
10	1253	201	.71	.586	.046	.717	.011	.670	.289
11	1259	211	.81	.651	.047	.820	.009	.780	.382

FG based (MC-Low, 63 plots)

Simulation				Estimated Migration					
				One-stage est.		Inference method		G_{st}-statistic	
dataset	Nr. sp.	Nr. sing	m	m1	SD	m2	SD	m3	SD
1	1077	141	.21	.238	.033	.216	.015	.209	.044
2	1081	145	.25	.287	.034	.254	.015	.247	.0459
3	1109	140	.31	.313	.039	.320	.016	.311	.0770
4	1143	166	.35	.339	.042	.351	.016	.327	.075
5	1164	158	.41	.427	.052	.413	.015	.386	.099
6	1152	177	.45	.440	.049	.461	.015	.429	.125
7	1169	182	.51	.478	.058	.530	.015	.489	.153
8	1209	162	.55	.557	.058	.551	.014	.509	.169
9	1212	193	.61	.578	.060	.624	.013	.588	.220
10	1241	213	.71	.542	.043	.712	.012	.653	.257
11	1254	224	.81	.375	.059	.824	.009	.760	.325

EC based (MC-High, 72 plots)

Simulation				Estimated Migration							
				One-stage est.		Inference method		G_{st}-statistic		2-stage (Etienne)	
dataset	Nr. sp.	Nr. sing	m	m1	SD	m2	SD	m3	SD	m4	SD
1	1589	205	.21	.280	.025	.2154	.014	.205	.043	.204	.017
2	1659	244	.25	.264	.030	.254	.014	.241	.056	.245	.022
3	1699	249	.31	.469	.053	.313	.014	.296	.075	.308	.022
4	1718	266	.35	.528	.055	.359	.014	.336	.096	.344	.025
5	1736	256	.41	.620	.060	.419	.013	.399	.132	.417	.030
6	1776	297	.45	.648	.061	.460	.013	.435	.157	.449	.036
7	1803	308	.51	.665	.060	.510	.013	.472	.173	.496	.029
8	1804	283	.55	.712	.053	.556	.013	.525	.212	.547	.040
9	1852	336	.61	.712	.046	.618	.018	.587	.250	.610	.041
10	1891	368	.71	.729	.036	.724	.010	.695	.326	.713	.046
11	1922	371	.81	.625	.022	.819	.008	.808	.418	.810	.045

Table S2. Results belonging to figure S7 with mean $\Phi1_{obs}$ and $\Phi1_{exp}$ for all plots of the datasets Guyana/Suriname (GS; 67), French Guiana (FG; 63) and Ecuador (EC; 72).

	Mean FA	*Mean Nr. species*	*Mean $\phi_1$obs*	*Mean $\phi_1$exp*	*Ratio*	*Sd. ratio*
GS	20	58	22	19	1.198	.33
FG	67	142	69	59	1.169	.11
EC	78	146	75	68	1.115	.17

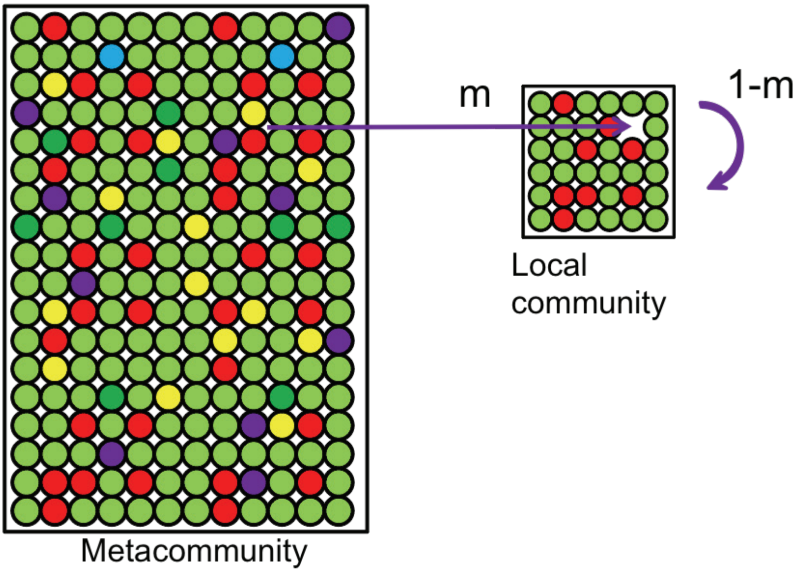


Figure S1. Schematic view of the two spatio-temporal communities and their connection as implemented in the original Unified Neutral Theory of Biodiversity and Biogeography by Hubbell (29)

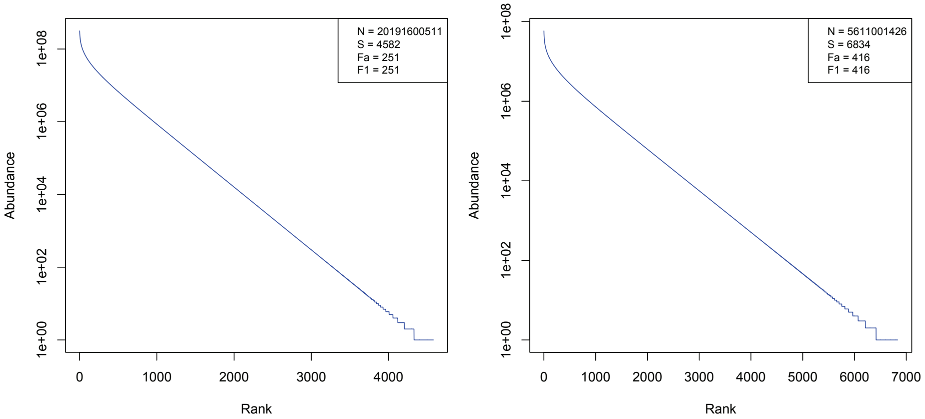


Figure S2. Rank abundance curve for both MC-low (left) and MC-high (right). Numbers are based estimates for the number of individuals and species from (14).

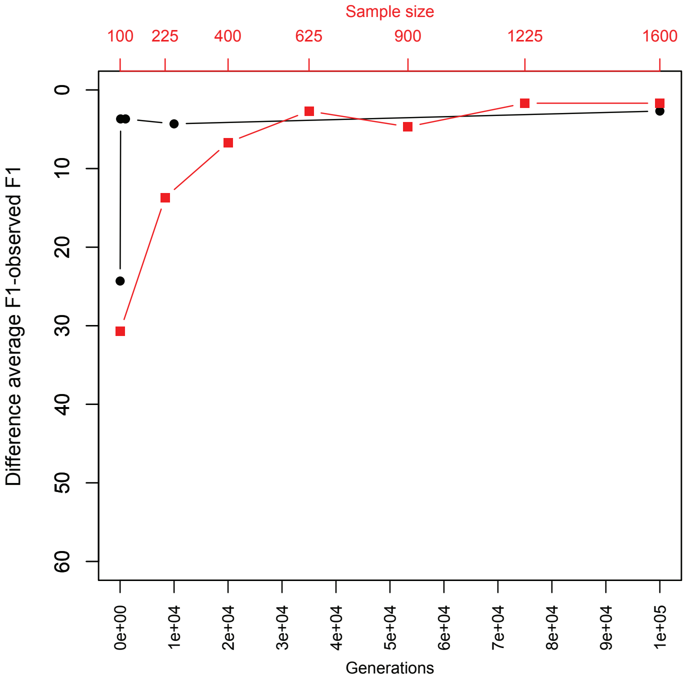


Figure S3. Stabilization time of the sampling model without dispersal limitation (*m*=1) based on both sample size (x1 axis above, red label) and number of generations (x2 axis below, black label). On the y-axis is the difference between the expected number of singletons (F1) and observed F1, which was calculated using Fisher's Logseries (98).

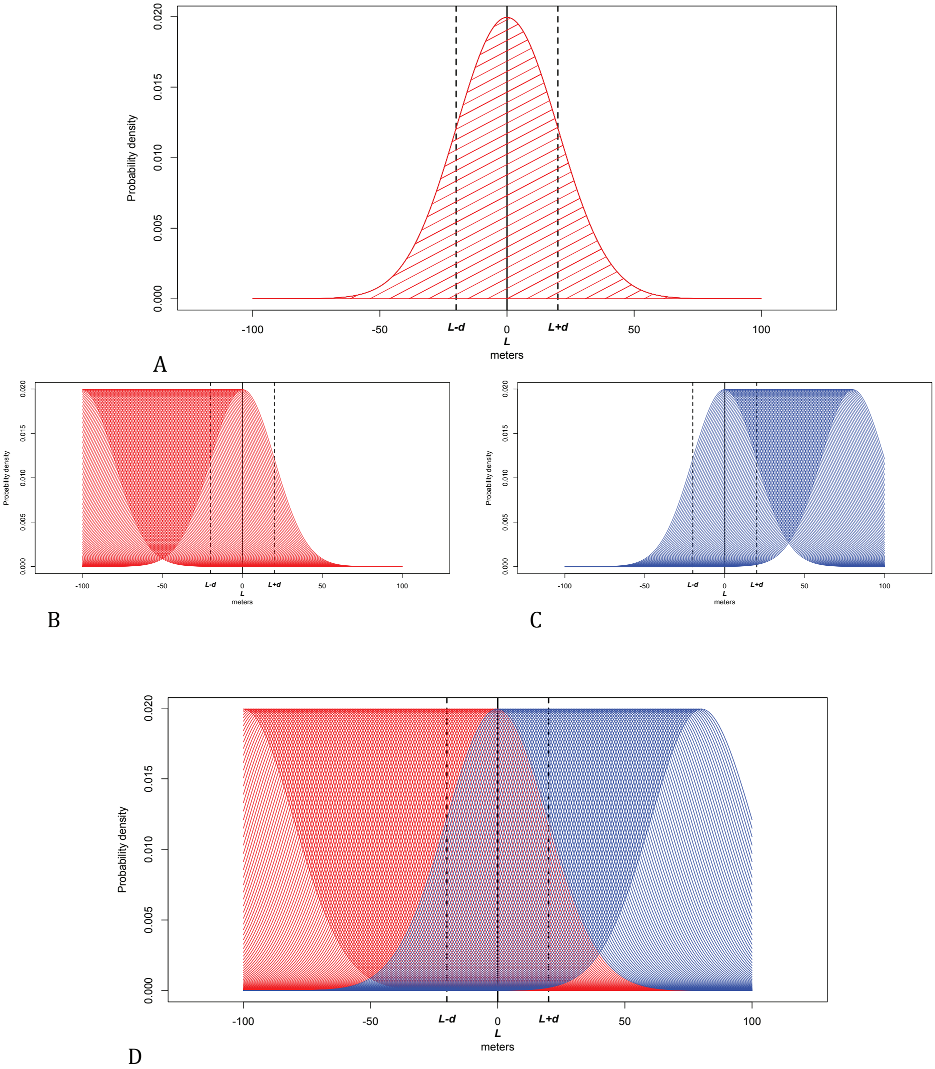


Figure S4. Different situations of the distribution of Probability Density Kernels following a Guassian Distribution with L the edge of a plot and d the mean dispersal distance used as the standard deviation (σ) of the distribution and range given by L +- 4σ to ensure a full range (as 3σ covers 99.73% of the values around the mean). A) A single dispersal kernel for propagules produced by a parent on the edge of the plot, B) All dispersal kernels from parents residing within the plot, C) All dispersal kernels from parents residing outside the plot and D) The combined plot of B and C.

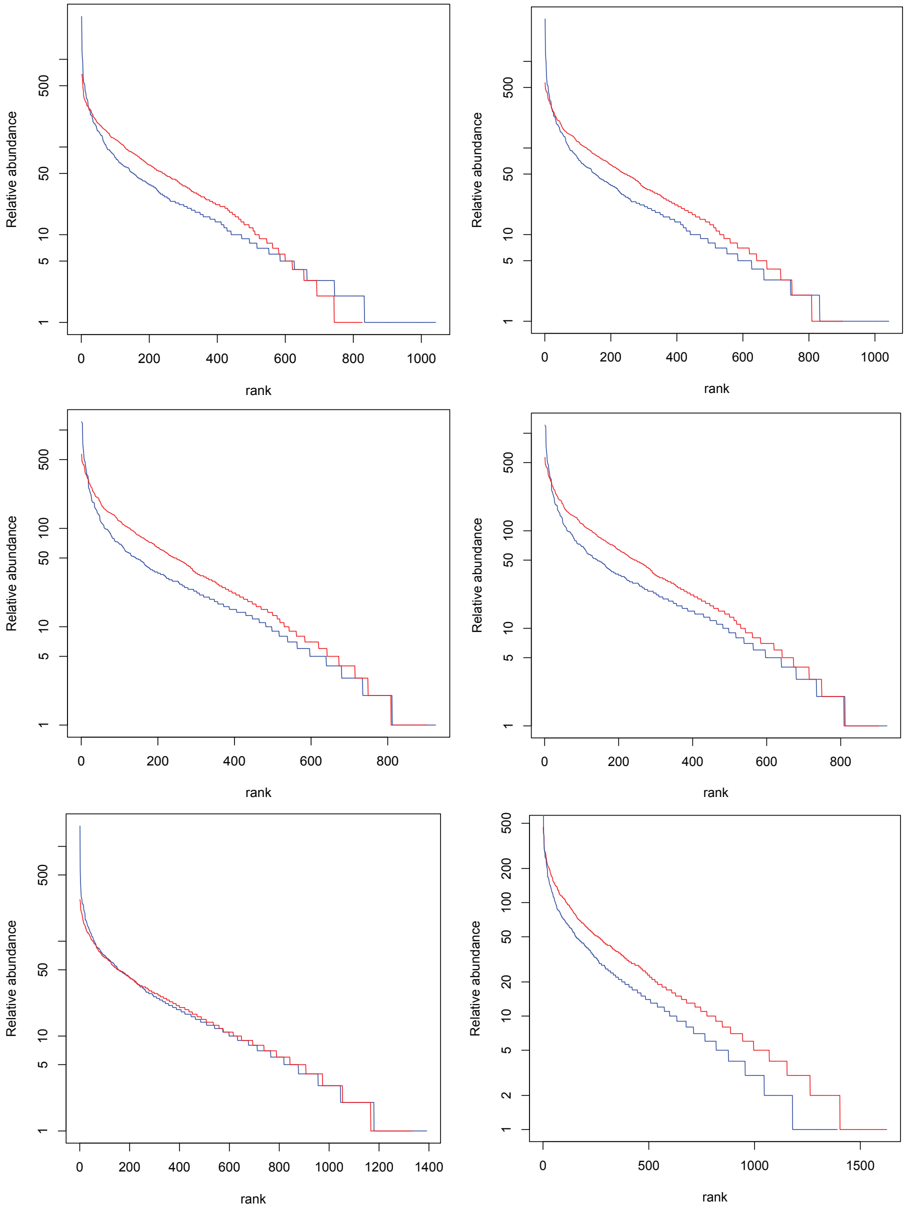


Figure S5. Rank abundance curves (Guyana/Suriname top, French Guiana middle, Ecuador bottom) for the simulated datasets (red) in comparison with those of the actual field data plotted in the same graph (blue). Migration parameter used for sampling the simulated sets is based on the Gst statistic (left) and Inference method (right).

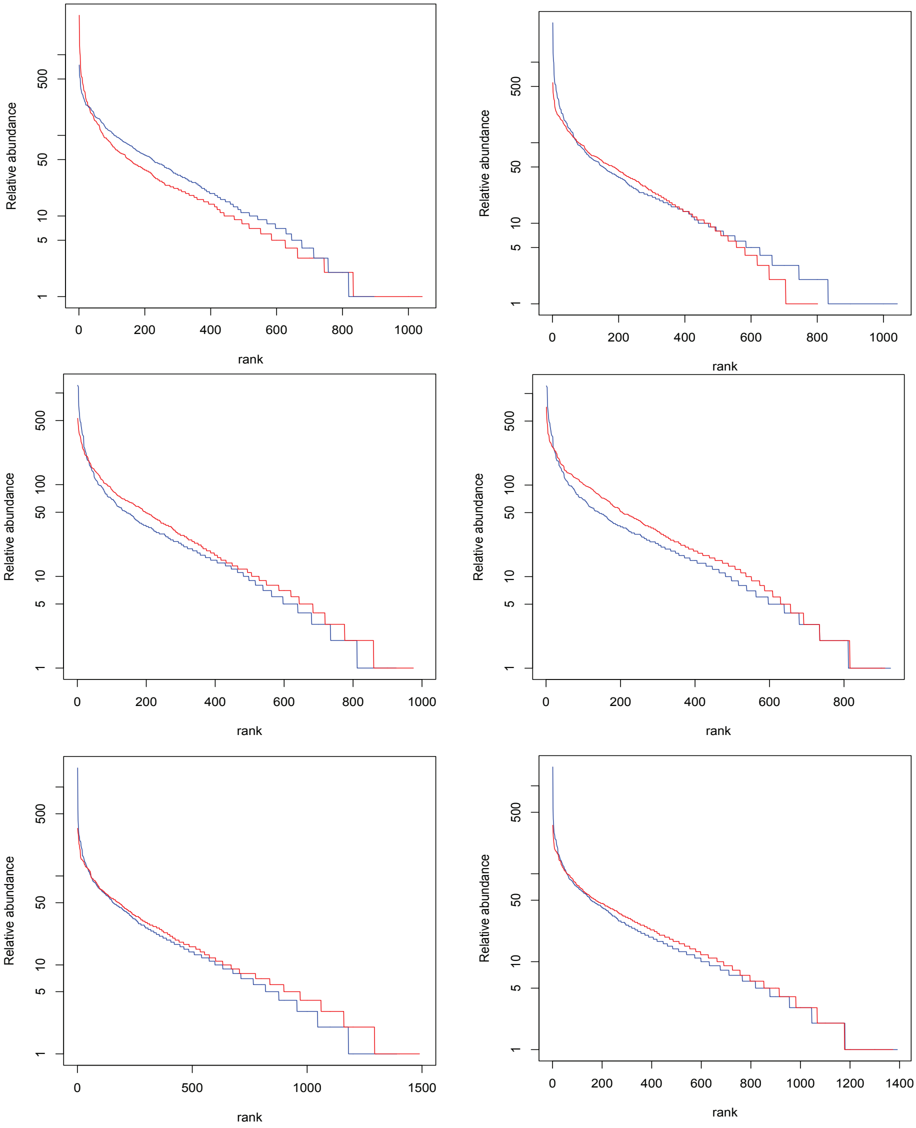


Figure S6. Rank abundance curves (Guyana/Suriname top, French Guiana middle, Ecuador bottom) for the simulated datasets (red) in comparison with those of the actual field data plotted in the same graph (blue). Migration parameter used for sampling the simulated sets is based on Etienne's two-stage estimation method (202) (left) and the (corrected) Plot Geometry method from Chisholm & Lichstein (138) (right).

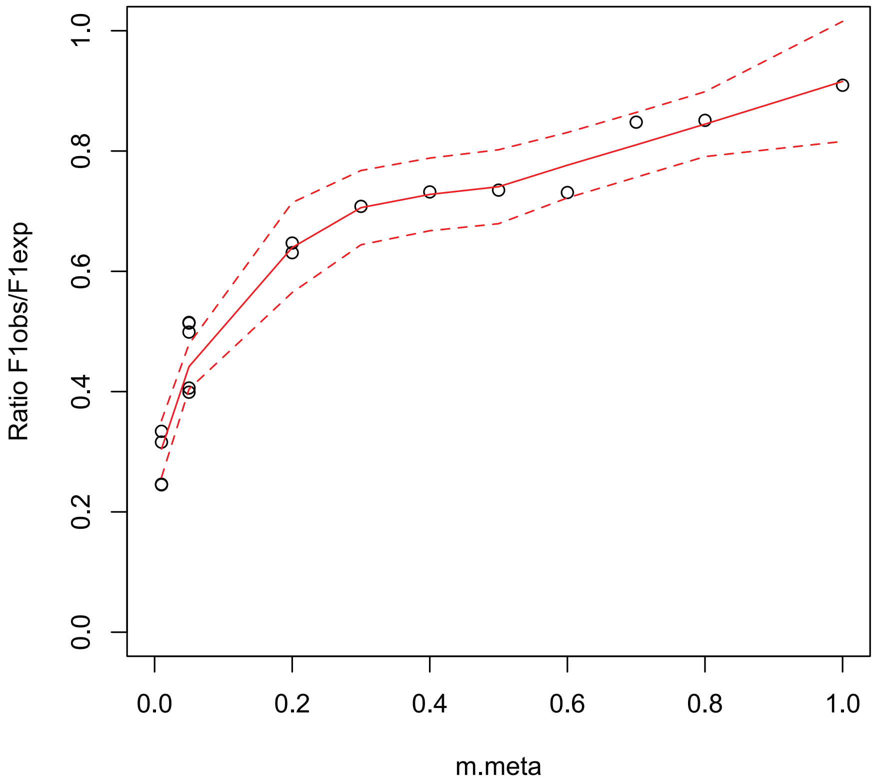


Figure S7. The calculated ratio of $\Phi 1_{obs}/\Phi 1_{exp}$ plotted against the given migration probability from the metacommunity (*m.meta*) in the semi spatially explicit model as described in the Supporting Information. Dashed lines indicate the 95% confidence interval for the loess model.

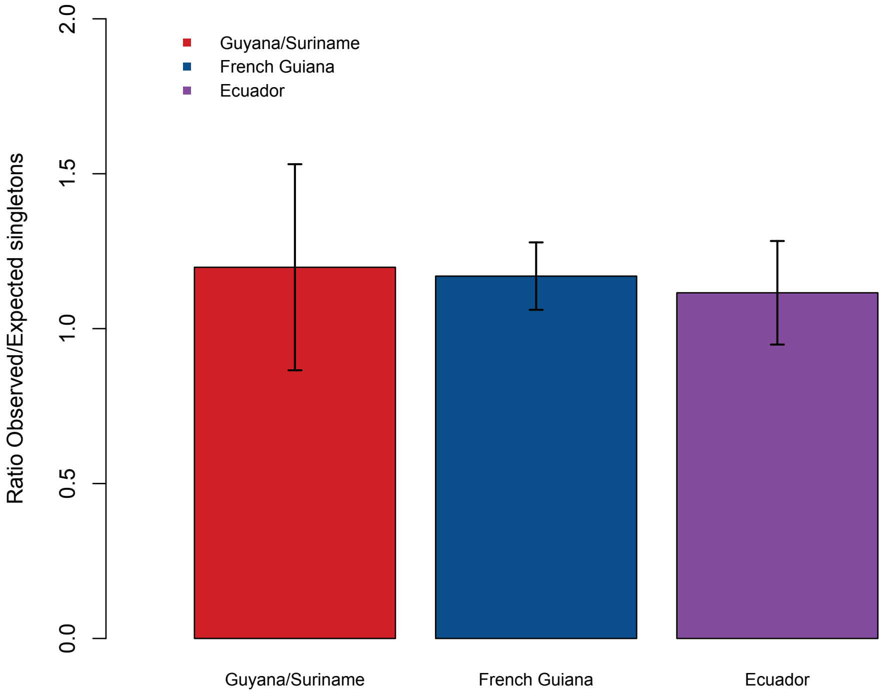
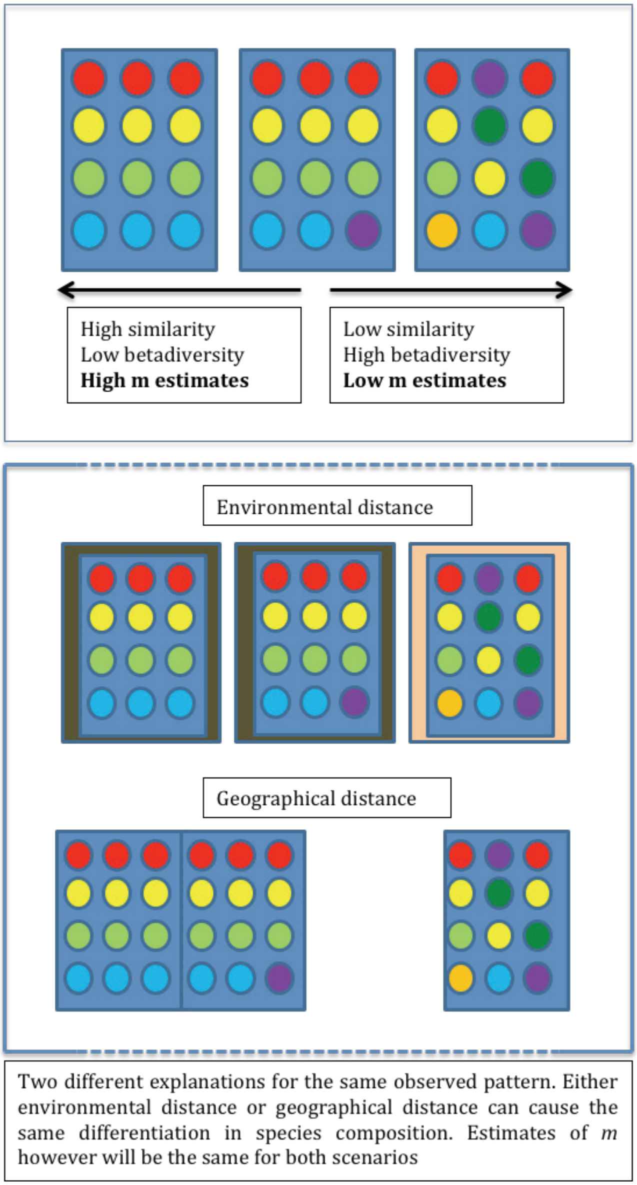
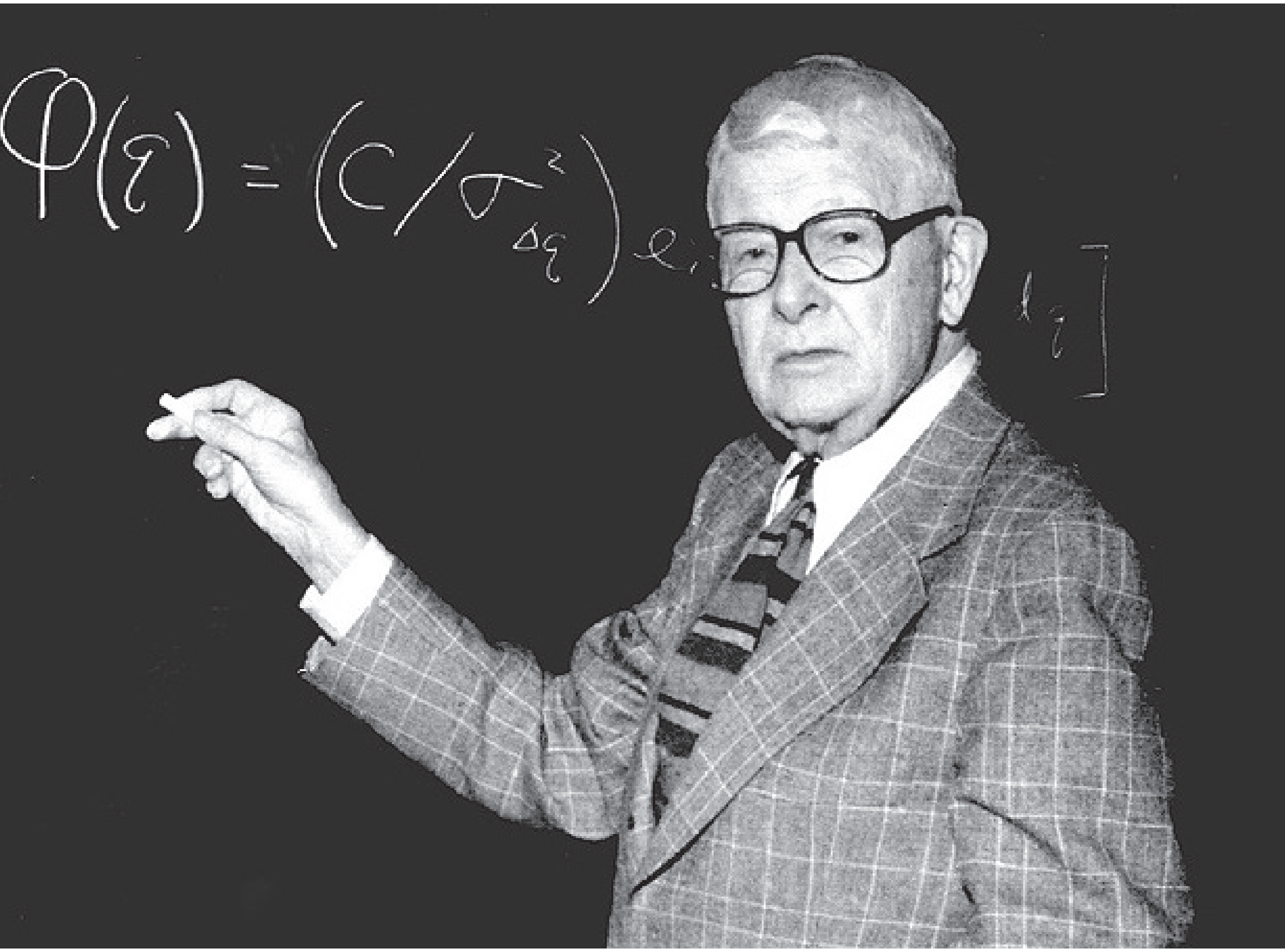


Figure S8. Calculation of the mean ratio of observed singletons versus the expected amount of singletons per field dataset as explained in the main text of the Supporting Information. Error bars shown are the standard deviations of the mean.



Two different explanations for the same observed pattern. Either environmental distance or geographical distance can cause the same differentiation in species composition. Estimates of *m* however will be the same for both scenarios

Figure S9. Schematic view of how both environmental distance and geographical distance can cause similar patterns in differentiation of species composition. Although these two mechanisms are each others opposite in terms of the niche versus neutral discussion, spatially implicit neutral models cannot differentiate between the two mechanisms responsible.



The Darwinian process of continued interplay of a random and a selective process is not intermediate between pure chance and pure determinism, but qualitatively utterly different from either in its consequences.

Sewall Wright (1889 - 1988)

Chapter Six

Adding biological reality to general predictions of neutral theory reveals scaling issues between local and regional patterns of diversity

(Currently under Review)

Edwin Pos[1,2], Juan Ernesto Guevara[3,4], Jean-François Molino[5], Daniel Sabatier[5], Olaf S. Bánki[6], Nigel C.A. Pitman[7], Hugo F. Mogollón[8], Roosevelt García-Villacorta[9,10], David Neill[11], Oliver L. Phillips[12], Carlos Cerón[13], Marcos Ríos Paredes[14], Percy Núñez Vargas[15], Nállarett Dávila[16], Anthony Di Fiore[17], Gonzalo Rivas-Torres[18,19], Raquel Thomas-Caesar[20], Corine Vriesendorp[7], Kenneth R. Young[21], Milton Tirado[22], Ophelia Wang[23], Rodrigo Sierra[22], Italo Mesones[3], Roderick Zagt[24], Rodolfo Vasquez[25], Manuel Augusto Ahuite Reategui[26], Walter Palacios Cuenca[27], Elvis H. Valderrama Sandoval[28,29], Hans ter Steege[2,30]*

[1]Ecology & Biodiversity Group, Utrecht University, Padualaan 8, Utrecht, 3584 CH, The Netherlands
[2]Biodiversity Dynamics, Naturalis Biodiversity Center, PO Box 9517, Leiden, 2300 RA, The Netherlands
[3]School of Biological Sciences, Yachay Tech, Hacienda San José s/n, San Miguel de Urcuquí, Ecuador
[4]Keller Science Action Center, The Field Museum, 1400 S. Lake Shore Drive, Chicago, IL 60605-2496, USA
[5]AMAP, IRD, Cirad, CNRS, INRA, Université de Montpellier, TA A-51/PS2, Bd. de la Lironde, Montpellier, 34398, France
[6]Naturalis Biodiversity Center, PO Box 9517, Leiden, 2300 RA, The Netherlands
[7]Science and Education, The Field Museum, 1400 S. Lake Shore Drive, Chicago, IL, 60605-2496, USA
[8]Endangered Species Coalition, 8530 Geren Rd., Silver Spring, MD, 20901, USA
[9]Institute of Molecular Plant Sciences, University of Edinburgh, Mayfield Rd, Edinburgh, EH3 [5]LR, UK
[10]Royal Botanic Garden Edinburgh, 20a Inverleith Row, Edinburgh, Scotland, EH3 5LR, UK
[11]Ecosistemas, Biodiversidad y Conservación de Especies, Universidad Estatal Amazónica, Km. 2 1/2 vía a Tena (Paso Lateral), Puyo, Pastaza, Ecuador
[12]School of Geography, University of Leeds, Woodhouse Lane, Leeds, LS2 9JT, UK
[13]Escuela de Biología Herbario Alfredo Paredes, Universidad Central, Ap. Postal 17.01.2177, Quito, Pichincha, Ecuador
[14]Servicios de Biodiversidad EIRL, Jr. Independencia 405, Iquitos, Loreto, 784, Perú
[15]Herbario Vargas, Universidad Nacional de San Antonio Abad del Cusco, Avenida de la Cultura, Nro 733, Cusco, Cuzco, Peru
[16]Biologia Vegetal, Universidade Estadual de Campinas, Caixa Postal 6109, Campinas, SP, 13.083-970, Brazil
[17]Department of Anthropology, University of Texas at Austin, SAC 5.150, 2201 Speedway Stop C3200, Austin, TX, 78712, USA
[18]Colegio de Ciencias Biológicas y Ambientales-COCIBA & Galapagos Institute for the Arts and Sciences-GAIAS, Universidad San Francisco de Quito-USFQ, Quito, Pichincha, Ecuador
[19]Department of Wildlife Ecology and Conservation, University of Florida, 110 Newins-Ziegler Hall, Gainesville, FL, 32611, USA
[20]Iwokrama International Programme for Rainforest Conservation, Georgetown, Guyana

[21] *Geography and the Environment, University of Texas at Austin, 305 E. 23rd Street, CLA building, Austin, TX, 78712, USA*
[22] *GeoIS, El Día 369 y El Telégrafo, 3° Piso, Quito, Pichincha, Ecuador*
[23] *Environmental Science and Policy, Northern Arizona University, Flagstaff, AZ, 86011, USA*
[24] *Tropenbos International, Lawickse Allee 11 PO Box 232, Wageningen, 6700 AE, The Netherlands*
[25] *Jardín Botánico de Missouri, Oxapampa, Pasco, Peru*
[26] *Medio Ambiente, PLUSPRETOL, Iquitos, Loreto, Peru*
[27] *Herbario Nacional del Ecuador, Universidad Técnica del Norte, Quito, Pichincha, Ecuador*
[28] *Department of Biology, University of Missouri, St. Louis, MO, 63121, USA*
[29] *Facultad de Biologia, Universidad Nacional de la Amazonia Peruana, Pevas 5ta cdra, Iquitos, Loreto, Peru*
[30] *Systems Ecology, Free University, De Boelelaan 1087, Amsterdam, 1081 HV, Netherlands*

Abstract

Neutral models of ecology are often used as null models, testing the relative importance of niche versus neutral processes in shaping diversity. Most versions of neutral models, however, focus only on regional scale predictions and neglect local level contributions. Using a semi-spatially explicit neutral model and a unique dataset from the most species-rich forest region on Earth, we add biological reality to general predictions of neutral theory by combining regional and local-level perspectives and testing the scalability of predictions. We find that accurate simultaneous predictions of both regional and local patterns are not attainable. Specifically, predictions of patterns in species dominance at local levels while maintaining regional species richness are not feasible. We show that although there are clear relationships between species composition and both spatial and environmental distances, there is also a clear differentiation of species able to attain dominance with and without restriction to specific habitats. We hypothesize that lack of ecological equivalence accounts for this failure of scaling predictions either up or down and that this is just as likely due to competitive differentiation in terms of tolerance to pathogens or herbivores, regardless of adaptation to specific habitats, as it is to competitive exclusion related to specific evolved niche differentiation.

Significance statement

Lack of attempts at testing the scalability of predictions from neutral theory has left the scientific ecological community wanting for a more biologically realistic testing. Using a novel modeling approach and a unique tree inventory dataset of over 200 hectares we show that neutral theory suffers severe scaling issues, unable to attain accurate local and regional predictions simultaneously. Our results are not only a clear indication that non-neutral processes other than dispersal limitation must be at work, at least at the local level, but are also a clear warning regarding the use and interpretation of neutral models, both previous and newly developed.

Introduction

Why are some species dominant and others rare? Posed by Charles Darwin, this question remains among the most important in ecology (13) and its answer frames our fundamental understanding of community assembly. Classical Hutchinsonian ecology emphasizes deterministic processes based on niche-thinking and environmental heterogeneity. Neutral theory (NT), emerging from the theory of Island Biogeography (53), lottery models (62) and earlier work by population geneticists (64), argues for stochastic processes and environmental stochasticity. The goal of the latter is not to test the hypothesis that species do not differ or that interactions play no role at all. Instead it was put forward as a null model to test if these interactions and differences between species matter to the assembly of ecological communities. In a way it is similar to the Hardy Weinberg theorem in population genetics (222, 223), testing assumptions regarding the evolution of populations. NT likewise tests assumptions regarding the dynamics of communities. The first neutral models of ecology were spatially implicit, with recruitment from either within a local community or from a metacommunity (Supporting Information (SI) Chapter 6: Fig. S1A). These models fail, however, to correctly estimate migration from a spatially explicit world (139), even though they generate accurate predictions of community structure. Considering the overwhelming evidence that migration is in all probability very important (74, 156), spatially explicit models were developed to study the relative importance of migration and neutral processes in determining community structure (70, 173, 224, 225). These models generated good predictions for species Rank Abundance Distributions (RADs) and Species Area Relationships (SPARs) (29) but focused only on explaining such patterns at regional scales. A lack of previous attempts to combine both regional and local scale predictions has prevented a proper validation of fundamental predictions of NT. Here we combine regional and local results of a neutral spatially semi-explicit model (139) using parameters based on species characteristics to test if there is a biologically sound prediction at regional scales, following from accurate predictions at local scales. If migration is the main process determining community structure (reflecting mainly the neutral perspective), our model should approach empirical data accurately both at regional and local scales. If, however, model results deviate substantially from empirical data on either scale, key assumptions of the model are violated and other processes must be more dominant or at least strongly complementary to migration. To explain such a potential discrepancy and to identify model assumption violations we performed a number of different (multivariate) analyses on the empirical data, complementary to the simulations and studied distribution of dominant species. Using empirical data from 223 hectares' worth of plots in the Amazon, covering 4493 species and 120.322 individual trees, we simulate forests on the order of 8000 hectares, with 400-500 individuals per hectare. With the Amazon being one of the most diverse forests of the world in terms of tree species (14, 137, 226), such a large dataset allows us to test the model at different spatial scales and different communities in terms of diversity.

Results

No single model parameter setting was capable of reproducing correct patterns at both regional and local scales fitting empirical observations simultaneously. Although regional RAD patterns (Fig. 6.1), total number of species and Fisher's alpha of the total sample (Table 6.1 and Fig. 6.2) showed good, although not significant fits for two out of three datasets (Guyana/Suriname and French Guiana) the simulation output did not approach the empirical data for Maximum Dominance Distribution (MDD: the distribution of species with the highest number of individuals per plot) at plot level for any of the three datasets (Figs. 6.1, 6.2). For Guyana/Suriname there was a significant difference between the predicted and field RADs, although maximum distance (D, derived from the Kolmogorov-Smirnov test) between the two distributions was small (Fig. 6.1). There was also a relatively small yet significant difference in the mean number of species per plot but no significant differences in the mean number of singletons (species with only one individual) (Table 6.1). Simulated and empirical regional RADs were also significantly different for French Guiana (again with small maximum distance), with a significant difference in both the mean number of species per plot and mean number of singletons per plot at local scales. For Ecuador/Peru, simulations yielded a much less diverse sample than the empirical data resulting in a strong significantly different RAD yielding

Table 6.1 Table comparing simulated (Sim) and empirical datasets (Field) in terms of number of species, singletons and Fisher's Alpha (both total and mean per plot), ** indicate significance levels at P ≤ .01, * at p ≤ .001.**

	Guyana/Suriname		*French Guiana*		*Ecuador/Peru*	
	Sim	**Field**	**Sim**	**Field**	**Sim**	**Field**
Mean nr species	110***	84***	114***	157***	113***	168***
Total nr of species	1227	1042	1212	1204	2247	3018
Mean nr singletons	34	33	36***	78***	40***	88***
Total nr singletons	215	210	212	208	462	998
Mean FA per plot	46**	31**	48***	76***	58***	101***
FA of total sample	243	199	244	242	489	716

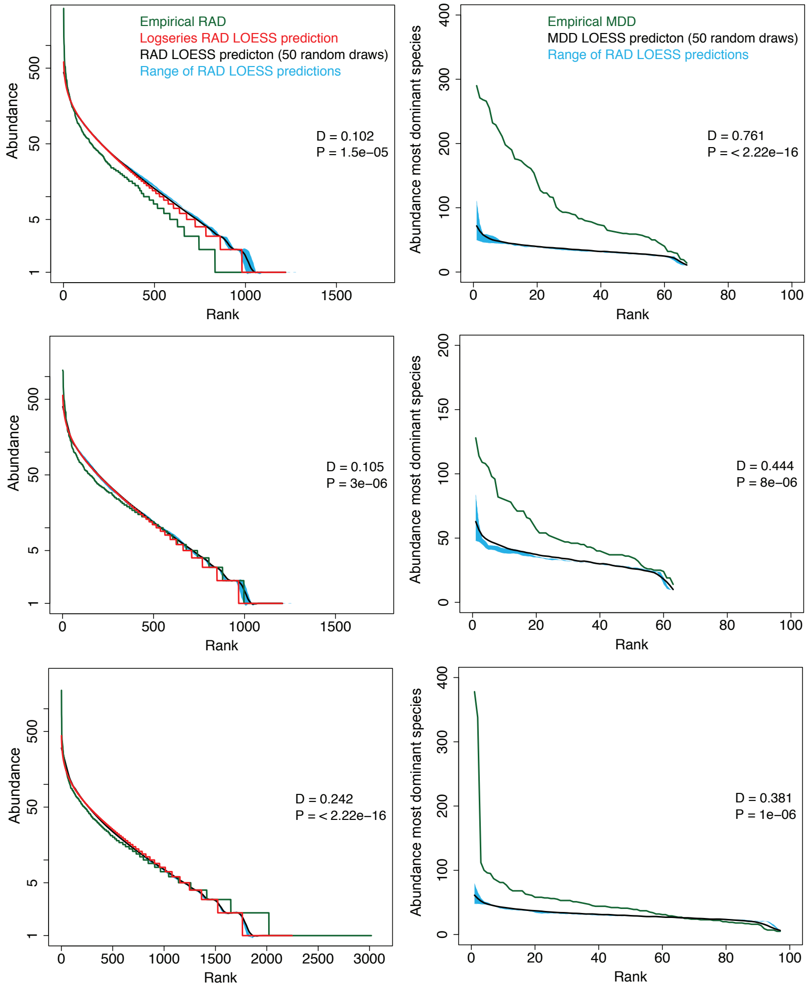


Fig. 6.1. The Rank Abundance Distribution (RAD, left) and Maximum Dominance Distributions (MDD, right) for tree species in 223 Amazon forest plots from Guyana/Suriname (top), French Guiana (middle) and Ecuador/Peru (bottom). Green lines indicate field data, black the simulated data based on the spatially semi-explicit model and red the fitted logseries based on the simulated distribution of individuals over the species. Blue shading indicates upper and lower RAD based on 25 sampling iterations of the total simulated forest. For the RADs, x-axis indicates the rank from most abundant to least abundant species, with the y-axis showing the actual abundances of the species for the ith rank. For the MDD graphs, the x-axis reflects the ranking of plots and the y-axis the maximum dominance of the most abundant species for each plot. Green depicts results from the empirical data, black the simulation output based on the same empirical dataset, blue shading indicates maximum and minimum values of distributions after 25 sampling iterations. Simulation output shows a much more even distribution of maximal dominance over all plots in comparison to the empirical dataset for all three regional datasets.

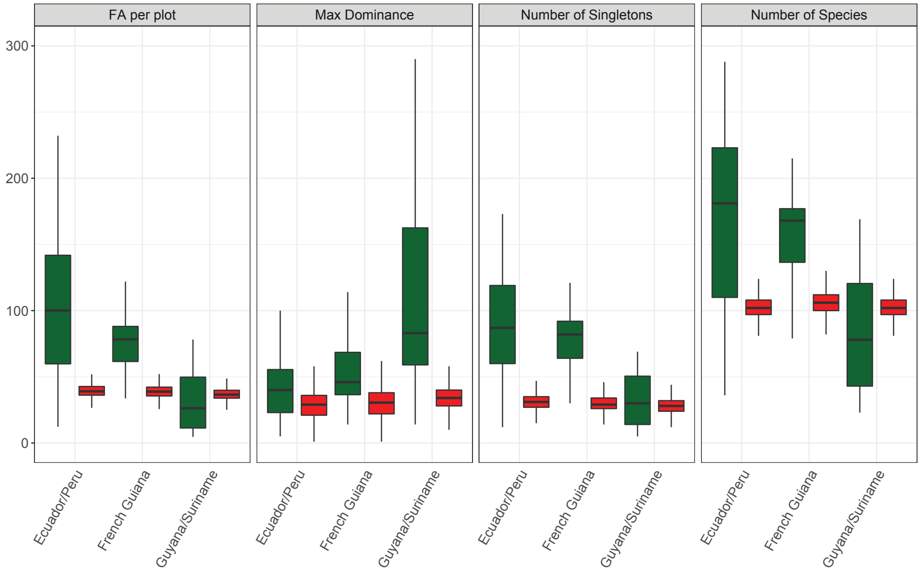


Fig. 6.2. Boxplots summarizing features of quantitative variables of composition for both simulation and field data) for Guyana/Suriname (top), French Guiana (middle) and Ecuador/Peru (bottom). Statistics are shown by the labels for the plots from the simulation (red) and from the actual field data (green) after a single sampling iteration. Whiskers of boxplots indicate minimum or maximum values (excluding outliers), hinges reflect lower and upper quartiles, and bold stripes reflect median values.

less than half of the species found in the empirical data as well with a maximum distance over twice as large as for the other two datasets. There were also significant differences in the mean number of species and singletons per plot. All RADs at a regional scale showed the familiar logseries (Fig. 6.1) although comparisons of mean Fisher's alpha (FA) per plot did reveal significant differences for all datasets. Regional total FA indicated close comparisons for both Guyana/Suriname and French Guiana whereas Ecuador/Peru again showed larger differences between observed and simulated values. From the RAD it can clearly be seen that primarily the very common species are responsible for distances for all three datasets yet larger distances between empirical and simulated RAD of Ecuador/Peru are most due to differences in the tail of rare species.

As simulations were unable to attain realistic patterns in dominance distribution of species we redid simulations using near null migration (m <<.1) to mimic extreme ecological drift at the local level. We also performed simulations at the other extreme of near unity (m = .9) migration, mimicking a panmictic community. This clearly

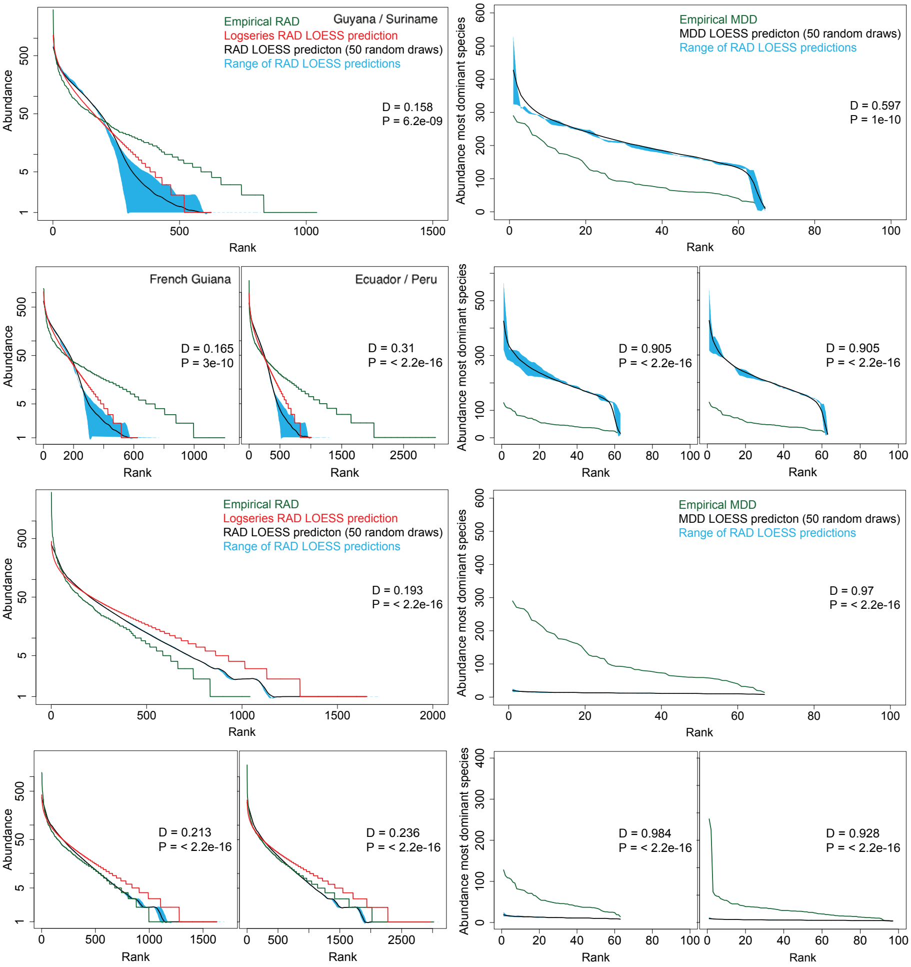


Fig. 6.3 Relative abundance (left) and Maximum dominance distribution (MDD: right) for two limiting cases of near null (top) and near unity migration (below). Distributions are shown in each corner for Guyana Suriname (top), French Guiana (left) and Ecuador/Peru (right) Distributions clearly show the disagreement between predictions both for relative abundances and the maximum dominance, with migration set near null yielding accurate regional predictions but losing local predictions as based on MDD with for migration set to unity yielding the opposite results.

showed the disagreement between regional and local predictions of the RADs (Fig. 6.3). The first resulted in MDD shapes approaching the empirical data (yet too even and too rich and still significantly different) but at the cost of regional diversity where RAD agreement was lost. With migration probabilities set near unity, regional predicted patterns of RAD again showed stronger approximation with empirical data (although richer) but MDDs were almost flat, i.e. individuals were too even distributed over the species and significantly different.

Analyses of composition. For all three datasets there were significant correlations between spatial distance and composition dissimilarity with relatively high r statistics from the Mantel tests for Guyana/Suriname (0.3101) and French Guiana (0.6723) whereas for Ecuador this was considerably lower (0.2073)(SI Chapter 6: Table S1). Dissimilarity of composition was also compared with environmental distance matrices where local ecology was approximated by Euclidean distances for annual rainfall and a binary distance index of 0 or 1 for forest type (SI Chapter 6: S2). For Guyana/Suriname this yielded a weaker r statistic of .1176 for the former but a similar r statistic of .2961 for the latter, both significant. For French Guiana, only comparisons between local ecology and species distances were available as all plots are from the same forest type, yielding an r statistic of .1713, also significant. For Ecuador, in comparing species distances with local ecology yielded an r statistic of .1742, with forest type vs. species distances yielding .3122, both also significant. NMDS also showed distinct grouping for all three different subsets with high agreement between plotted values and observed dissimilarities (all $R^2 > .95$) (SI Chapter 6: Fig. S3). Guyana/Suriname showed strong groups based on both country and forest type. French Guiana showed strong overlapping groups based on geographical subdivisions. Ecuador/Peru, with analyses performed separately for all forest types combined and only TF to show separation on country of origin on axes more discretely, yielded clear visible segregation along the first axis for both forest type and country of origin. A one-way ANOVA based on the scores of the first or second axis yielded significant differences for segregation of both country and forest type for Guyana/Suriname. Segregation of geographical subdivision along the second axis of the NMDS for French Guiana also proved to be highly significant as well as segregation of country and forest type for Ecuador/Peru, both along first axis in the separate analyses.

Discussion

Incorporating dispersal in a realistic way and being able to model a considerably large area, we were able to predict diversity at regional scales but unable to predict diversity patterns at local scales and vice versa. This disagreement between regional and local predictions suggests that even if regional patterns follow neutral predictions, suggesting neutral dynamics with a significant role for dispersal (e.g. (35)), local patterns may deviate strongly and indicate non-neutral dynamics. In the least, they suggest a severe scalability issue of neutral theory. At regional scales, simulation results of both Guyana/Suriname and French Guiana were very similar to the actual field data. There were only small differences in the total number of species, total FA over all plots or the distribution of species and singletons over the sampled plots according to the RAD. For western Amazonian plots, however, the simulation yielded a much less diverse sample at the regional scale, not only in terms of total number of species (almost 1000 species fewer than the field data) but also for total FA (almost half of the field data). Simulated data also showed much narrower ranges of these values compared to the field data for any dataset, indicative of much more similar distributions across plots in comparison with field data (Fig. 6.2). In addition, local community structure showed a much more even distribution of species per plot than the field data (Fig. 6.1). Our results suggests that with estimates of dispersal limitation based on species characteristics, neutral theory can neither predict the high dominance of some species observed in any of the field datasets (even though they approximate regional patterns quite good) nor the excessive diversity of Western Amazonian forests reflected in the large tail of rare species. Only with severely unrealistic dispersal limitation, patterns in maximal dominance at local scales can be approximated, but at the loss of diversity in comparison with empirical data at regional scales. As mentioned earlier, similar to the Hardy Weinberg principle from population genetics testing the null hypothesis of no evolutionary change when assumptions are not violated (222, 223), NT can be used as the null hypothesis of ecological equivalence in ecology. When the assumptions of NT, with ecological equivalence, birth and death rates being proportional to abundance in either LC or MC and a saturated landscape being the main assumptions, are indeed true we would expect predicted patterns to follow those observed in the field. If, however, for some reason any of these (or other assumptions) are not met, predictions deviate from empirical data. Our results indicate deviation from the model predictions occurred at two scales, the regional scale for the Ecuador/Peru data and the local scale, with the latter independent of the dataset used and show that at least one assumption is violated.

Rejection of NT on a regional level. We hypothesize potential violations are threefold: 1) differences in environmental heterogeneity, even within forest types, and life history strategy among species, 2) geographical distance between plots not being equal and 3) the laws of probability. Western Amazonia in general has much richer soils, lower wood density and smaller seed size in comparison with the Guiana Shield and is much more diverse in term of species (209). These differences in fertility and different life history strategy (indicated by wood density and seed mass differences) might allow higher and differential turnover rates of individuals and hence a higher diversity than predicted by NT where the assumption is strict ecological equivalence between species on the individual level (227). This in turn could lead to the higher in general diversity of Ecuador and Peru versus the Guiana Shield. Such signals within each dataset, where environmental heterogeneity obviously is to be found, also potentially account for the significant differences in all datasets, even with small distances between simulated and empirical RADs. The average geographical distance between plots in the empirical dataset is also larger for Ecuador and Peru in comparison with Guyana, Suriname and French Guiana (mean distance of 195 km for Ecuador/Peru versus 161 for the Guiana Shield datasets) adding to the turnover of species within the sampling scheme and reinforcing this difference in diversity resulting in the larger distance between simulated and empirical RADs. The last potential cause for deviation of predictions at regional scale is perhaps not so much a violation of the models assumption, but rather an indirect result of the modeling process, reinforcing the earlier mentioned violations. When simulations start, each plot shares roughly the same logseries of the total forest. The process of ecological drift then slightly changes this logseries for each plot separately causing some species to become more abundant whereas others become less abundant. As this is random across all plots and we have a large number of plots (8000), the law of large numbers will cause the logseries to be preserved in the total sample, even though each separate community might deviate substantially. A similar pattern is observed in population genetics in allele frequencies across communities. Separate samples starting with similar allele frequencies under influence of drift show that the average frequency over all plots does not change, even though each separate sample might show fixation or loss of the allele (228). The same could be happening in simulations of neutral models: even though each separate sample is under influence of ecological drift, perhaps losing some species and fixating others, adding up all plots results in an average RAD that hardly changes and might look like the one observed in the field. However, it is hiding the fact that each separate plot is quite different, both in terms of composition and structure in comparison with field data. More interestingly, patterns of MDD over all the plots showed remarkably different results compared to regional patterns, with no congruence between field data and simulation output even in extreme cases (Fig. 6.3). The summed regional

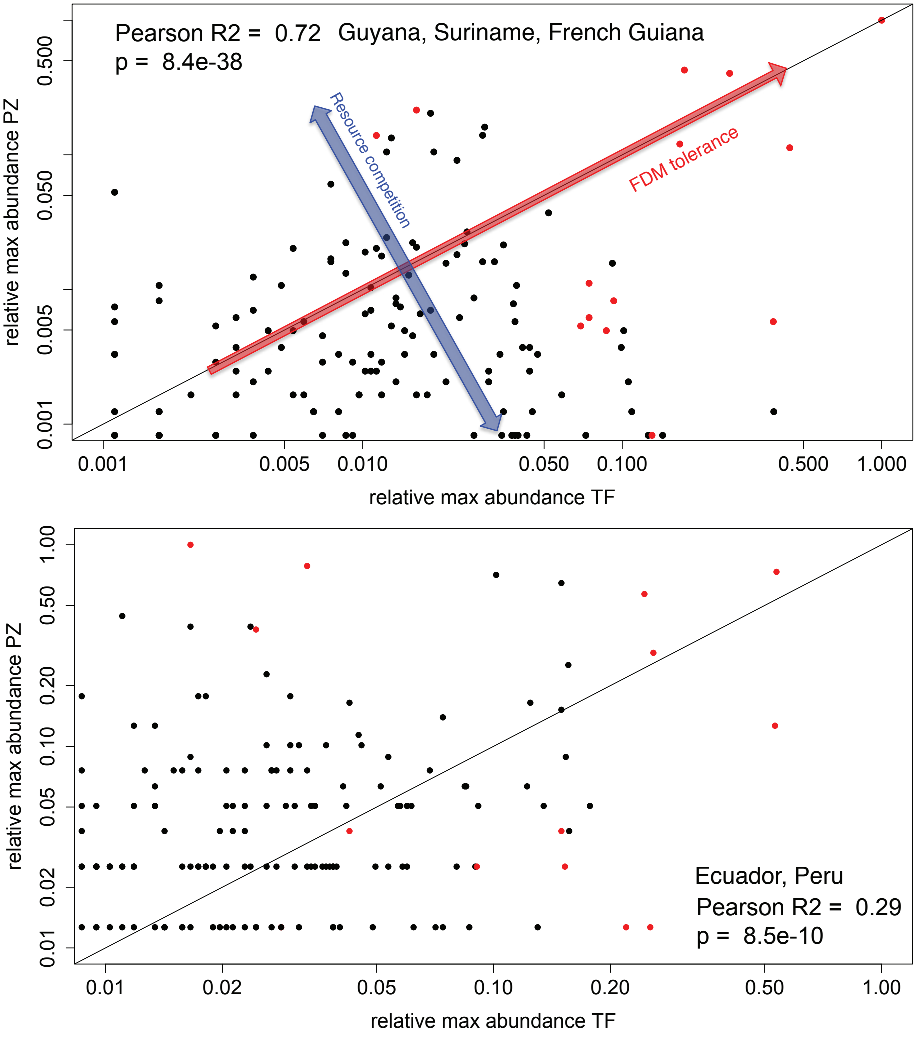


Fig. 6.4. Correlations between species identity and relative abundance corrected for total abundance and standardized over all plots for Guyana, Suriname and French Guiana combined (top) and Ecuador with Peru (bottom). Points indicate standardized relative abundance of species occurring in both Terra Firme (x-axis) and Podzol forests (y-axis). Red dots indicate species that attain maximal dominance within any plot. Pearson rank correlation coefficients are noted, including the estimated significance levels. Arrows indicate two categories in which species can attain dominance: mainly resource competition in combination with tolerance to frequency dependent mortality, limiting dominance to a single forest type (blue arrow) or on either forest type indicative of only tolerance to Frequency Dependent Mortality but a lesser degree of resource competition (red arrow).

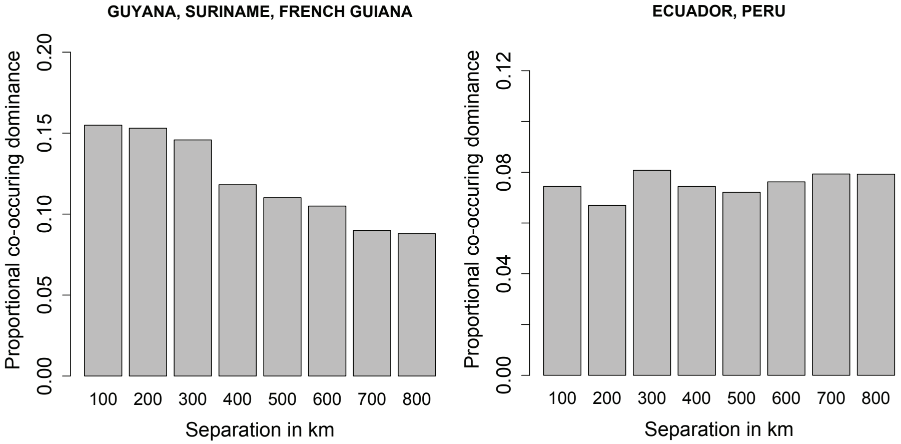


Fig. 6.5 Proportion of co-occurring dominant species plotted against distance classes between plots. X-axis shows distance classes between which plots comparisons are made, proportions of co-occurrence indicates the proportion of plots within the distance class of each other share the same species being the most dominant.

rank abundance distributions then would suggest neutrality by their good fits (at least for the Guiana Shield), which appears not to exist at the local scale.

Rejection of NT on a local level. At local levels we show strong deviation in predicted patterns of MDD versus those observed in empirical data. Even at regional levels, some species are more abundant than either predictions or estimations using a logseries distribution (Fig. 6.1). This suggests some species are better competitors in some way, reaching higher abundances than predicted by NT at both scales. This clearly is a violation of one of the key assumptions of NT, ecological equivalence, which would predict a much more even distribution. Mantel tests supported this view and in accordance, NMDS also showed clear segregation of plot community composition based on both geographical and environmental proxies for all three datasets used. These results would indicate that at least in terms of composition there is a strong effect of both environmental filtering and dispersal limitation, violating at least partly the assumption of ecological equivalence. Although this would be expected for communities of different forest types, many of the dominant species are clearly not restricted to a single forest type. This leaves the question whether being the better competitor related to abiotic conditions is making species more dominant than predicted by NT or that perhaps a greater ability to withstand pests, pathogens and herbivores could account for this pattern.

Identifying key violations of ecological equivalence. If environmental filtering and subsequent selection on certain traits based on abiotic conditions would account mostly for dominance of species we would expect species dominant within one forest type are not necessarily dominant in others as selective regimes would be different. To test this, we correlated species identity and relative abundance, corrected for total abundance and standardized over all plots for the two major forest types found in each dataset (Terra Firme and Podzols). This was done within geographical subsets to account for dispersal limitation effects (i.e. the three datasets separate), similar to an approach studying distribution of species in Peru and Ecuador performed earlier (167). This showed there was a strong correlation between relative abundances with a highly significant Pearson rank correlation of .72 for Guyana/Suriname and French Guiana combined. For Ecuador this yielded a weaker correlation of .29 although still highly significant (Fig. 6.4). Interestingly, species that attain maximal dominance in any plot (indicated by red) fall in two distinct categories: those dominant on a single forest type or those attaining dominance across forest types. We show this is not due to a mass effect of clustering due to limited dispersal leading to the same dominant species in nearby plots as only between 8 and 15% of dominant species co-occurred between plots, with only a slight decrease in the proportion of co-occurrence over larger distances (Fig. 6.5). Species apparently can attain dominance in two ways: being a good competitor in a specific environment driven by for instance resource competition (confined by forest type in this case, indicated by the blue arrow in Fig. 6.4) or being a good competitor regardless of the abiotic environment (red arrow). In either case they must have tolerance to frequent dependent mortality (FDM tolerance, as indicated in Fig. 6.4). In terms of violating the assumption of ecological equivalence, we now also have two explanations accounting for the excessive dominance of species related to different scales: 1) species either outcompete other species, adding to deviance at regional scale patterns as different species are dominant in different areas or 2) they are better competitors in terms of escaping from frequency dependent mortality, resilience to pathogens or predators (229, 230) or specialization (14), regardless of habitat. Processes such as competitive exclusion (231, 232) based on environmental filtering or even severe ecological drift (233), however, are less likely to account for the majority of these patterns in MDD on the local level. The first would result in either the same set of species reaching higher abundances more likely than expected by chance (at least within forest types and over distances) while the second would result in lower regional diversity as shown by our model predictions. Our findings are supported by recent studies on correlation of species richness on different taxonomic levels at a global scale (234, 235), also indicating violation of neutral theory assumptions, as the number of species per family and individuals within species were highly correlated across different continents.

Conclusions

Our results indicate a severe scaling issue in predictions of neutral theory. It fails to scale either up or down while maintaining good predictions for both local and regional predictions of community structure. We hypothesize this is due to model assumption violation at different scales, in particular that of ecological equivalence as indicated by the disagreement between distributions of maximum dominance of species across species. In addition, we show that even though (summed) regional patterns in diversity from neutral models may be accurate, there is no guarantee that local plot dynamics and hence the mechanisms behind community composition are also neutral.

Materials and Methods

Spatially semi-explicit models: modeling the green mass. We used a mechanistic model (236) simulating not only separate plots and their direct interaction (SI Chapter 6: Fig. S1), but also the intermediate green matrix connecting these plots (not unlike the analytic network approach by Economo & Keitt (28)). Although we often look at the forest using only a relatively small sample of plots, it is this intermediate green matrix which plays a vital role in determining species composition of each local plot, acting as a bridge for exchanging species between the plots being sampled. The model is built up as a Rubik's cube, with each column of the cube representing a single plot within a forest with its individuals stacked as the individual blocks (SI Chapter 6: Fig. S2). The different colors of blocks represent different species and the number of individuals (i.e. amount of stacked blocks) is based on an average amount per plot as observed in the empirical set used for comparison. Creating the forests starts with each block (i.e. each individual) being assigned to a species by randomly sampling from a hypothetical metacommunity. This metacommunity follows a logseries, which has been shown to be the best approximation for describing species richness of hyper diverse communities (2, 14, 137, 196). The logseries is parameterized using the actual field data to which the simulation is being compared, similar to an earlier study (5). With each time step of the simulation, the forest is allowed to change with one individual in each plot randomly chosen for replacement. Replacements can come from either of five categories: 1) the plot itself (local recruitment), 2) adjacent plots, 3) the entire forest, 4) the hypothetical metacommunity or 5) a speciation event, which creates a new species neither present in the forest nor in the metacommunity. We estimate probability of migration from adjacent plots using the Corrected Plot Geometry method (138, 236) and mean dispersal distance based on phenotypic characteristics, see below. Migration probability of each subsequent category is calculated as 10 percent of the former, e.g. if the migration probability from adjacent plots is estimated at .071, that from the entire forest is set at .0071 and from the hypothetical metacommunity at 0.00071. We calculated speciation as in the original UNTB: *theta/(2*J)* with theta equal to Fisher's α (29) and J the total number of individuals in the forest. Parallel processing using either multiple cores on one processor or a cluster using the packages foreach, *doParallel* and *doSnow* (237, 238) allows multiple forests to be simulated at once, all drawing from the same larger hypothetical metacommunity. These separate forests are indirectly connected as they draw from the same metacommunity, essentially simulating vagrant dispersal from a larger species pool. This allows for much faster computation of a large area. Each step of the simulation itself is explained in chronological order in the Supporting Information of Chapter 6 (S1).

Field datasets. Three independent datasets were used: Guyana/Suriname combined, French Guiana and Ecuador/Peru, also combined. All identifications within each dataset were harmonized and are independent and non-overlapping (214). Each dataset consisted of plots having all trees ≥10 cm Diameter at Breast Height (DBH) inventoried. Species ID's were standardized to the W3 Tropicos database within each dataset, using TNRS (14, 94). The Guyana/Suriname set consisted of 67 plots all of one hectare in size, yielding 37.446 individual trees distributed among 1042 morpho-species. French Guiana is comprised of 63 plots, ranging between .40 and 1 hectares in size (.40 ha 1 plot, .50 ha 3 plots, .80 ha 1 plot, .98 ha 2 plots and 1 ha 56 plots) accounting for 35.075 individuals belonging to 1204 morpho-species. Ecuador/Peru having 93 plots, ranging in size from .2 to 1 hectares in size (.1 hectares 2 plots, .2 ha 1 plot, .25 ha 6 plots, .5 ha 1 plot and 1 ha 87 plots) accounts for 47.801 individuals and 3018 morpho-species. A map of the locations of all plots is provided in the Supporting Information (SI Chapter 6, Fig. S4).

Parameterizing the model. The mean dispersal distance for each dataset to be implemented in the Corrected Plot Geometry method (236, 239) was calculated by assigning a mean dispersal distance depending on the category, based on literature (210, 240, 241) (SI chapter 6: Table S2). This was done for each plot and ultimately averaged over all plots per dataset (see below). As a control, we also simulated the forests for a range of combinations where the total summed amount of migration was randomly divided over all the different categories. A table of all used parameters is provided in the Supporting Information (SI chapter 6: Table S2). To test for the influence of severe and absent ecological drift on the difference between local and regional patterns of diversity, we also implemented near null (m <<.1) and near-unity parameters of migration (m = .9).

Sampling and analyses. After the simulations, a number of plots equal to the amount of plots in the dataset used for comparison were sampled randomly from the forest. Shapes of the RADs for both the simulation output and the empirical data were compared using the non-parametric Kolmogorov–Smirnov test (242) as it allows for a goodness-of-fit test between two distributions without assuming any theoretical distribution and calculates the statistical distance between the two distributions. D values reported indicate maximum distance between the two distributions, with p-values indicating the probability of such a D statistic being larger or equal to the observed value. The mean number of species in the total sample, number of singletons and Fisher's alpha (98) were compared using the non-parametric Wilcoxon rank sum test (243). We posited that if forest dynamics are similar to our neutral model, these aspects of the empirical and simulated datasets should also be similar. Any substantial deviation would represent non-dispersal related influences on species composition. Thus we treat the model as a null-model, much like the Hardy-Weinberg theorem in population genetics (222, 223), with only dispersal as a mechanistic driver. In addition to studying the regional patterns in diversity we did the same for local patterns studying the average number of species per plot and the ranking in dominance of these species per local community over the whole dataset. To complement these comparisons, we performed three different analyses to study the relative importance of geographical distance and environmental filtering. These were Non Metric Multidimensional Scaling (NMDS) (112–114) using the Morisita index of diversity (244) as distance measure and a correlation analysis between environmental, geographic and species distance matrices using Mantel tests (104, 105), using the same distance measure. All are explained in more detail in the Supporting Information (SI chapter 6: S2).

Acknowledgments

We thank Pablo Stevenson for contributing data on dispersal syndromes used in estimating dispersal distances. We also thank Jari Oksanen for discussions on the calculation of the Morisita index and Henry Horn for valuable feedback during the writing process. Funding: Daniel Sabatier and Jean-François Molino are supported by an "Investissement d'Avenir" grant managed by Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-0025). Author Contributions: EP and HtS designed the experiment. EP wrote R-scripts, analyzed the results.

Supporting Information

S1 Steps of the Simulation

1) Setting up the metacommunity. The first step of the simulation process is creating a metacommunity having similar characteristics to the actual field data with which to compare the results afterwards. For this we made use of Fisher's Logseries, as most forests of the Amazon show a near exact fit with this distribution model (2, 14, 137). Assuming the larger surrounding metacommunity also follows a logseries and that the datasets are reflective of this metacommunity we can derive the relative abundance distribution from the expected number of species and individuals by solving for Fisher's Logseries parameters alpha and x. These are then used to construct each term of the logseries (i.e. singletons, doubletons etc.) until all individuals have been distributed for each separate empirical dataset, similar to earlier published results (139, 245).

2) Creating the forest and filling it up with individuals. The forest itself is created by setting up a square lattice of size N (Number of plots), where each XY-coordinate represents a local community of stacked individuals as an array instead of one individual (Fig. S2), similar to Kimura's stepping-stone model from population genetics (51). The size of the array, i.e. the number of individuals, depends on the average amount of individuals found per plot of the field data (*Jp*). The forest is filled up with individuals belonging to specific species for the first time step of the simulation process. This is done at random (i.e. mixed) by sampling the metacommunity created earlier for each available slot in the forest. By doing so, the model at t = 0 is a direct representative sample of the metacommunity with the probability of selecting a species a direct result of its abundance in the metacommunity.

3) Running the neutral game: migration and speciation. After the metacommunity is created and the entire forest of *N*Jp* individuals is filled up, the next step is to specify the recruitment parameters. Probabilities of migration in each category are based on the mean mode of dispersal based on the dominant dispersal mode of the majority of species present in the respective communities. For each plot, for all known species within the plot, the dispersal syndrome was verified, if known (courtesy Pablo Stevenson). Plots were categorized into predominantly zoochorous, zoo/synzoochorous, explosive dehiscence/zoochorous or anemo/zoochorous as these were the most occurring combinations of syndromes. We calculated the percentage of each dispersal syndrome present within a plot and grouping was then based on either a sole category comprising over 75% of all individuals or, if this was not the case, a combination yielding over 50% of all individuals in the plot. This mean dispersal distance belonging to this grouping is based on previous literature (see main text), this distance is then implemented in the Corrected Plot Geometry method (see main text). The probability of speciation is determined from Hubbell's equation for the fundamental diversity number: $\theta = 2Jmv$, resulting in $v = \theta/2*(N*Jp)$ with theta assumed to be similar to the forests Fisher's alpha. The sampling and replacing of individuals is carried out sequentially, starting with the four corners, followed by all edges and lastly the remaining plots on the lattice, this way we avoid having to implement any torus and are able to incorporate edge effects as they would also occur in a real forest (e.g. on the edge of a river, a cliff or coastline). Each simulation was run for 150,000 time steps.

S2. Analysis

Multivariate analysis - NMDS. A Non-Metric Multidimensional Scaling (NMDS) using MetaDMS and Morisita index of diversity (244) from the package *vegan* (99) was performed on each of the three datasets to study patterns of community composition and its relation to either environmental characteristics or geographical distances. For each of the datasets the NMDS was performed separately, after which points in the two-dimensional space were colored to represent environmental characteristics or geographical separation. If either environmental or geographic distances were mostly related to floristic composition, plots should be separated accordingly in the ordination space. NMDS attempts to find the best rank-order agreement between floristic similarities and distances between points in the ordination space (112–114). As such, NMDS does not fit axes based on eigenvalues as many other ordination techniques do but instead represents a coordinate system for the computed space. MetaMDS is a specific NMDS approach which, in addition to the standard MDS, centers the origin of the ordination space on the averages of the axes and then uses principal components so that most variation is represented by the first axes, then the second, etcetera (143). If migration is the most important factor determining species composition, we expect grouping of points will be strongly related to geographical origin (i.e. country or region). In contrast, if environmental filtering is stronger, grouping is expected to be more strongly related to forest types.

Environmental and geographical distance matrices - Mantel tests. To determine whether geographical distance (being neutral) or environmental similarity (being more niche dependent) would be correlated with similarity across plots we also performed Mantel tests (104, 105). Geographical distance matrices were based on actual Euclidean distance derived from latitude and longitude coordinates for all plots. For the floristic species similarity matrices we used the same Morisita diversity index as mentioned earlier. Environmental distance matrices were based either on mean annual rainfall or forest type. With the first, distance is calculated as the absolute Euclidean distance. For the latter, a binary index was used for measuring similarity with 0 indicating the plots both are of the same forest type and 1 when this is not the case. For the Mantel tests we also subdivided all plots into separate groups depending on the dominant dispersal mode within the plots. Distance matrices and Mantel tests were performed using the distance matrix function in vegan's vegdist and the mantel function (143).

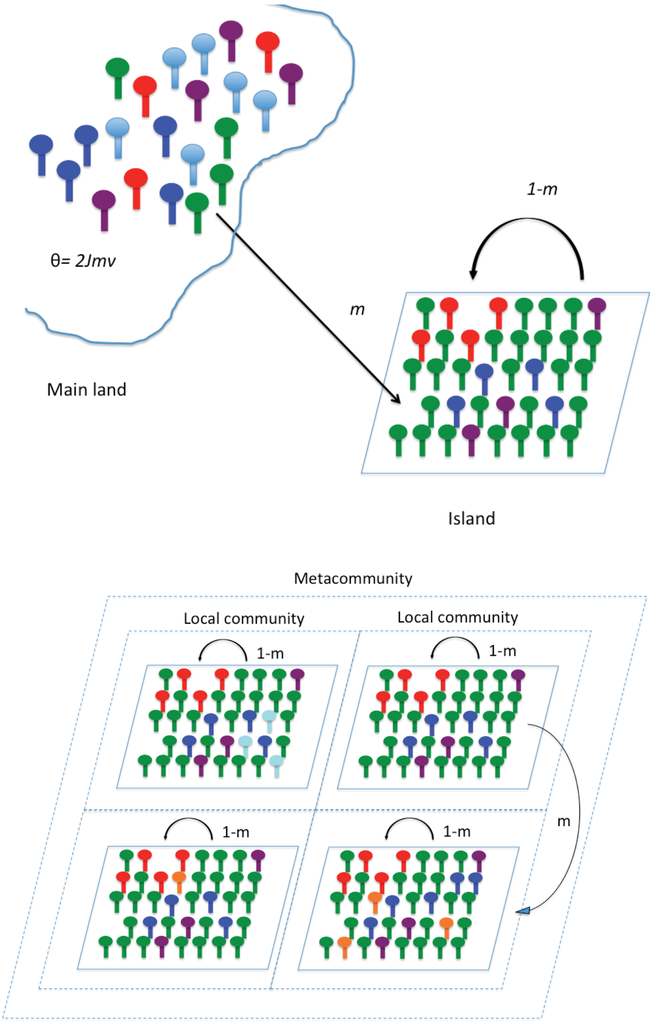


Fig. S1. Above) Original neutral model adapted to the theory of Island Biogeography (53) with a larger metacommunity as the mainland and a smaller local community as the island. Relative abundance distributions of the metacommunity are given by Hubbell's fundamental biodiversity number theta (29) (equals *2Jmv*, with *Jm* the size of the metacommunity and *v* the speciation rate). Migration determines the relation between the metacommunity and local community with 1 minus the migration probability giving the probability of a local recruit after a death in the local community. **Right) Later adaptation to the original neutral model with the metacommunity as the sum of a collection of local communities.** Local communities are connected by the same migration parameter although this is now migration from plot to plot. It is, however, still an approximation of migration, as the intermediate plots are not taken into account and migration still acts as an ecological aggregated parameter, incorporating not only dispersal but also filtering and recruitment.

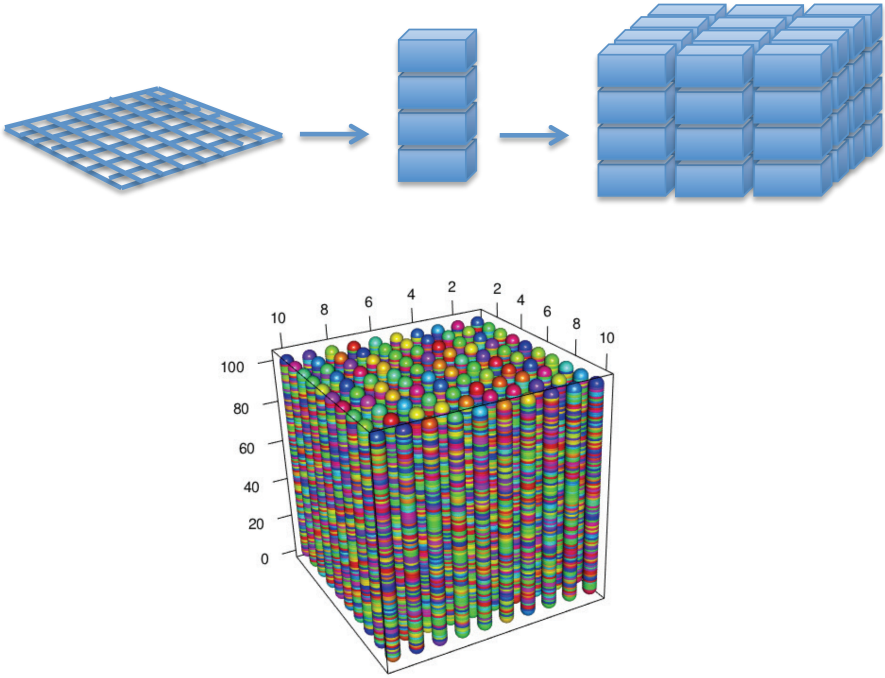


Fig. S2. New proposed mechanistic model build up as a Rubik's cube, with each column of the cube representing a single plot within a forest with its individuals stacked as the individual blocks. Each unique x-y coordinate represents a single plot with z individuals, bottom shows the 3D view of the forest with the individuals stacked as an array of per plot (example forest of 10x10 plots with each plot having a 100 individuals).

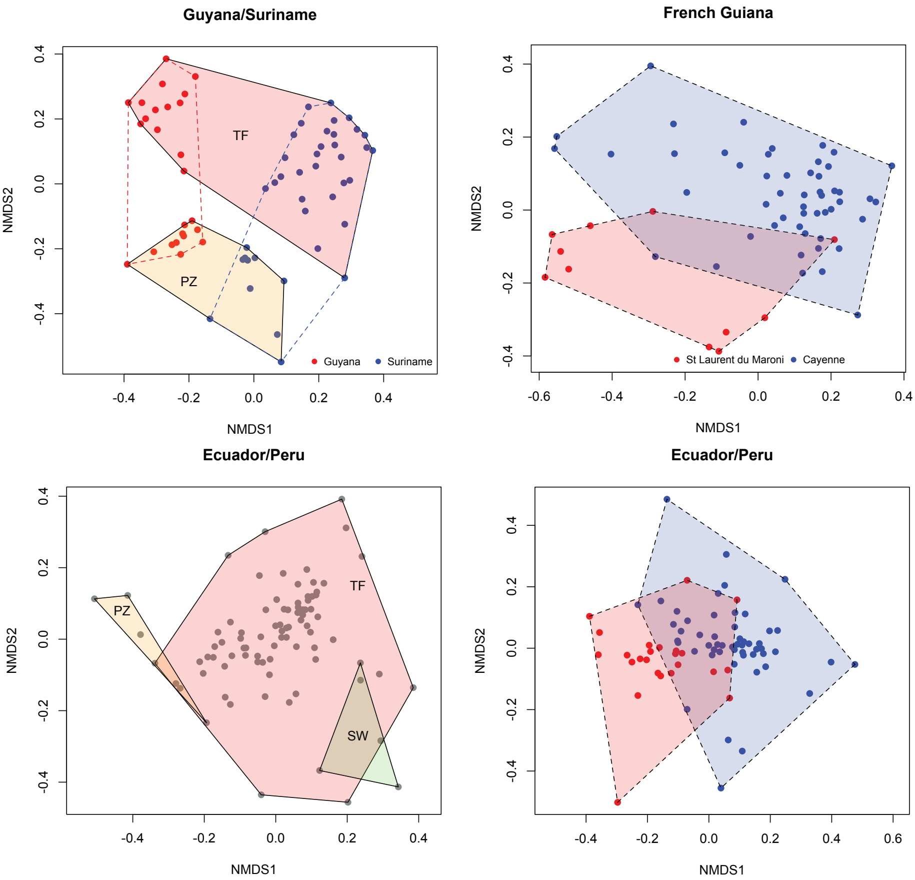


Fig. S3 Results of the Non Metric Multidimensional Scaling for Guyana and Suriname. Polygon coloring and grouping is based on forest type (terra firme TF, podzol PZ, swamp SW). As French Guiana only has a single forest type included (terra firme) the polygon is excluded. Dashed lines indicate grouping based on country of origin or region of origin for French Guiana (Cayenne or St Laurent du Maroni). First axis segregation for Guyana/Suriname on country was highly significant (F(1,63) =243 and p < 2.2 e^{-16}) as well as the second axis segregation on forest type (F(1,63) =107 and p = 3.6 e^{-15}). Segregation of geographical subdivision along the second axis for French Guiana also was highly significant (F(1,60) =28 and p = 1.7e^{-6}) as well as segregation along first (forest type) and second (country) for Ecuador/Peru (F(2,82) =15 and p = 3.1 e^{-6}, F(1,75) =54 and p = 1.8 e^{-10}).

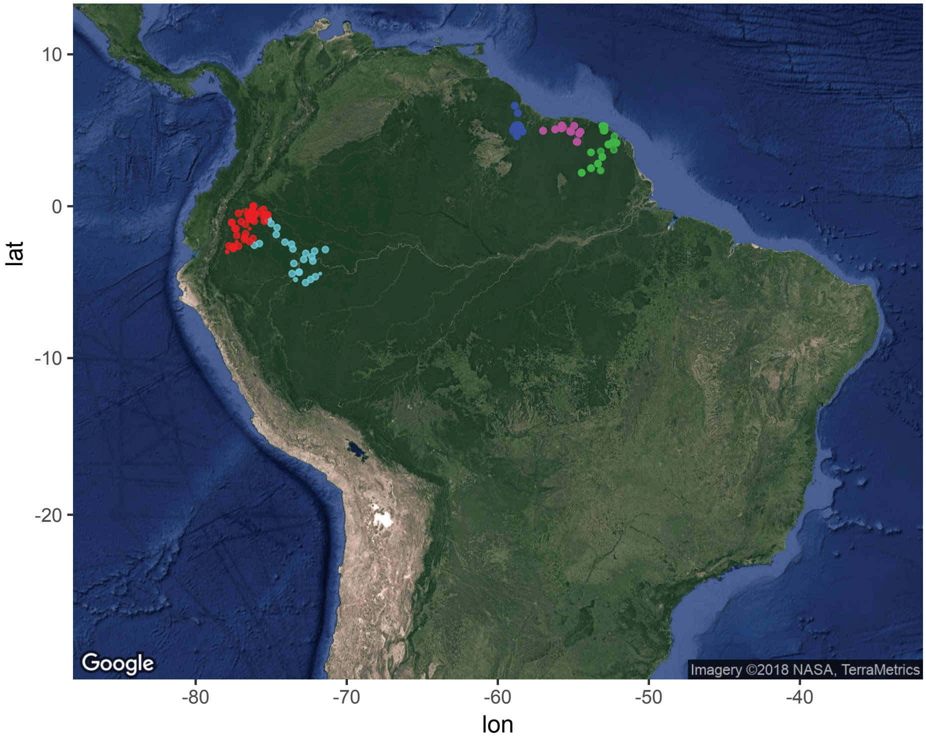


Fig. S4 Map showing the spatial location of all plots from the different geographical subsets of Guyana (blue), Suriname (pink), French Guiana (green), Ecuador (red) and Peru (light blue).

Table S1. Results from the Mantel tests for each separate dataset with and without specific grouping based on dominant dispersal syndrome. Asterisks indicate significance at the .05 level.

OSB	Spatial	Local ecology	Forest type
All species	.3101*	.1176*	.2961*
Dispersal ability			
Zoochory	.5519*	.08003	.4056*
Anemo/zoochory	.5686*	.2904*	.2351*
Explo/zoochory	.2297*	.08632	.4464*

DS	Spatial	Local ecology
All species	.6723*	.1713*
Dispersal ability		
Zoochory	.5913*	.4647*
Zoo/Synzoochory	.6322*	.4586*
Anemo/zoochory	.6110*	.4925*
Explo/zoochory	.2824	.3102*

JEG	Spatial	Local ecology	Forest type
All species	.2073*	.1742*	.3122
Dispersal ability			
Zoochory	.2297*	.2139*	.2347
Zoo/Synzoochory	.2401	.3308*	.3404*
Anemo/zoochory	.6068	.4887*	.6553
Hydro/zoochory	.3513*	.3716*	.9186*

Table S2. Mean dispersal distances in meters based on literature for the various dispersal syndromes. Distances were used to calculate the average dispersal distance per plot based on the grouping category of the predominant dispersal syndrome as approximation using the Corrected Plot Geometry method. Below: mean migration probabilities for each category per dataset.

syndrome	**min**	**max**	**mean**
zoochorous	3.16	100	25.12
zoo/synzoochorous	4.39	100	52.19
explosive dehiscence/zoochorous	3.16	52.8	27.98
anemo/zoochorous	6.58	100	53.29

Dataset	**local recr.**	***m* adjacent**	***m* forest**	***m* meta**	**speciation**
Guyana/Suriname	.84394	.141	.0141	.00141	.000563
French Guiana	.83217	.1512	.01512	.001512	.000565
Ecuador/Peru	.82229	.1601	.01601	.001601	.000638

Information is something that can be used to remove uncertainty.

Claude E. Shannon (1916 - 2001)

Chapter Seven

Rolling the dice or struggling for survival, using Maximum Entropy to unravel drivers of community composition

(Currently under Review)

Edwin Pos, Luiz de Souza Coelho, Diogenes de Andrade Lima Filho, Rafael P. Salomão, Iêda Leão Amaral, Francisca Dionízia de Almeida Matos, Carolina V. Castilho, Oliver L. Phillips, Juan Ernesto Guevara, Marcelo de Jesus Veiga Carim, Dairon Cárdenas López, William E. Magnusson, Florian Wittmann, Mariana Victória Irume, Maria Pires Martins, Daniel Sabatier, José Renan da Silva Guimarães, Jean-François Molino, Olaf S. Bánki, Maria Teresa Fernandez Piedade, Nigel C.A. Pitman, Abel Monteagudo Mendoza, José Ferreira Ramos, Joseph E. Hawes, Everton José Almeida, Luciane Ferreira Barbosa, Larissa Cavalheiro, Márcia Cléia Vilela dos Santos, Bruno Garcia Luize, Evlyn Márcia Moraes de Leão Novo, Percy Núñez Vargas, Thiago Sanna Freire Silva, Eduardo Martins Venticinque, Angelo Gilberto Manzatto, Neidiane Farias Costa Reis, John Terborgh, Katia Regina Casula, Euridice N. Honorio Coronado, Juan Carlos Montero, Beatriz S. Marimon, Ben-Hur Marimon Jr., Ted R. Feldpausch, Alvaro Duque, Chris Baraloto, Nicolás Castaño Arboleda, Julien Engel, Pascal Petronelli, Charles Eugene Zartman, Timothy J. Killeen, Rodolfo Vasquez, Bonifacio Mostacedo, Rafael L. Assis, Jochen Schöngart, Hernán Castellanos, Marcelo Brilhante de Medeiros, Marcelo Fragomeni Simon, Ana Andrade, José Luís Camargo, Layon O. Demarchi, William F. Laurance, Susan G.W. Laurance, Emanuelle de Sousa Farias, Maria Aparecida Lopes, José Leonardo Lima Magalhães, Henrique Eduardo Mendonça Nascimento, Helder Lima de Queiroz, Gerardo A. Aymard C., Roel Brienen, Juan David Cardenas Revilla, Flávia R.C. Costa, Adriano Quaresma, Ima Célia Guimarães Vieira, Bruno Barçante Ladvocat Cintra, Pablo R. Stevenson, Yuri Oliveira Feitosa, Joost F. Duivenvoorden, Hugo F. Mogollón, Natalia Targhetta, Leandro Valle Ferreira, James A. Comiskey, Freddie Draper, José Julio de Toledo, Gabriel Damasco, Nállarett Dávila, Roosevelt García-Villacorta, Aline Lopes, Alberto Vicentini, Janaína Costa Noronha, Flávia Rodrigues Barbosa, Rainiellen de Sá Carpanedo, Thaise Emilio, Carolina Levis, Domingos de Jesus Rodrigues, Juliana Schietti, Priscila Souza, Alfonso Alonso, Francisco Dallmeier, Vitor H.F. Gomes, Jon Lloyd, David Neill, Alejandro Araujo-Murakami, Luzmila Arroyo, Fernanda Antunes Carvalho, Fernanda Coelho de Souza, Dário Dantas do Amaral,* *Kenneth J. Feeley, Rogerio Gribel, Marcelo Petratti Pansonato, Daniel Praia, Jos Barlow, Erika Berenguer, Joice Ferreira, Paul V.A. Fine, Toby Alan Gardner, Marcelino Carneiro Guedes, Eliana M. Jimenez, Juan Carlos Licona, Maria Cristina Peñuela Mora, Carlos A. Peres, Boris Villa, Carlos Cerón, Terry W. Henkel, Paul Maas, Marcos Silveira, Juliana Stropp, Raquel Thomas-Caesar, Tim R. Baker, Doug Daly, Kyle G. Dexter, John Ethan Householder, Isau Huamantupa-Chuquimaco, Toby Pennington, Marcos Ríos Paredes, Alfredo Fuentes, Jose Luis Marcelo Pena, Miles R. Silman, Sebastián Tello, Jerome Chave, Fernando Cornejo Valverde, Anthony Di Fiore, Renato Richard Hilário, Juan Fernando Phillips, Gonzalo Rivas-Torres, Tinde R. van Andel, Patricio von Hildebrand, Edelcilio Marques Barbosa, Luiz Carlos de Matos Bonates, Hilda Paulette Dávila Doza, Ricardo Zárate Gómez, Therany Gonzales, George Pepe Gallardo Gonzales, Jean-Louis Guillaumet†, Bruce Hoffman, André Braga Junqueira, Yadvinder Malhi, Ires Paula de Andrade Miranda, Linder Felipe Mozombite Pinto, Adriana Prieto, Agustín Rudas, Ademir R. Ruschel, Natalino Silva, César I.A. Vela, Vincent A. Vos, Egleé L. Zent, Stanford Zent, Bianca Weiss Albuquerque, Angela Cano, Diego F. Correa, Janaina Barbosa Pedrosa Costa, Bernardo Monteiro Flores, Milena Holmgren, Marcelo Trindade Nascimento, Alexandre A. Oliveira, Hirma Ramirez-Angulo, Maira Rocha, Veridiana Scudeller, Rodrigo Sierra, Milton Tirado, Maria Natalia Umaña Medina, Geertje van der Heijden, Emilio Vilanova Torre, Corine Vriesendorp, Ophelia Wang, Kenneth R. Young, Manuel Augusto Ahuite Reategui, Cláudia Baider, Henrik Balslev, Sasha Cárdenas, Luisa Fernanda Casas, William Farfan-Rios, Cid Ferreira, Reynaldo Linares-Palomino, Casimiro Mendoza, Italo Mesones, Armando Torres-Lezama, Ligia Estela Urrego Giraldo, Daniel Villarroel, Roderick Zagt, Miguel N. Alexiades, Karina Garcia-Cabrera, Lionel Hernandez, William Milliken, Walter Palacios Cuenca, Susamar Pansini, Daniela Pauletto, Freddy Ramirez Arevalo, Adeilza Felipe Sampaio, Elvis H. Valderrama Sandoval, Luis Valenzuela Gamarra, Gerhard Boenisch, Jens Kattge, Nathan Kraft, Aurora Levesley, Karina Melgaço, Georgia Pickavance, Lourens Poorter, Hans ter Steege*

† Deceased 01-2018

Affiliations: For full list of Authors and Affiliations see the Supporting Information

Abstract

Understanding drivers of community assembly remains important in ecology and attempts to resolve it range from truly deterministic to completely neutral. We apply Maximum Entropy to disentangle dynamics of Amazonian tree communities without invoking a-priori assumptions. We use over 2000 hectares of tree inventory plots and functional traits associated with a broad range of ecological challenges. We show an overall low, but strong, environmentally dependent effect of functional traits on genus level composition. We also show much stronger effects of dispersal from the regional taxonomic pool, accompanied by a strong spatial pattern that depends on geographical distance. Our results significantly contribute to the debate between neutral and niche paradigms and suggest direction for future studies on the governing dynamics of ecological communities.

Introduction

What drives community assembly? Approaches to answer this long standing question in ecology have varied from completely deterministic, or niche based (16–26) to completely neutral (29, 74, 246–248) and almost everything in between (e.g. near-neutral, continuum or emergent-neutral: 17–20). Most models are based on some prior described reasoning of a functional mechanism driving community dynamics. Inference of governing dynamics depend on the type of model used, neutral or deterministic, and whether or not model outputs fit with field observations. However, such a priori reasoning is also at the center of the question itself, making it a causality dilemma: choosing whether to use a deterministic, neutral or hybrid model of community assembly. But when the choice has been made, do results reflect the choice or actual community dynamics?

Choosing has a practical basis, e.g. deterministic models are popular as they allow precise inferences regarding dynamics. They are, however, quickly overloaded with parameters, challenging their use, especially in hyper diverse communities. As a response, more simplified (neutral) models emerged as null-hypotheses to test against actual empirical observations. Without invoking complex sets of parameters, these run on simple but strict rules of demographic stochasticity. But even such models require complex parameterization and often unrealistic assumptions (139, 226). In addition, they have difficulties disentangling predictions at different scales (215) as well as faced by many empirical objections (34, 159). Given all this, it may be helpful to explore different methods of quantifying niche versus neutral processes.

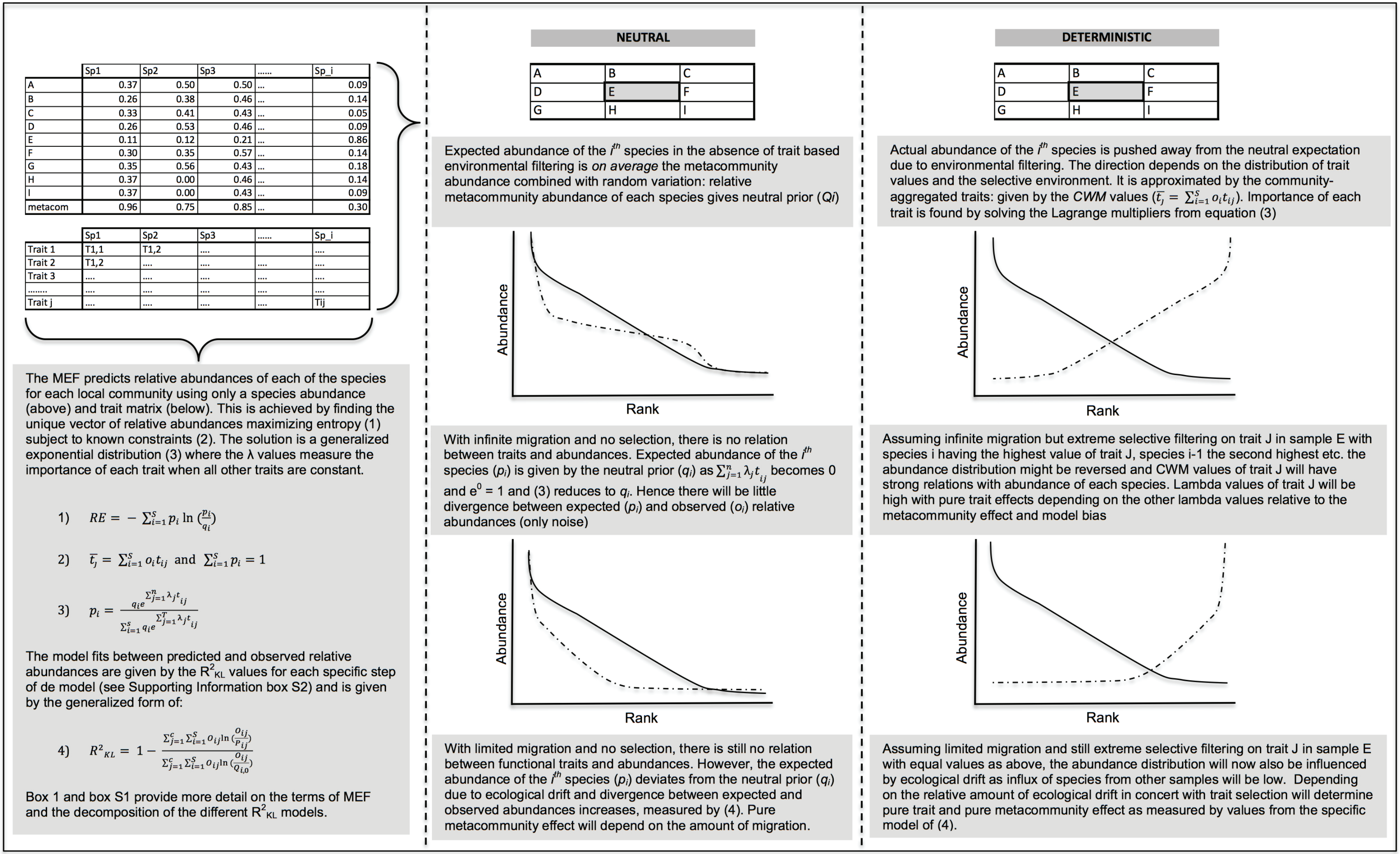
The Maximum Entropy Formalism (MEF) provides such a different method (40, 79, 251). In contrast with models as mentioned above, this makes no inference regarding population dynamics and no a-priori assumptions in terms of functional dynamics to generate predictions of relative abundances. Instead, it mathematically derives relative abundances in the form of Bayesian probabilities for each entity in a sample (see Fig. 7.1 and SI chapter 7: boxes S1-S3). Predictions of these relative abundances for each sample (i.e. each local community) are such that they need to agree with any constraint we might know, e.g. functional traits or the abundance distribution of taxa in the total sample (i.e. the regional metacommunity), but are otherwise maximally uninformative. In other words, no other constraints beyond those we know are implied on these probabilities. If demographic rates are indeed determined by heritable traits in specific environmental context (a deterministic view), over time specific traits leading to higher fitness will lead to greater relative abundances of taxa possessing these traits within these environments. If this assumption holds true, we should be able to derive accurate predictions of local community composition from pure trait effects based solely on community-weighted means of these taxa-specific traits. If relative abundances are, however, more driven by the limited migration of individuals regardless of functional traits (a neutral perspective), the regional relative abundances should provide a better prediction relative to using only the traits as constraints (i.e. a pure metacommunity effect). Given the many studies on dispersal kernels and the accompanying distance decay of similarity (65) the effect of the regional relative abundances of taxa should also be inversely related to geographic distance from the sample for which predictions of relative abundance distributions are made using the MEF. Although it is nearly impossible to include all traits that might be important as constraints, here we use a comprehensive list of functional traits reflecting a broad spectrum of ecological challenges (Table 7.1). Hence, any unidentified causes of variation in abundances should be left to demographic stochasticity or ecological challenges not captured in these functional traits. And even though there are many different tests available to link trait variation to abundances, turnover between habitats or environments and the distance decay of similarities between samples, they cannot quantify the importance of these constraints relative to each other. The MEF, however, is capable of and designed to do exactly this.

Here we present the application and adaptation of the MEF to quantify signals of natural selection over time and migration from a regional pool without making any a priori assumptions regarding community dynamics. We apply this formalism to the largest and richest rainforest of our world, the Amazon, using a tree inventory database well over 2000 plots (14, 41). We first quantify the relative importance of niche and neutral processes in structuring community composition for different forest types using the MEF and secondly identify which, if any, traits are most important in structuring composition for the different forest types. Finally, we provide estimates of the actual potential metacommunity size of Amazonian tree genera for local communities by studying spatial patterns of relative metacommunity effect to trait effects. The above allows us to significantly advance the study of relative importance of niche versus neutral processes.

Results

Pure trait based filtering of community composition accounted for 21% on average of the information contained in the observed relative abundances for the total dataset. Filtered by forest type this was (on average) 35% for white sand forests, 23% and 21% for várzea and igapó, 33% for swamp and 19% for terra firme forests (see SI chapter 7: Table S1 for a detailed decomposition). Dispersal filtering based on the metacommunity prior (i.e. a neutral prior, taking total Amazonian abundances but not the traits into account) accounted for on average 56% for the combined dataset (in the same order for the separate forest types this yielded 51%, 50%, 53%, 53% and 58%). The hybrid model (including both traits and the metacommunity prior) performed slightly better for the combined dataset (average 60%) and for each specific forest type separate at 60%, 55%, 56%, 56% and 62%. The above is also reflected in the predictive ability of the maximum entropy model for the observed relative abundances. Using only the functional traits as constraints for the MEF calculations and a uniform prior (i.e. a metacommunity without structure) this resulted in a Pearson's R^2 of 0.39 whereas using both the traits and the neutral prior this increased to 0.68. For the regional summed genus pool the increase in predictive ability was even higher, yielding Pearson's R^2 values of 0.27 and 0.99 respectively. This shows a very high predictive ability of the maximum entropy model when taking both traits and the regional abundances as prior into account (SI chapter 7: Fig. S1).

When inferring biologically whether niche or neutral processes are more important, the explanatory power is regarded relative to the model bias, to compensate for any relationship between regional abundances (i.e. relative abundances in the total metacommunity) and local trait constraints (see also Fig. 7.1 and SI chapter 7: boxes S2 and S3). This lowered the proportion of information accounted for considerably and yielded average pure metacommunity effects of 44% for the overall dataset and for each forest type: 29%, 37%, 40%, 29% and 47%. The pure trait effect, although also lowered substantially, did appear to be strongly dependent on forest type. When taken relative to effects of demographic stochasticity (i.e. the unexplained effects and model bias), the pure trait effects accounted for only 5% for the combined dataset on average with for each forest type 11%, 6%, 3%, 5% and 4% (Fig. 7.2 and SI chapter



	Sp1	Sp2	Sp3	……	Sp_i
A	0.37	0.50	0.50	…	0.09
B	0.26	0.38	0.46	…	0.14
C	0.33	0.41	0.43	…	0.05
D	0.26	0.53	0.46	…	0.09
E	0.11	0.12	0.21	…	0.86
F	0.30	0.35	0.57	…	0.14
G	0.35	0.56	0.43	…	0.18
H	0.37	0.00	0.46	…	0.14
I	0.37	0.00	0.43	…	0.09
metacom	0.96	0.75	0.85	…	0.30

	Sp1	Sp2	Sp3	……	Sp_i
Trait 1	T1,1	T1,2	….		….
Trait 2	T1,2	….	….	….	….
Trait 3	….	….	….	….	….
……..	….	….	….	….	….
Trait j	….	….	….	….	Tij

The MEF predicts relative abundances of each of the species for each local community using only a species abundance (above) and trait matrix (below). This is achieved by finding the unique vector of relative abundances maximizing entropy (1) subject to known constraints (2). The solution is a generalized exponential distribution (3) where the λ values measure the importance of each trait when all other traits are constant.

1) $RE = -\sum_{i=1}^{S} p_i \ln\left(\frac{p_i}{q_i}\right)$

2) $\overline{t_j} = \sum_{i=1}^{S} o_i t_{ij}$ and $\sum_{i=1}^{S} p_i = 1$

3) $p_i = \frac{q_i e^{\sum_{j=1}^{n} \lambda_j t_{ij}}}{\sum_{i=1}^{S} q_i e^{\sum_{j=1}^{T} \lambda_j t_{ij}}}$

The model fits between predicted and observed relative abundances are given by the R^2_{KL} values for each specific step of de model (see Supporting Information box S2) and is given by the generalized form of:

4) $R^2{}_{KL} = 1 - \frac{\sum_{j=1}^{c} \sum_{i=1}^{S} o_{ij} \ln\left(\frac{o_{ij}}{P_{ij}}\right)}{\sum_{j=1}^{c} \sum_{i=1}^{S} o_{ij} \ln\left(\frac{o_{ij}}{Q_{i,0}}\right)}$

Box 1 and box S1 provide more detail on the terms of MEF and the decomposition of the different R^2_{KL} models.

NEUTRAL

A	B	C
D	E	F
G	H	I

Expected abundance of the i^{th} species in the absence of trait based environmental filtering is *on average* the metacommunity abundance combined with random variation: relative metacommunity abundance of each species gives neutral prior (*Qi*)

With infinite migration and no selection, there is no relation between traits and abundances. Expected abundance of the i^{th} species (p_i) is given by the neutral prior (q_i) as $\sum_{j=1}^{n} \lambda_j t_{ij}$ becomes 0 and $e^0 = 1$ and (3) reduces to q_i. Hence there will be little divergence between expected (p_i) and observed (o_i) relative abundances (only noise)

With limited migration and no selection, there is still no relation between functional traits and abundances. However, the expected abundance of the i^{th} species (p_i) deviates from the neutral prior (q_i) due to ecological drift and divergence between expected and observed abundances increases, measured by (4). Pure metacommunity effect will depend on the amount of migration.

DETERMINISTIC

A	B	C
D	E	F
G	H	I

Actual abundance of the i^{th} species is pushed away from the neutral expectation due to environmental filtering. The direction depends on the distribution of trait values and the selective environment. It is approximated by the community-aggregated traits: given by the *CWM* values ($\overline{t_j} = \sum_{i=1}^{S} o_i t_{ij}$). Importance of each trait is found by solving the Lagrange multipliers from equation (3)

Assuming infinite migration but extreme selective filtering on trait J in sample E with species i having the highest value of trait J, species i-1 the second highest etc. the abundance distribution might be reversed and CWM values of trait J will have strong relations with abundance of each species. Lambda values of trait J will be high with pure trait effects depending on the other lambda values relative to the metacommunity effect and model bias

Assuming limited migration and still extreme selective filtering on trait J in sample E with equal values as above, the abundance distribution will now also be influenced by ecological drift as influx of species from other samples will be low. Depending on the relative amount of ecological drift in concert with trait selection will determine pure trait and pure metacommunity effect as measured by values from the specific model of (4).

Fig. 7.1 Community assembly and the Maximum Entropy Formalism. Schematic depiction of the ingredients and workings of the MEF in relation to community assembly processes. Left panel shows a species abundances per sample matrix and a functional trait matrix per species, bottom half of the panel outlines the MEF basic calculations. Middle and right panel show different scenarios of either neutral or deterministic dynamics under infinite or limited migration. Dashed lines indicate local abundance distribution, solid line indicates the (summed) regional abundance distribution.

Table 7.1. Overview of the functional traits used as constraints. Mean and standard deviation (SD) are calculated on the data after the predictive mean matching algorithm (percentage of estimated values is given by MICE (%)). Associated challenge indicates different aspects of life history and selective environment that are related to the specific functional traits, sources are given in the table. For specific methodology of measurement protocols and calculation for each trait we refer to the original sources of the data: Adalardo de Oliveira (unpublished data), L. Poorter (unpublished data), J. Lloyd (unpublished data), Van der Sande and Mazzei (unpublished data), Van der Sande and Poorter (unpublished data), (272), (273), (274), (227), (275), (276), (277), (278) and the TRY database (279).

Functional trait	Units	Mean	SD	MICE (%)	Associated challenge
Wood density (*WD*)	g/cm^3	0.625	0.170	0.30	Longevity(*15*)
Seed Mass Class (*SMC*)	categorical (1-8)	4.272	1.377	0.31	Dispersal, Fecundity, Establishment(*15*
Specific Leaf Area (*SLA*)	mm^2/mg	15.019	5.922	0.41	Establishment, Plasticity, Disturbance(
Leaf nitrogen content (*N*)	mg/g	22.310	7.290	0.41	Photosynthetic capacity(*15*)
Leaf phosphorus content (*P*)	mg/g	1.018	0.773	0.50	Limited available P for metabolism(*16*)
Leaf carbon content (*C*)	mg/g	466.642	38.131	0.54	Herbivore resistance (C:N)(*17*)
Latex	1=no, 2 =yes	0.239	0.427	0.46	Herbivore resistance(*18*)
Resin	1=no, 2 =yes	0.143	0.351	0.58	Herbivore resistance(*18*)
Root Nodules (*Nodules*)	1=no, 2 =yes	0.084	0.278	0.00	Nitrogen fixation(*19*)
Ectomycorrhiza (*EctoMyco*)	1=no, 2 =yes	0.011	0.105	0.00	Organic N fixation (*20*), heavy metal pc
Aluminum accumulation (*AlAcc*)	1=no, 2 =yes	0.046	0.210	0.03	Heavy metal pollution(*22*)
Fleshy Fruits (*Fleshy*)	1=no, 2 =yes	0.574	0.495	0.07	Dispersal (*specificity*)(*23*)
Winged seeds (*Wings*)	1=no, 2 =yes	0.224	0.417	0.39	Dispersal (*limitation*)(*23*)

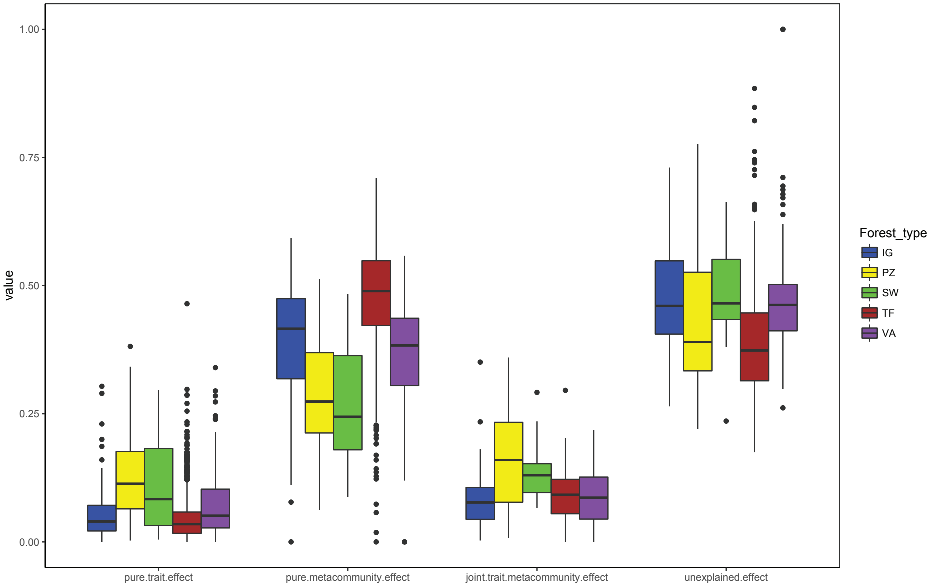


Fig. 7.2. Visual representation of pure trait, pure metacommunity, hybrid model and the remaining unexplained information for each separate forest type. Abbreviations indicate different types: white sand (PZ), várzea (VA), Igapó (IG), swamp (SW) and terra firme (TF). Boxplots show the median value of each pure effect over all samples (i.e. inventory plots), with lower and upper hinges corresponding to the first and third quartiles (the 25th and 75th percentiles). Whiskers extends from the hinge to the largest or smallest value no further than 1.5 * IQR from the hinge. Points beyond this range are plotted individually.

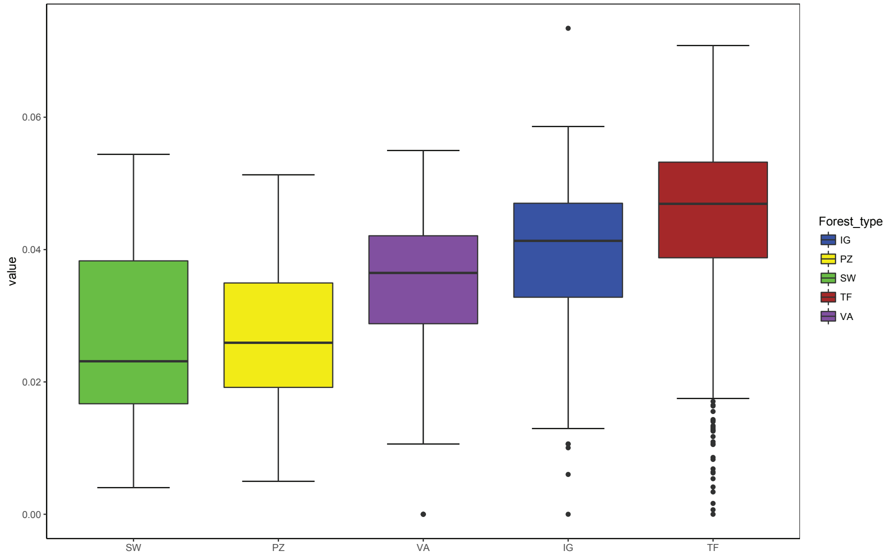


Fig. 7.3. Boxplots showing ratio of pure metacommunity to pure trait effect by forest type, ordered by median value. Ratio was calculated per plot by dividing the pure metacommunity effect by (10*pure trait effects+1). High values indicate a high relative importance of metacommunity effects, and low values indicate a high relative importance of trait effects. A clear pattern can be seen, from with white sand and swamp forests having the lowest pure metacommunity and trait effects ratio, indicating stronger trait effects relative to the metacommunity effects and vice versa for terra firme and igapó forests.

7: Table S1). We also see a clear trend dependent on forest type when pure trait and metacommunity effects are both taken into account and looked at relative to each other (Fig. 7.3). White sand and swamp forests clearly have weak metacommunity effects relative to the trait effects (indicating a stronger selective environment), whereas terra firme forests show the opposite with a stronger metacommunity to trait effect. Spatially, results also indicate that for the Amazonian interior, the influx of taxa from surrounding areas as reflected by the pure metacommunity effect is more important in explaining variation of Amazonian composition than trait effects, whereas trait effects are more important along the edges (Fig. 7.4). Most likely this is not due to differentiation of climatically driven regional patterns as this is taken into account in the four step MEF model calculations and would yield stronger trait effects, even at large spatial scales.

Direction and strength of selection. The absolute value and sign of the lambda values for traits (i.e. the relative strength of the effect on local abundance, see methods) indicate both the strength and direction of selection. Positive values indicate that taxa having these traits also have higher abundances, whereas negative values indicate the opposite. Comparisons between forest types showed that all traits showed strong significant difference when compared between forest types (Fig. 7.5). A number of functional traits associated with low nutrient conditions (e.g. ectomycorrhiza) and life history strategies suited for protection against herbivores (e.g. latex and high leaf C content) were clearly subjected to positive selection in nutrient poor environments (white sand), indicated by the positive lambda values. In contrast, having fleshy fruits and high leaf N and P content were clearly negatively associated with abundance for these soils. Likewise, the ability to accumulate aluminum was positively selected for on those soils commonly associated with higher aluminum content such as igapó (showing strong effects) and terra firme soils (showing a minor, yet positive effect). In contrast, it was strongly negatively selected for on the other soils, with negative lambda values for white sand, várzea and swamp forests. Traits such as SLA, having nodules or winged fruits also showed strong patterns dependent on forest type.

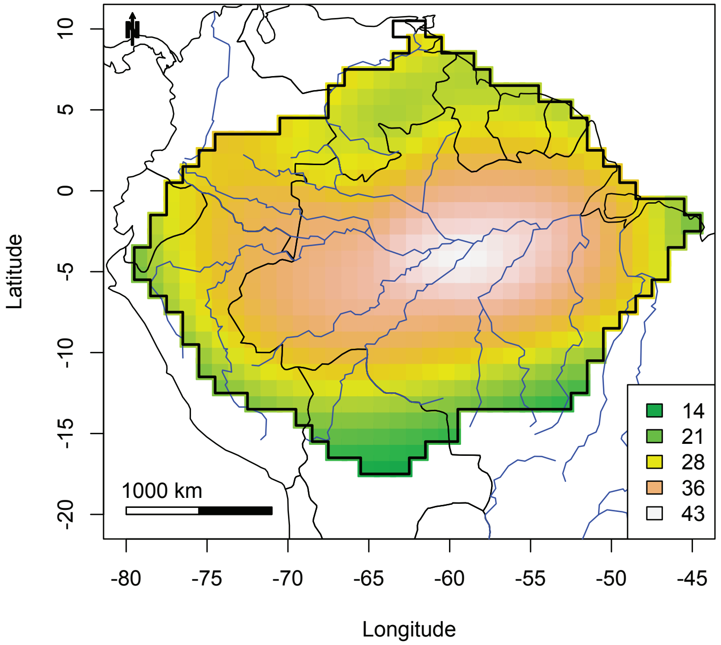


Figure 7.4. Spatial gradient in pure trait relative to pure metacommunity effect. Map showing the ratio between the pure metacommunity effect and the pure trait effects for each plot. Ratio was calculated per plot by dividing the pure metacommunity effect by (10*pure trait effects+1). Values for projection on the map using a loess regression were multiplied by 1000 to allow clearer differentiation. Squares show the predictions from loess regression (color depending on value). Map clearly shows interior of the Amazon having weaker trait effects relative to metacommunity effects whereas on the edges of the Amazon this pattern is reversed.

Metacommunity size. The effect of the regional species pool relative to the model bias and trait effect appeared to be strongly related to geographical distance and dependent on forest type (Fig. 7.6). For the total dataset there was a 19% decrease of the mean information explained purely by the metacommunity prior when the prior was scaled up from 100 (mean 64%, SD 12%) to 3500 km (mean 45%, SD 10%). For the separate forest types, although the initial pure metacommunity effect varied, the decline appeared remarkably similar with a mean 22% decrease in pure metacommunity effect for white sand (50% to 28%), 19% for várzea (59% to 40%), 31% for igapó (71% to 40%) and 37% for swamp forests (57% to 20%) with terra firme forests having a smaller decline of approximately 18% (65% to 47%). Clearly, there is an initial relative steep decline the first 1000 km followed by a shallower decline the next 1500 km. After this second boundary at 2500 km there is virtually no visible decay anymore in pure metacommunity effect, regardless of forest type.

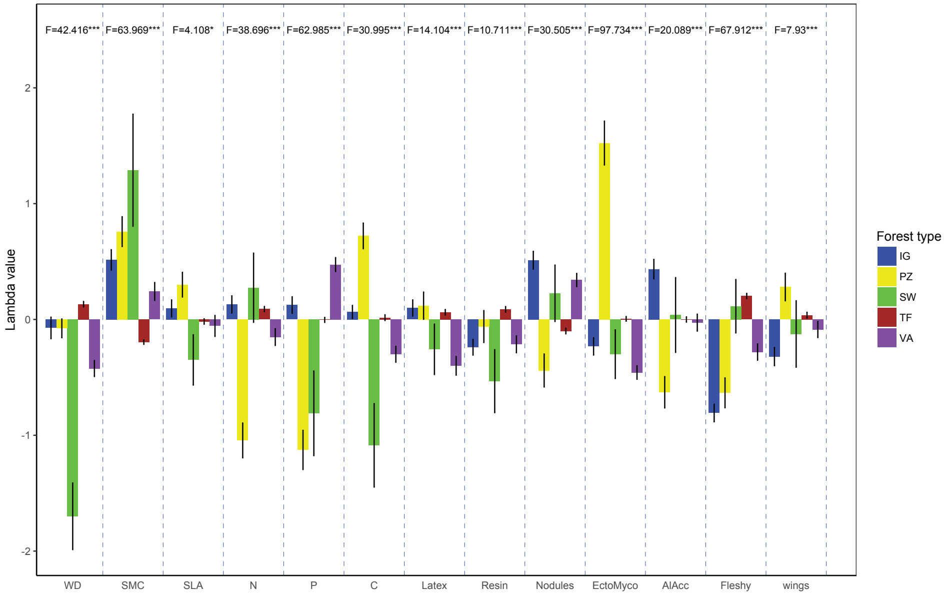


Figure 7.5. Mean lambda values with standard error bars for each functional trait and compared between forest types. Positive values indicate positive selection, reflected in a strong association between higher trait values and higher abundances; negative values reflect the opposite with high trait values associated with lower abundances. Differences between forest types were tested with a one way analysis of variance with significance levels corresponding to: * p < .05, ** p < .01 and *** p < .001. Abbreviations of functional traits stand for WD wood density, SMC seed mass class, SLA specific leaf area, N P and C are nitrogen, phosphorus and carbon leaf content. Latex, Resin, Nodules are presence absence of said traits. EctoMyco ectomycorrhiza, AlAcc the ability to accumulate aluminum, both also presence or absence. Fleshy indicates having fleshy fruits and wings the presence or absence of winged seeds.

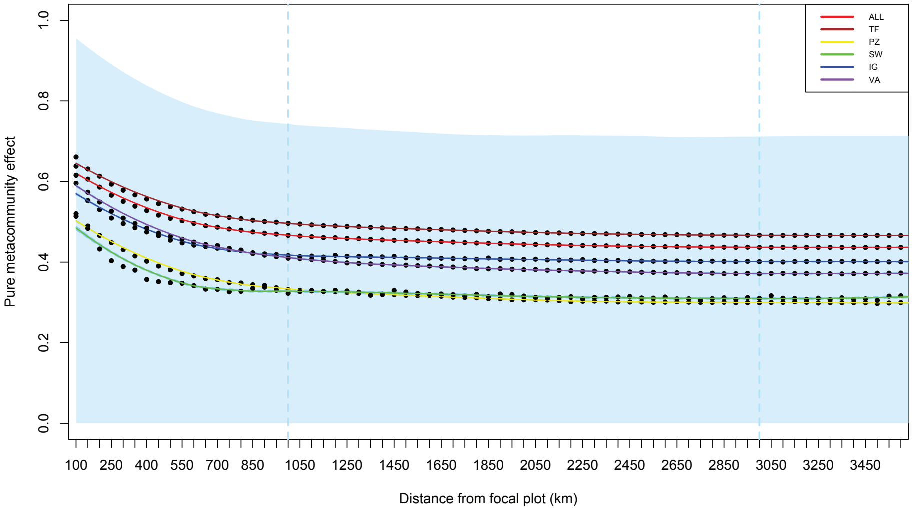


Figure 7.6. Distance decay of pure metacommunity effect. X-axis represents radius of metacommunity prior; i.e. first 100 km consists of just a few plots and at 3800 km all plots are taken into account. Colors indicate different forest types with abbreviations as in main text. Lines indicate predictions from loess regression based on all points. Blue vertical dashed lines indicate 1000 and 2500 km boundary points. Blue shading reflects maximum values for that distance of total dataset.

Projecting these decays geographically using the ratio of pure metacommunity effect at the start (i.e. 100 km) and at the second boundary of decline (i.e. 2500 km) showed a clear pattern: plots in the Amazonian interior have the most gradual declines (yielding higher ratios) compared to plots along the edges of the Amazon (Fig. 7.7). There is an obvious risk that when metacommunity size is increased, this also includes more environmental heterogeneity, potentially confounding results. However, if this were the case, the metacommunity prior (q_i from Fig. 7.1 and SI chapter 7: box S2) would also change. As the pure metacommunity effect is the explained information that remains after correcting for any trait effects and the pure trait effects is the explained information that remains after correcting for the pure metacommunity effect (SI chapter 7: box S3) this confounding effect should then be accompanied by an increase in pure trait effect for each sample. This is not what is observed, not even within the different forest types (SI chapter 7: Fig. S3). Instead, the trait effect gradually goes up and then remains constant.

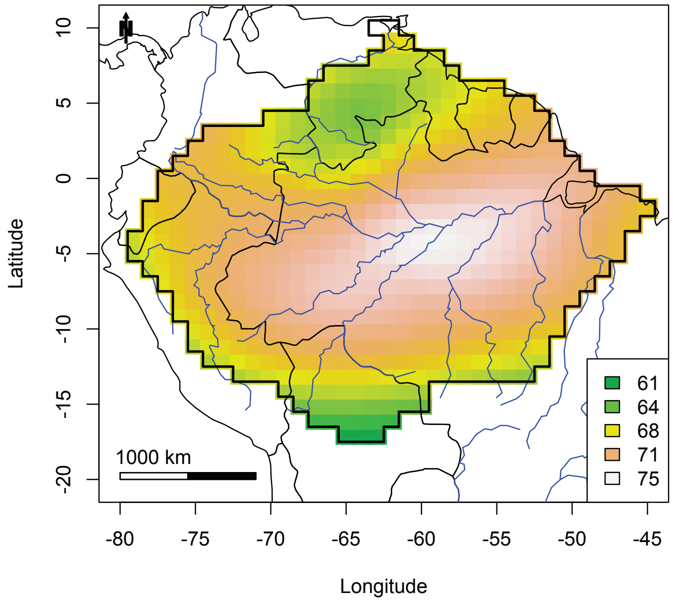


Figure 7.7. Spatial pattern of the distance decay of pure metacommunity effect. Map showing spatial patterns of the inverse of the absolute power law's exponent best describing the distance decay of pure metacommunity effect (Fig. 7.6) for each plot. Values for projection on the map using a loess regression were multiplied by 10 to allow clearer differentiation, legend indicates values of predictions from loess regression (squares). Map clearly shows interior of the Amazon having weaker declines of metacommunity importance over distances whereas on the edges of the Amazon this pattern is reversed.

Discussion

The underlying principles of the MEF follow from a well-founded theoretical body of evolutionary biology (i.e. natural selection of beneficial traits), ecology (i.e. migration of individuals) and population dynamics (40, 79, 251, 252). It enabled us to quantitatively disentangle the dynamics of community structure for tropical forest communities at genus level, i.e. determine the relative importance of niche versus neutral processes and to study their relationship at large spatial scales on genus level taxonomy. Our results show that pure trait based filtering relative to regional abundances explained 11% of forest composition for white sand forests with an average explained proportion of only 5% when all forest types are taken together. The influence of dispersal limitation, reflected by the pure metacommunity effect, was 4 times higher with a maximum of 47% for terra firme forests at an average value of 44% for all forest types combined, almost tenfold larger. The joint information taking both the traits and regional abundances into account showed a maximum explained proportion of 12% for white sand forests and an overall average explained proportion of 7% for all forest types combined. Assuming a sufficient number of functional traits were taken into account, these results clearly indicate that neither neutral nor niche processes, as taken into account by these functional traits, can be solely responsible for community composition. We first discuss our findings in light of explaining community dynamics and hypothesize on what is missing from the niche or neutral perspective; second we discuss this in relation to geographical distance and finally propose what should be added to our ecological toolbox in order to gain a better understanding of community assembly.

Governing dynamics of community composition. Signals of quantitative selection in functional traits caused by long-term evolutionary change was found to be highest for white sand forests, whereas its counterpart in the form of the dispersal mass effect from the regional pool of genera had the second lowest value (only swamp forests had slightly lower values). Looking at other forest types we see the same pattern, where the pure metacommunity effect is always stronger than the pure trait effects. White sand forests, having extremely nutrient poor soils and therefore presumably a much stronger selective environment than any of the other forest types, clearly support a deterministic view of community composition. Terra firme forests, however, reflective of a less strong selective environment in terms of resource availability, showed the opposite, with less than half of the pure trait effect in comparison with white sand forests (even when rarefied to accommodate for different sample sizes). In addition, white sand forests have a smaller connected surface area and accompanying smaller number of genera in comparison with terra firme forests, adding to the calculated stronger trait effects (14).

It should be noted that within forest type heterogeneity was not taken into account as this was mixed into a single environmental class. This might cause an underestimation of the deterministic effect but as of yet cannot be corrected for at this scale. Detailed analyses of lambda values also gave indications which traits were important for selective advantages between forest types. The strength and direction of selection indicated a clear selective pressure for a different life history strategy of growth versus protection. Traits associated with protection against herbivores such as latex (253) and high leaf carbon content were clearly positively associated with higher abundances on white sand soils, whereas traits indicative of soil fertility, investment in growth and photosynthetic ability such as high foliar concentrations of P and N (254) showed strong negative associations. The ability to accumulate aluminum was also strongly positively associated on the more nutrient but also often aluminum enriched soils of terra firme and igapó. There were also traits that showed no specific (strong) signal of selection on certain forest types (either positive or negative), such as wood density on terra firme or ectomycorrhiza on várzea (see Fig.

7.5 for all lambda values). Interestingly, terra firme also showed the smallest lambda values overall (either positive or negative). This may be indicative of either more pronounced demographic stochasticity or ecological drift eliminating the association between traits and relative abundance or lower effects of selection in general. Again, this might also be due to mixing heterogeneous microenvironments into a single environmental class. Support for such heterogeneity within terra firme forests having influence on distribution of functional traits on valleys or plateaus has recently been found (255). It should be noted that a significant part of the trait data was estimated using the predictive mean matching (Table 7.1) which also might account for counterintuitive signals such as the positive selection of SLA on nutrient poor white sand soils. The general pattern, however, does indicate a clear signal of differential trait selection between forest types relative to effects of the metacommunity for a number of functional traits. This pattern is most likely a combined effect of differential abiotic environments resulting in quantitative selection of traits over time (i.e. strong filtering for white sand and swamp forests) and a scale dependency of the potential pool of recruits resulting in differential influence of the regional pool (i.e. large for terra firme forests but smaller for white sand forests where the selective environment has more influence). This is also supported by the variation in strength and sign of the lambda values (Fig. 7.5).

Explaining the unexplained effects. The relatively strong selective filter of white sand forests is supported by the fact that these forests also had a relative low unexplained proportion of 48% while having the lowest metacommunity effects. For other forest types this unexplained proportion was quite high with a maximum of 56% for swamp forests and an average of 44%. These unexplained effects could be attributed to two separate causes. First, demographic stochasticity could weaken any link between functional traits measured and regional abundances. This would mean almost half of the information contained in relative abundances are the result of random population dynamics and are not structurally governed. The alternative and perhaps more likely hypothesis is that this is due to important functional traits, reflective of processes not taken into account in this study. Traits such as SLA, nodulation, leaf C and N concentration or wood density and seed mass are likely reflective of differential life history strategy associated with competition effects in terms of growth (35–37) with latex or resin often associated with protection against herbivory (253). The still relatively high values of unexplained information, however, indicate this competition effect may be less important in driving community composition in comparison with other processes for which we had no functional traits, such as the co-evolutionary arms races with pests and pathogens (58, 59, 230). The first supports results from earlier studies, finding no relation between similar traits as used in this study and the hyperdominance of species (14) as well as findings that similar species can reach high dominance in different habitats (215) while the second argues for perhaps more detailed study of trait combinations (258). It might also be the case that these functional traits are in fact not as good predictors of competitive ability at all (259, 260). However, in light of the large amount of functional traits and that despite low proportions of explanatory power there were strong signals of selection of these traits, they can be used to approximate such life history strategies. The above therefor argues for other processes in concert with resource competition or herbivore defense to play important roles in structuring communities.

Spatial effects of metacommunity importance. Although the initial explanatory power of the metacommunity prior might differ between forest types, the decay pattern is very similar for each forest type (Fig. 7.6). Only for terra firme the first decline appeared to be more gradual. This arguably is related to these forests having the largest relative surface area of the Amazon, giving these forests potentially an almost continuous metacommunity without gaps allowing for a gradual decline. The overall pattern, however, still remains the same, with an initial steep decline up to approximately 1000 km followed by a slow gradual decrease stabilizing after roughly 2500 km. Even terra firme forests adhere to this first 1000 km boundary. Our findings provide a quantitative and mechanistic explanation for the often-observed distance decay in similarity of tropical forests where we see almost the exact same pattern (SI chapter 7: Fig. S2). As the effects of either traits or the metacommunity are measured in the goodness-of-fit predictions on local relative abundances, this implies that at small spatial scales the surrounding regional abundances provide better estimators than functional traits and at larger scales this shifts to the traits. At small spatial scales, local communities share similar environmental conditions leaving only dispersal and drift acting in changing community composition, at least for genus level taxonomy. Again, for species level analyses any micro environmental gradients might prove to show selection at local scales (255, 261, 262), but as of yet this high resolution data is not available. As the potential regional pool is increased, however, more and more environmental heterogeneity is introduced and differences in composition are being driven more and more by differences in selective pressures compared to direct dispersal, lowering the pure metacommunity effect (although it still remains relatively high at large distances). This spatial trade-off of course only holds for the traits used in this study, differences in microhabitat or herbivore/pathogen composition might change this relationship. Nevertheless, due to limited dispersal there is an apparent clear switch between migration and environmental filtering being the most dominant process shaping communities. This effect seems to be universal across forest types and that it does not change anymore at certain distances but instead remains constant indicates the effect of dispersal potentially occurs over very large distances. Crossing these large distances at ecological (short-term) temporal scales could be accounted for by vagrant long distance dispersal (263). Considering these calculations are done at community and genus level and do not measure single dispersal events but rather the effect of long-term dispersal on community composition much deeper in time, we argue there is a strong effect of dispersal at evolutionary (long-term) temporal scales. Values ranging between approximately 20 - 50% metacommunity effect suggest more than a dispersal event every now and then. Instead, it argues for prolonged mixing of forests at large geographical scales. This would indicate Amazonian forests are perhaps more mixed on genus level than expected based on for instance average actual dispersal distances. Such an evolutionary metacommunity is also supported by recent findings demonstrating a lack of geographical phylogenetic structure of lineages for Amazonian tree genera, although this was based on species level calculations (264).

Explaining Amazonian tree diversity patterns. Many models have been proposed for explaining patterns in Amazonian tree diversity. Some rely on very simple null models such as the mid-domain effect (265, 266). Here we show support that high diversity of the Amazon interior could be explained by influx of recruits due to large (overlapping) ranges, causing high mixing as indicated by the still relatively high pure metacommunity effects over large distances. The mid-domain effect, however, would also predict lower species richness for the edges due to lower range overlap (assuming a closed community). This is not the case, as there is a strong species richness gradient from West (rich) to Eastern Amazonian forests (poor) (115). The lower metacommunity effect for the edges then is most likely not due to less absolute influx of genera, rather less influx from the Amazonian tree community. Influx from the species-rich Andes biome could account for the high diversity (267), yet low Amazonian metacommunity effect for Western Amazonian forests. In contrast, South Eastern parts of Amazonia receive influx from species-poor biomes (i.e. the Cerrado) resulting in lower diversity but also low metacommunity effect.

Conclusions

Our results significantly advance the debate between the neutral or niche discussion for at least genus level taxonomy and show that there is a clear spatial and environmental dependency of these two aspects of community assembly. However, there is still much to be explored due to the large unexplained effects and analyses on finer taxonomic scales (i.e. species level) could resolve these issues. The relatively large effects of the regional pool of recruits over great distances do suggest an important role for dispersal and mixing of Amazonian trees in community assembly at evolutionary time-scales (at least for genera), either via stepping-stones or prolonged slow mixing over large distances, especially for the interior of the Amazon.

Materials and Methods

Shipley et al (2006) introduced the ecological application of the MEF with the goal to predict local abundances of taxa within a sample based on information of functional traits and the abundance of taxa in other areas (i.e. regional metacommunity abundances) (79). It is based on what we know and what we do not know regarding these taxa. The first is described by certain constraints and prior information based on empirical data (e.g. values of certain functional traits and the regional abundance of taxa) and by the nature of the states we are looking at (e.g. abundances of the taxa in the local community). The latter is described by the remaining uncertainty, quantified by the entropy; the greater the uncertainty, the greater is the entropy. The purpose of MEF is to find probabilities for all possible states in the system and ultimately the most likely relative abundance of genera in such a way that the distribution still maximizes entropy while continuously agreeing with the constraints and prior information.

Empirical data. The ATDN network (41) consists of over 2000 tree inventory plots distributed over both the Amazon basin and the Guiana Shield, hereafter collectively referred to as Amazonia. Of the entire ATDN database only those plots were used with trees ≥ 10 cm DBH (diameter at breast height) leaving 2011 plots with a mean of 558 individuals per plot identified to at least genus level. Most plots used are 1 ha in size (1414) with 492 being smaller (minimum size of .1 ha) and 105 larger (maximum size of 80 ha). Genus IDs have been standardized to the W3 Tropicos database (268) using the Taxonomic Name Resolution Service (TNRS (94)). After filtering based on above criteria and solving any nomenclature issues 1.121.935 individuals distributed over 828 genera remained.

Functional traits. 13 different functional traits were used as constraints: wood density, seed mass class, Specific Leaf Area (SLA), leaf Nitrogen, Phosphorus and Carbon content and whether genera possessed latex, resin, root nodules, ectomycorrhiza, whether they were aluminum accumulators and whether fruits were fleshy or seeds had wings (for protocol and measurements see the original sources of the data as shown in Table 7.1). Trait values were computed as genus-level means of species values if known within the genus and considered constant for each genus. Genus level of taxonomy was used as the available trait database had the most information on this taxonomic level. Doing calculations on species level would mean assuming many species have the same average trait value as many species specific trait values are unknown, which could confound any results. In addition to effects on relations with functional traits, if there are many species with an extremely restricted range and they do not only have low probabilities but simply zero of reaching some local communities this potentially also leads to an overestimated size of the actual regional species pool and leads to prediction errors. Following an earlier approach, unknown values for traits were estimated by Multiple Imputation with Chained Equations (MICE) using the package *mice* available for the R statistical environment (269). Predictive mean matching (pmm setting) uses all available data as predictors in estimating the most likely values for missing data. All trait values were transformed to Community Weighted Means (CWM)of each trait (J) for each plot (K) as

$$\bar{T}_{JK} = \sum_{i=1}^{S} t_{ij} ra_{ik}$$

with ra the relative abundance of the ith genus in the kth plot, following earlier uses of the MEF (27), but now on genus level. Table 7.1 provides details on functional traits used and reports units of measurement, mean and standard deviations as well as the percentage of estimated values using the predictive mean matching.

MEF Procedure, ecological inference and predictions

Predictions of relative abundances and inference of proportions of explained information for each plot by either traits, the relative abundances of the regional pool or a combination were obtained by applying the *maxent2* function (251), an updated version of the *maxent* function currently in the FD package in the R environment (270). In the MEF, constraints are incorporated using the CWM values, reflective of the traits possessed by the "average individual" in the local community, and the prior information such as the regional relative abundances (see SI chapter 7: box S1 for an overview of important terms). CWM values are assumed to be reflective of continuous selective pressures over time and space. If there is a fitness advantage of having certain traits there should be a strong relation between certain CWM values and the relative abundances. On the other hand, if relative abundances are not determined by the selective advantage of having a specific suit of traits, then knowing these trait values will not give any further information already known from the regional prior. In the most likely case, both are operating at some level and knowing both traits and the regional relative abundances will in this case provide an increase in information for predicting relative abundances in local communities. To decompose these aspects, the MEF procedure is formulated by a four step model with each step randomizing a different aspect (e.g. traits or regional abundances) to determine its effect on abundance predictions: 1) Given genus-specific traits and their CWM, fit the data assuming a uniform metacommunity prior, i.e. each genus is assumed to have an equal abundance on average in the regional pool of recruits. 2) Permutate the fit found in step 1 using the observed relative abundances and the specific traits by randomly shuffling the traits and the genera using the *maxent.test* function (271). 3) Calculate new fit using the CWM plus the observed metacommunity prior, i.e. using the actual relative abundances of genera summed over all samples and finally 4) permutate this fit again similar to step 2 by randomly shuffling traits, genera and their metacommunity prior using the *maxent.test* function. Each step finds predicted relative abundance values for each genus in each local community while maximizing entropy given the CWM values and specific priors for that model.

The proportion of uncertainty in observed relative abundances explained by each model is given by the Kullback-Leibler divergence R^2_{KL} values, a generalization of the classic R^2 goodness of fit values (30). The above four steps and the specific R^2_{KL} values generate all the necessary information to calculate the pure trait, pure metacommunity, joint metacommunity-trait and the unexplained effects as proportions of the total biologically relevant information for each plot (see SI chapter 7 for details: box S2 and S3). Analyses were performed on the entire dataset and the dataset filtered according to forest type (white sand, várzea, igapó, swamp and terra firme). Quantitative predictive ability for predictions from step 1 and 3 for the entire dataset was analysed using a linear least square regression with reported R^2 value equal to the Pearson correlation coefficient between the observed and predicted relative abundances defined as one minus the ratio of the error sum of squares to the total sum of squares. The R^2_{KL} basically has all of the same properties as this standard model R^2 but in addition to being able to quantify the proportion of total information of the dependent variable accounted for in the model it is able to decompose this in the various components of the four step model and is calculated using actual observed and predicted relative abundances instead of the error sum of squares (see also (251)). Because sampling size (i.e. number of plots) differed considerably between forest types (swamp 28, white sand 111, igapó 176, Várzea 277 and terra firme 1419) data were rarefied to the smallest sample size (i.e. 28) and calculations permutated 25 times to identify for any sampling effects. Results from this rarefaction procedure indicated no significant change in results and total dataset was used for all analyses.

Strength and direction of selection

The MEF generates predictions of relative abundance as a function of its traits reflected in the CWM values and a series of constants (λjk: the Lagrange Multipliers). Each multiplier quantifies the association between a unit of change for a particular trait *j* and a proportional change in the predicted relative abundance p_{ik} (the ith genus in the kth community) considering all other traits are constant. Positive values indicate that entities with larger trait values for this specific trait in general also are associated with higher abundances (positive selection), negative values indicate the opposite with higher trait values associated with lower abundances (negative selection). Values more or less equal to zero indicate no true association and hence it could be assumed there has been no selective pressure for this particular trait. Studying these lambda values then gives information on both the strength and direction of selection. Lambda values for each trait were compared between forest types using a One Way Analysis of Variance (ANOVA).

Estimation of metacommunity size

To estimate the potential range of dispersal of genera for each local community, apart from vagrant dispersal, we adapted the MEF procedure to run in a loop. Each iteration, the size of the regional pool is increased in concentric circles of a fixed radius around the local community for each plot. The surface area then covered by the circle and plots therein constitute the prior of regional abundances. If this subset of the actual regional pool is very small (i.e. the first circles) we expect metacommunity effects to be high relative to the effects of traits as the environment will be more homogeneous. This will, however, most likely shift to the relative effect of functional traits being stronger relative to the size of the regional pool as the size of the regional pool of recruits and composition of the metacommunity prior changes relative to the sample. In addition, if there is no strong selection filter operating there should be no difference in this decrease of the metacommunity importance in relation to distance, as they would potentially share the same regional pool. If this, however, does differ strongly between the forest types it would be indicative of selection preventing the sharing of the same pool. Regardless of forest type, at some point the pure metacommunity effect will not change anymore from the previous to the next circle and we assume we have reached the outer range of the potential regional pool of recruits from which point on only vagrant dispersal still causes input of new genera and the effect of the metacommunity stabilizes. The relationship between the pure metacommunity effect and radius of metacommunity size was analyzed using a smoothing loess regression (function *loess* and *predict*; R-package *stats* (51)). The fit from the loess regression was subsequently used to predict values of metacommunity effect based on geographical distance. Again, analyses were performed on both the entire dataset and the dataset filtered to forest type. The ratio of pure metacommunity effect at 100 and 2500 km was then used to project spatial patterns across the Amazon for each plot and interpolated using a loess regression (Fig. 7.7). The same interpolation and projection was done for the ratio between pure trait and pure metacommunity effect (Fig. 7.4).

Acknowledgements

Funding: Daniel Sabatier and Jean-François Molino are supported by an "Investissement d'Avenir" grant managed by Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-0025). Author Contributions: E.T. Pos and H. ter Steege designed the experiment. E.T.Pos wrote R-scripts, analyzed the results and took the lead in writing the manuscript. H. ter Steege supervised the writing and provided regular feedback both for the manuscript and the interpretation of the results. All other authors provided feedback on the manuscript and provided their data from the Amazon Tree Diversity Network. Authors E.T. Pos to L.V. Gamarra provided tree inventory data, authors G. Boenisch, J. Kattge, N. Kraft, A. Levesley, K. Melgaço, G. Pickavance, L. Poorter provided data on functional traits, C. Baraloto, J. Lloyd, A. A. Oliveira and H. ter Steege provided both tree inventory and functional trait data. Competing interests: Authors declare no competing interests. Data and materials availability: R scripts are available upon request from E.T. Pos.

Supporting Information

Authors

Edwin Pos*1,2, Luiz de Souza Coelho3, Diogenes de Andrade Lima Filho3, Rafael P. Salomão4,5, Iêda Leão Amaral3, Francisca Dionízia de Almeida Matos3, Carolina V. Castilho6, Oliver L. Phillips7, Juan Ernesto Guevara8,9, Marcelo de Jesus Veiga Carim10, Dairon Cárdenas López11, William E. Magnusson12, Florian Wittmann13,14, Mariana Victória Irume3, Maria Pires Martins3, Daniel Sabatier15, José Renan da Silva Guimarães10, Jean-François Molino15, Olaf S. Bánki2, Maria Teresa Fernandez Piedade16, Nigel C.A. Pitman17, Abel Monteagudo Mendoza18, José Ferreira Ramos3, Joseph E. Hawes19, Everton José Almeida20, Luciane Ferreira Barbosa20, Larissa Cavalheiro20, Márcia Cléia Vilela dos Santos20, Bruno Garcia Luize21, Evlyn Márcia Moraes de Leão Novo22, Percy Núñez Vargas23, Thiago Sanna Freire Silva24, Eduardo Martins Venticinque25, Angelo Gilberto Manzatto26, Neidiane Farias Costa Reis27, John Terborgh28, Katia Regina Casula27, Euridice N. Honorio Coronado29,7, Juan Carlos Montero30,3, Beatriz S. Marimon31, Ben-Hur Marimon Jr.31, Ted R. Feldpausch32,7, Alvaro Duque33, Chris Baraloto34, Nicolás Castaño Arboleda11, Julien Engel15,34, Pascal Petronelli35, Charles Eugene Zartman3, Timothy J. Killeen36, Rodolfo Vasquez18, Bonifacio Mostacedo37, Rafael L. Assis16, Jochen Schöngart16, Hernán Castellanos38, Marcelo Brilhante de Medeiros39, Marcelo Fragomeni Simon39, Ana Andrade40, José Luís Camargo40, Layon O. Demarchi16, William F. Laurance41, Susan G.W. Laurance41, Emanuelle de Sousa Farias42,43, Maria Aparecida Lopes44, José Leonardo Lima Magalhães45,46, Henrique Eduardo Mendonça Nascimento3, Helder Lima de Queiroz47, Gerardo A. Aymard C.48, Roel Brienen7, Juan David Cardenas Revilla3, Flávia R.C. Costa3, Adriano Quaresma16, Ima Célia Guimarães Vieira4, Bruno Barçante Ladvocat Cintra16,7, Pablo R. Stevenson49, Yuri Oliveira Feitosa50, Joost F. Duivenvoorden51, Hugo F. Mogollón52, Natalia Targhetta53, Leandro Valle Ferreira4, James A. Comiskey54,55, Freddie Draper56,34, José Julio de Toledo57, Gabriel Damasco58, Nállarett Dávila59, Roosevelt García-Villacorta60,61, Aline Lopes16, Alberto Vicentini12, Janaína Costa Noronha62, Flávia Rodrigues Barbosa62, Rainiellen de Sá Carpanedo62, Thaise Emilio63,12, Carolina Levis64,65, Domingos de Jesus Rodrigues62, Juliana Schietti3, Priscila Souza3, Alfonso Alonso55, Francisco Dallmeier55, Vitor H.F. Gomes4,66, Jon Lloyd67, David Neill68, Alejandro Araujo-Murakami69, Luzmila Arroyo69, Fernanda Antunes Carvalho12,70, Fernanda Coelho de Souza12,7, Dário Dantas do Amaral4, Kenneth J. Feeley71,72, Rogerio Gribel73, Marcelo Petratti Pansonato3,74, Daniel Praia16, Jos Barlow75, Erika Berenguer76, Joice Ferreira46, Paul V.A. Fine58, Toby Alan Gardner77, Marcelino Carneiro Guedes78, Eliana M. Jimenez79, Juan Carlos Licona30, Maria Cristina Peñuela Mora80, Carlos A. Peres81, Boris Villa16, Carlos Cerón82, Terry W. Henkel83, Paul Maas84, Marcos Silveira85, Juliana Stropp86, Raquel Thomas-Caesar87, Tim R. Baker7, Doug Daly88, Kyle G. Dexter89,61, John Ethan Householder13, Isau Huamantupa-Chuquimaco23, Toby Pennington32,61, Marcos Ríos Paredes90, Alfredo Fuentes91,92, Jose Luis Marcelo Pena93, Miles R. Silman94, Sebastián Tello92, Jerome Chave95, Fernando Cornejo Valverde96, Anthony Di Fiore97, Renato Richard Hilário57, Juan Fernando Phillips98, Gonzalo Rivas-Torres99,100, Tinde R. van Andel101, Patricio von Hildebrand102, Edelcilio Marques Barbosa3, Luiz Carlos de Matos Bonates3, Hilda Paulette Dávila Doza90, Ricardo Zárate Gómez103, Therany Gonzales104, George Pepe Gallardo Gonzales90, Jean-Louis Guillaumet†105, Bruce Hoffman106, André Braga Junqueira107, Yadvinder Malhi108, Ires Paula de Andrade Miranda3, Linder Felipe Mozombite Pinto90, Adriana Prieto109, Agustín Rudas109, Ademir R. Ruschel46, Natalino Silva110, César I.A. Vela111, Vincent A. Vos112,113, Egleé L. Zent114, Stanford Zent114, Bianca Weiss Albuquerque16, Angela Cano49, Diego F. Correa49,115, Janaina Barbosa Pedrosa Costa78, Bernardo Monteiro Flores116, Milena Holmgren117, Marcelo Trindade Nascimento118, Alexandre A. Oliveira74, Hirma Ramirez-Angulo119, Maira Rocha16, Veridiana Scudeller66, Rodrigo Sierra120, Milton Tirado120, Maria Natalia Umaña Medina49,121, Geertje van der Heijden122, Emilio Vilanova Torre119,123, Corine Vriesendorp17, Ophelia Wang124, Kenneth R. Young125, Manuel Augusto Ahuite Reategui126, Cláudia Baider127,74, Henrik Balslev128, Sasha Cárdenas129, Luisa Fernanda Casas129, William Farfan-Rios94, Cid Ferreira3, Reynaldo Linares-Palomino130, Casimiro Mendoza131,132, Italo Mesones58, Armando Torres-Lezama119, Ligia Estela Urrego Giraldo33, Daniel Villarroel69, Roderick Zagt133, Miguel N. Alexiades134, Karina Garcia-Cabrera94, Lionel Hernandez38, William Milliken63, Walter Palacios Cuenca135, Susamar Pansini27, Daniela Pauletto136, Freddy Ramirez Arevalo137, Adeilza Felipe Sampaio27, Elvis H. Valderrama Sandoval138,137, Luis Valenzuela Gamarra18, Gerhard Boenisch139, Jens Kattge140, Nathan Kraft141, Aurora Levesley7, Karina Melgaço7, Georgia Pickavance7, Lourens Poorter65, Hans ter Steege,101,142

Affiliations

1Ecology & Biodiversity Group, Utrecht University, Padualaan 8, Utrecht, 3584 CH, The Netherlands

2Naturalis Biodiversity Center, PO Box 9517, Leiden, 2300 RA, The Netherlands

3Coordenação de Biodiversidade, Instituto Nacional de Pesquisas da Amazônia - INPA, Av. André Araújo, 2936, Petrópolis, Manaus, AM, 69067-375, Brazil

4Coordenação de Botânica, Museu Paraense Emílio Goeldi, Av. Magalhães Barata 376, C.P. 399, Belém, PA, 66040-170, Brazil

5Programa de Pós-Graduação em Agricultura e Ambiente, Universidade Estadual do Maranhão Cidade Universitária Paulo VI, Avenida Lourenço Vieira da Silva, 1000, São Luís, MA, 65055- 310, Brazil

6EMBRAPA – Centro de Pesquisa Agroflorestal de Roraima, BR 174, km 8 – Distrito Industrial, Boa Vista, RR, 69301-970, Brazil

7School of Geography, University of Leeds, Woodhouse Lane, Leeds, LS2 9JT, UK

8School of Biological Sciences, Yachay Tech, Hacienda San José s/n, San Miguel de Urcuquí, Ecuador

9Keller Science Action Center, The Field Museum, 1400 S. Lake Shore Drive, Chicago, IL, 60605-2496, USA

10Departamento de Botânica, Instituto de Pesquisas Científicas e Tecnológicas do Amapá - IEPA, Rodovia JK, Km 10, Campus do IEPA da Fazendinha, Amapá, 68901-025, Brazil

11Herbario Amazónico Colombiano, Instituto SINCHI, Calle 20 No 5-44, Bogotá, DC, Colombia

12Coordenação de Pesquisas em Ecologia, Instituto Nacional de Pesquisas da Amazônia - INPA, Av. André Araújo, 2936, Petrópolis, Manaus, AM, 69067-375, Brazil

13Dep. of Wetland Ecology, Institute of Geography and Geoecology, Karlsruhe Institute of Technology - KIT, Josefstr.1, Rastatt, D-76437, Germany

14Biogeochemistry, Max Planck Institute for Chemistry, Hahn-Meitner Weg 1, Mainz, 55128, Germany

15AMAP, IRD, Cirad, CNRS, INRA, Université de Montpellier, INRA, TA A-51/ PS2, Bd. de la Lironde, Montpellier, F-34398, France

16Coordenação de Dinâmica Ambiental, Instituto Nacional de Pesquisas da Amazônia - INPA, Av. André Araújo, 2936, Petrópolis, Manaus, AM, 69067-375,

Brazil

17Science and Education, The Field Museum, 1400 S. Lake Shore Drive, Chicago, IL, 60605-2496, USA

18Jardín Botánico de Missouri, Oxapampa, Pasco, Peru

19Applied Ecology Research Group, Anglia Ruskin University, East Road, Cambridge, CB1 1PT, UK

20ICNHS, Universidade Federal de Mato Grosso, Av. Alexandre Ferronato, 1200, Sinop, MT, 78557-267, Brazil

21Departamento de Ecologia, Universidade Estadual Paulista - UNESP – Instituto de Biociências – IB, Av. 24 A, 1515, Bela Vista, Rio Claro, SP, 13506-900, Brazil

22Divisao de Sensoriamento Remoto – DSR, Instituto Nacional de Pesquisas Espaciais – INPE, Av. dos Astronautas, 1758, Jardim da Granja, São José dos Campos, SP, 12227-010, Brazil

23Herbario Vargas, Universidad Nacional de San Antonio Abad del Cusco, Avenida de la Cultura, Nro 733, Cusco, Cuzco, Peru

24Departamento de Geografia, Universidade Estadual Paulista -UNESP – Instituto de Geociências e Ciências Extas – IGCE, Bela Vista, Rio Claro, SP, 13506-900, Brazil

25Centro de Biociências, Departamento de Ecologia, Universidade Federal do Rio Grande do Norte, Av. Senador Salgado Filho, 3000 , Natal, RN, 59072-970, Brazil

26Departamento de Biologia, Universidade Federal de Rondônia, Rodovia BR 364 s/n Km 9,5 - Sentido Acre, Unir, Porto Velho, RO, 76.824-027, Brazil

27Programa de Pós- Graduação em Biodiversidade e Biotecnologia PPG- Bionorte, Universidade Federal de Rondônia, Campus Porto Velho Km 9,5 bairro Rural, Porto Velho, RO, 76.824-027, Brazil

28Center for Tropical Conservation, Duke University, Nicholas School of the Environment, Durham, NC, 27708, USA

29Instituto de Investigaciones de la Amazonía Peruana (IIAP), Av. A. Quiñones km 2,5, Iquitos, Loreto, 784, Perú

30Instituto Boliviano de Investigacion Forestal, Av. 6 de agosto #28, Km. 14, Doble via La Guardia, Casilla 6204, Santa Cruz, Santa Cruz, Bolivia

31Programa de Pós-Graduação em Ecologia e Conservação, Universidade do Estado de Mato Grosso, Nova Xavantina, MT, Brazil

32Geography, College of Life and Environmental Sciences, University of Exeter, Rennes Drive, Exeter, EX4 4RJ, UK

33Departamento de Ciencias Forestales, Universidad Nacional de Colombia, Calle 64 x Cra 65, Medellín, Antioquia, 1027, Colombia

34International Center for Tropical Botany (ICTB) Department of Biological Sciences, Florida International University, 11200 SW 8th Street, OE 243, Miami, FL, 33199, USA

35Cirad UMR Ecofog, AgrosParisTech,CNRS,INRA,Univ Guyane, Campus agronomique, Kourou Cedex, 97379, France
36Agteca-Amazonica, Santa Cruz, Bolivia
37Facultad de Ciencias Agrícolas, Universidad Autónoma Gabriel René Moreno, Santa Cruz, Santa Cruz, Bolivia
38Centro de Investigaciones Ecológicas de Guayana, Universidad Nacional Experimental de Guayana, Calle Chile, urbaniz Chilemex, Puerto Ordaz, Bolivar, Venezuela
39Prédio da Botânica e Ecologia, Embrapa Recursos Genéticos e Biotecnologia, Parque Estação Biológica, Av. W5 Norte, Brasilia, DF, 70770-917, Brazil
40Projeto Dinâmica Biológica de Fragmentos Florestais, Instituto Nacional de Pesquisas da Amazônia - INPA, Av. André Araújo, 2936, Petrópolis, Manaus, AM, 69067-375, Brazil
41Centre for Tropical Environmental and Sustainability Science and College of Science and Engineering, James Cook University, Cairns, Queensland, 4870, Australia
42Laboratório de Ecologia de Doenças Transmissíveis da Amazônia (EDTA), Instituto Leônidas e Maria Deane, Fiocruz, Rua Terezina, 476, Adrianópolis, Manaus, AM, 69060-001, Brazil
43Programa de Pós-graduação em Biodiversidade e Saúde, Instituto Oswaldo Cruz - IOC/FIOCRUZ, Pav. Arthur Neiva – Térreo, Av. Brasil, 4365 – Manguinhos, Rio de Janeiro, RJ, 21040-360, Brazil
44Instituto de Ciências Biológicas, Universidade Federal do Pará, Av. Augusto Corrêa 01, Belém, PA, 66075-110, Brazil
45Programa de Pós-Graduação em Ecologia, Universidade Federal do Pará, Av. Augusto Corrêa 01, Belém, PA, 66075-110, Brazil
46Embrapa Amazônia Oriental, Trav. Dr. Enéas Pinheiro s/nº, Belém, PA, 66095-100, Brazil
47Diretoria Técnico-Científica, Instituto de Desenvolvimento Sustentável Mamirauá, Estrada do Bexiga, 2584, Tefé, AM, 69470-000, Brazil
48Programa de Ciencias del Agro y el Mar, Herbario Universitario (PORT), UNELLEZ-Guanare, Guanare, Portuguesa, 3350, Venezuela
49Laboratorio de Ecología de Bosques Tropicales y Primatología, Universidad de los Andes, Carrera 1 # 18a- 10, Bogotá, DC, 111711, Colombia
50Programa de Pós-Graduação em Biologia (Botânica), Instituto Nacional de Pesquisas da Amazônia - INPA, Av. André Araújo, 2936, Petrópolis, Manaus, AM, 69067-375, Brazil
51Institute of Biodiversity and Ecosystem Dynamics, University of Amsterdam, Sciencepark 904, Amsterdam, 1098 XH, The Netherlands
52Endangered Species Coalition, 8530 Geren Rd., Silver Spring, MD, 20901, USA
53MAUA Working Group, Instituto Nacional de Pesquisas da Amazônia - INPA,

Av. André Araújo, 2936, Petrópolis, Manaus, AM, 69067-375, Brazil
54Inventory and Monitoring Program, National Park Service, 120 Chatham Lane, Fredericksburg, VA, 22405, USA
55Center for Conservation Education and Sustainability, Smithsonian Conservation Biology Institute, 1100 Jefferson Dr. SW, Suite 3123, Washington, DC, 20560-0705, USA
56Department of Global Ecology, Carnegie Institution for Science, 260 Panama St., Stanford, CA, 94305, USA
57Universidade Federal do Amapá, Ciências Ambientais, Rod. Juscelino Kubitschek km2, Macapá, AP, 68902-280, Brazil
58Department of Integrative Biology, University of California, Berkeley, CA, 94720-3140, USA
59Biologia Vegetal, Universidade Estadual de Campinas, Caixa Postal 6109, Campinas, SP, 13.083-970, Brazil
60Institute of Molecular Plant Sciences, University of Edinburgh, Mayfield Rd, Edinburgh, EH3 5LR, UK
61Tropical Diversity Section, Royal Botanic Garden Edinburgh, 20a Inverleith Row, Edinburgh, Scotland, EH3 5LR, UK
62ICNHS, Federal University of Mato Grosso, Av. Alexandre Ferronato 1200, Setor Industrial, Sinop, MT, 78.557-267, Brazil
63Natural Capital and Plant Health, Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AB, UK
64Programa de Pós-Graduação em Ecologia, Instituto Nacional de Pesquisas da Amazônia - INPA, Av. André Araújo, 2936, Petrópolis, Manaus, AM, 69067-375, Brazil
65Forest Ecology and Forest Management Group, Wageningen University & Research, Droevendaalsesteeg 3, Wageningen, P.O. Box 47, 6700 AA, The Netherlands
66Biology Dep., UFAM, Av General Rodrigo Octavio 6200, Manaus, AM, 69077-000, Brazil
67Faculty of Natural Sciences, Department of Life Sciences, Imperial College London, Silwood Park, South Kensington Campus, London, SW7 2AZ, UK
68Ecosistemas, Biodiversidad y Conservación de Especies, Universidad Estatal Amazónica, Km. 2 1/2 vía a Tena (Paso Lateral), Puyo, Pastaza, Ecuador
69Museo de Historia Natural Noel Kempff Mercado, Universidad Autónoma Gabriel Rene Moreno, Avenida Irala 565 Casilla Post al 2489, Santa Cruz, Santa Cruz, Bolivia
70Centro de Biociências, Departamento de Botânica e Zoologia, Universidade Federal do Rio Grande do Norte, Campus Universitário, Lagoa Nova, Natal, RN, 59078-970, Brazil

71Department of Biology, University of Miami, Coral Gables, FL, 33146, USA
72Fairchild Tropical Botanic Garden, Coral Gables, FL, 33156, USA
73Diretoria de Pesquisas Científicas, Instituto de Pesquisas Jardim Botânico do Rio de Janeiro, Rio de Janeiro, RJ, Brazil
74Instituto de Biociências - Dept. Ecologia, Universidade de Sao Paulo - USP, Rua do Matão, Trav. 14, no. 321, Cidade Universitária, São Paulo, SP, 05508-090, Brazil
75Lancaster Environment Centre, Lancaster University, Lancaster, Lancashire, LA1 4YQ, UK
76Environmental Change Institute, University of Oxford, Oxford, Oxfordshire, OX1 3QY, UK
77Stockholm Environment Institute, Stockholm, 104 51, Sweden
78Empresa Brasileira de Pesquisa Agropecuária, Embrapa Amapá, Rod. Juscelino Kubitschek km 5, Macapá, Amapá, 68903-419, Brazil
79Grupo de Investigación en Tecnologías de la Información y Medio Ambiente, Instituto Tecnológico de Antioquia - Institución Universitaria, Calle 78B No. 72A-220, Medellín, Colombia
80Universidad Regional Amazónica IKIAM, Km 7 via Muyuna, Tena, Napo, Ecuador
81School of Environmental Sciences, University of East Anglia, Norwich, NR4 7TJ, UK
82Escuela de Biología Herbario Alfredo Paredes, Universidad Central, Ap. Postal 17.01.2177, Quito, Pichincha, Ecuador
83Department of Biological Sciences, Humboldt State University, 1 Harpst Street, Arcata, CA, 95521, USA
84Taxonomy and Systematics, Naturalis Biodiversity Center, PO Box 9517, Leiden, 2300 RA, The Netherlands
85Museu Universitário / Centro de Ciências Biológicas e da Natureza / Laboratório de Botânica e Ecologia Vegetal, Universidade Federal do Acre, Rio Branco, AC, 69915-559, Brazil
86Institute of Biological and Health Sciences, Federal University of Alagoas, Av. Lourival Melo Mota, s/n, Tabuleiro do Martins, Maceio, AL, 57072-970, Brazil
87Iwokrama International Programme for Rainforest Conservation, Georgetown, Guyana
88New York Botanical Garden, 2900 Southern Blvd, Bronx, New York, NY, 10458-5126, USA
89School of Geosciences, University of Edinburgh, 201 Crew Building, King's Buildings, Edinburgh, EH9 3JN, UK
90Servicios de Biodiversidad EIRL, Jr. Independencia 405, Iquitos, Loreto, 784, Perú
91Herbario Nacional de Bolivia, Universitario UMSA, Casilla 10077 Correo Central,

La Paz, La Paz, Bolivia
92Missouri Botanical Garden, P.O. Box 299, St. Louis, MO, 63166-0299, USA
93Department of Forestry Management, Universidad Nacional Agraria La Molina, Avenido La Molina, Apdo. 456, La Molina, Lima, Peru
94Biology Department and Center for Energy, Environment and Sustainability, Wake Forest University, 1834 Wake Forest Rd, Winston Salem, NC, 27106, USA
95Laboratoire Evolution et Diversité Biologique, CNRS and Université Paul Sabatier, UMR 5174 EDB, Toulouse, 31000, France
96Andes to Amazon Biodiversity Program, Madre de Dios, Madre de Dios, Peru
97Department of Anthropology, University of Texas at Austin, SAC 5.150, 2201 Speedway Stop C3200, Austin, TX, 78712, USA
98Fundación Puerto Rastrojo, Cra 10 No. 24-76 Oficina 1201, Bogotá, DC, Colombia
99Colegio de Ciencias Biológicas y Ambientales-COCIBA & Galapagos Institute for the Arts and Sciences-GAIAS, Universidad San Francisco de Quito-USFQ, Quito, Pichincha, Ecuador
100Department of Wildlife Ecology and Conservation, University of Florida, 110 Newins-Ziegler Hall, Gainesville, FL, 32611, USA
101Biodiversity Dynamics, Naturalis Biodiversity Center, PO Box 9517, Leiden, 2300 RA, The Netherlands
102Fundación Estación de Biología, Cra 10 No. 24-76 Oficina 1201, Bogotá, DC, Colombia
103PROTERRA, Instituto de Investigaciones de la Amazonía Peruana (IIAP), Av. A. Quiñones km 2,5, Iquitos, Loreto, 784, Perú
104ACEER Foundation, Jirón Cusco N° 370, Puerto Maldonado, Madre de Dios, Peru
105Departement EV, Muséum national d'histoire naturelle de Paris, 16 rue Buffon, Paris, 75005, France
106Amazon Conservation Team, Doekhieweg Oost #24, Paramaribo, Suriname
107International Institute for Sustainability, Estrada Dona Castorina 124, Horto, Rio de Janeiro, RJ, 22460-320, Brazil
108Environmental Change Institute, Oxford University Centre for the Environment, Dyson Perrins Building, South Parks Road, Oxford, England, OX1 3QY, UK
109Instituto de Ciencias Naturales, Universidad Nacional de Colombia, Apartado 7945, Bogotá, DC, Colombia
110Instituto de Ciência Agrárias, Universidade Federal Rural da Amazônia, Av. Presidente Tancredo Neves 2501, Belém, PA, 66.077-901, Brazil
111Facultad de Ciencias Forestales y Medio Ambiente, Universidad Nacional de San Antonio Abad del Cusco, Jirón San Martín 451, Puerto Maldonado, Madre de Dios, Peru
112Universidad Autónoma del Beni José Ballivián, Campus Universitario Final Av.

Ejercito, Riberalta, Beni, Bolivia
113Regional Norte Amazónico, Centro de Investigación y Promoción del Campesinado, C/ Nicanor Gonzalo Salvatierra N° 362, Riberalta, Beni, Bolivia
114Laboratory of Human Ecology, Instituto Venezolano de Investigaciones Científicas - IVIC, Ado 20632, Caracas, Caracas, 1020A, Venezuela
115School of Agriculture and Food Sciences - ARC Centre of Excellence for Environmental Decisions CEED, The University of Queensland, St. Lucia, QLD 4072, Australia
116University of Campinas, Plant Biology Department, Rua Monteiro Lobato, 255, Cidade Universitária Zeferino Vaz, Barão Geraldo, Campinas, São Paulo, CEP 13083-862, Brazil
117Resource Ecology Group, Wageningen University & Research, Droevendaalsesteeg 3a, Lumen, building number 100, Wageningen, Gelderland, 6708 PB, The Netherlands
118Laboratório de Ciências Ambientais, Universidade Estadual do Norte Fluminense, Av. Alberto Lamego 2000, Campos dos Goyatacazes, RJ, 28013-620, Brazil
119Instituto de Investigaciones para el Desarrollo Forestal (INDEFOR), Universidad de los Andes, Conjunto Forestal, C.P. 5101, Mérida, Mérida, Venezuela
120GeoIS, El Día 369 y El Telégrafo, 3° Piso, Quito, Pichincha, Ecuador
121Department of Biology, University of Maryland, College Park, MD, 20742, USA
122University of Nottingham, University Park, Nottingham, NG7 2RD, UK
123School of Environmental and Forest Sciences, University of Washington, Seattle, WA, 98195-2100, USA
124Environmental Science and Policy, Northern Arizona University, Flagstaff, AZ, 86011, USA
125Geography and the Environment, University of Texas at Austin, 305 E. 23rd Street, CLA building, Austin, TX, 78712, USA
126Medio Ambiente, PLUSPRETOL, Iquitos, Loreto, Peru
127Agricultural Services, Ministry of Agro-Industry and Food Security, Agricultural Services, Ministry of Agro-Industry and Food Security, The Mauritius Herbarium, Reduit, Mauritius
128Department of Bioscience, Aarhus University, Building 1540 Ny Munkegade, Aarhus C, Aarhus, DK-8000, Denmark
129Ciencias Biológicas, Universidad de Los Andes, Carrera 1 # 18a- 10, Bogotá, DC, 111711, Colombia
130Center for Conservation Education and Sustainability, Smithsonian's National Zoo & Conservation Biology Institute , National Zoological Park, 3001 Connecticut Ave, Washington, DC, 20008, USA
131FOMABO, Manejo Forestal en las Tierras Tropicales de Bolivia, Sacta, Cochabamba, Bolivia

132Escuela de Ciencias Forestales (ESFOR), Universidad Mayor de San Simon (UMSS), Sacta, Cochabamba, Bolivia
133Tropenbos International, Lawickse Allee 11 PO Box 232, Wageningen, 6700 AE, The Netherlands
134School of Anthropology and Conservation, University of Kent, Marlowe Building, Canterbury, Kent, CT2 7NR, UK
135Herbario Nacional del Ecuador, Universidad Técnica del Norte, Quito, Pichincha, Ecuador
136Instituto de Biodiversidade e Floresta, Universidade Federal do Oeste do Pará, Rua Vera Paz, Campus Tapajós, Santarém, PA, 68015-110, Brazil
137Facultad de Biologia, Universidad Nacional de la Amazonia Peruana, Pevas 5ta cdra, Iquitos, Loreto, Peru
138Department of Biology, University of Missouri, St. Louis, MO, 63121, USA
139Department of Biogeochemical Integration, Max-Planck-Institute for Biogeochemistry, P.O. Box 10 01 64, Jena, 07701, Germany
140Functional Biogeography, Max-Planck-Institute for Biogeochemistry, P.O. Box 10 01 64, Jena, 07701, Germany
141Department of Ecology and Evolutionary Biology, UCLA, 621 Charles E. Young Drive South, Box 951606, Los Angeles, CA, 90095, USA
142Systems Ecology, Free University, De Boelelaan 1087, Amsterdam, 1081 HV, Netherlands

* Corresponding author: e.t.pos@uu.nl
† Deceased 01-2018

Box S1. The different ingredients necessary for analyses using the MEF. A glossary of the most important terms used in the MEF analyses and throughout the main text to provide the necessary framework of understanding.

Entities
The basic unit of the MEF model can exist in different states. If the system under study is a collection of genera existing at a site, then each entity is a single genus.

States
Classification of different ways any entity can exist. In the same collection of taxa, states of each entity (i.e. genus) is its specific abundance at that site. Microstates are the exact arrangement in time and space for the states of the entities in the system. Macrostates are the description of entities among the possible states in the system under study without regard to the spatial or temporal arrangement of these entities. I.e. observing a relative abundance distribution, but not the actual dispersal and germination of individuals.

Traits, attributes or properties
Each entity possesses measurable properties whose values will probably differ between states. For example, genera differ in average wood density, seed mass, height etcetera.

Maximally uninformative prior
All the information concerning states before constraints are introduced. Called maximally uninformative as preferably all empirical information is introduced in the form of constraints as to have the maximal gain of information regarding the different traits.

Prior distribution
Prior distribution of expected states for the entities, which can be incorporated as a constraint in addition to the traits, being either the observed relative abundance of each entity in the summed sample (i.e. the metacommunity) or a maximally uninformed (uniform) distribution. The former would be a neutral prior (expected local abundance is equal to the abundance in the larger metacommunity).

Community-weighted means
The average trait value (i.e. measurable property such as wood density) of entities (such as genera) weighted by the relative abundance of each entity at a specific site.

Box S2 Mathematical description of the Maximum Entropy Formalism accompanying Fig. 7.1 of the main text. Left panel shows the necessary ingredients and basic formulation of the Maximum Entropy Formalism. Right side panel shows decomposition of the proportion of total deviance accounted for between observed and predicted relative abundances for each of the four-step solution.

The Maximum Entropy Formalism works on the basis of a conceptual model called the CATS (Community Assembly by Trait Selection) (323) and makes use of three inputs:

i) A trait matrix containing the measured functional traits of each of the S total genera in the total regional pool, these can be of either discrete or continuous form.
ii) A vector of n community weighted trait values, estimating the trait value of the average individual in the local community for each of the traits
iii) A prior probability distribution specifying the potential (hypothetical) contribution of the regional pool of recruits to the structure of local communities.

Using only these three sources of information, the model is able to predict relative abundances in the form of Bayesian probabilities for each of the entities for each local community without assuming any a priori relations or mechanisms. This is achieved by finding the unique vector of relative abundances maximizing entropy:

$$RE = -\sum_{i=1}^{S} p_i \ln\left(\frac{p_i}{q_i}\right)$$

with q_i the prior distribution (i.e. regional species pool abundance) and RE subject to the known constraints:

$$\overline{t_j} = \sum_{i=1}^{S} o_i t_{ij} \text{ and } \sum_{i=1}^{S} p_i = 1$$

The solution is a generalized exponential distribution where the λ values measure the importance of each trait when all other traits are constant:

$$p_i = \frac{q_i e^{\sum_{j=1}^{n} \lambda_j t_{ij}}}{\sum_{i=1}^{S} q_i e^{\sum_{j=1}^{T} \lambda_j t_{ij}}}$$

The final step is to measure the proportion of total deviance accounted for between observed and predicted relative abundances for each of the four-step solution. These are the R^2_{KL} values, a generalization of the classic R^2 index of maximum likelihood estimation using the Kullback-Leibler index (323):

1) $\bar{\mathbf{R}}^2_{\mathbf{KL}}$**(u)**: fit of model bias, the model null hypotheses given a uniform prior (i.e. equal distribution in the regional pool of recruits).

2) $\mathbf{R}^2_{\mathbf{KL}}$**(u, t)**: fit using again a uniform prior but including traits as constraints.

3) $\bar{\mathbf{R}}^2_{\mathbf{KL}}$**(m)**: fit using the metacommunity prior but excluding traits as constraints

4) $\mathbf{R}^2_{\mathbf{KL}}$**(m, t)**: fit using the metacommunity prior and including traits as constraints

The general form of the $\mathbf{R}^2_{\mathbf{KL}}$ divergence is calculated by:

$$R^2{}_{KL} = 1 - \frac{\sum_{j=1}^{c}\sum_{i=1}^{S} O_{ij} \ln\left(\frac{O_{ij}}{P_{ij}}\right)}{\sum_{j=1}^{c}\sum_{i=1}^{S} O_{ij} \ln\left(\frac{O_{ij}}{Q_{i,0}}\right)}$$

With the following parameters:

O_{ij} as the observed relative abundances of the ith genus in the jth community,
P_{ij} the accompanying predicted values for the specific model of the four solution step as described in the main text and,
$Q_{i,0}$ the predicted relative abundances given only the maximum uninformative prior.

Further details on the calculation of all separate R^2_{KL} values and accompanying pure trait, pure metacommunity, joint information and biologically unexplained information can be found in the SI (box S3).

Box S3 Detailed decomposition of the four-step solution from the MEF. Mathematical description of the decomposition based on the constraints and prior distributions (both uniform and neutral) for each of the steps from the four-step solution to measure the proportion of total deviance accounted for by each specific model from one of the four steps.

The purpose of using MEF is to decompose the deviance between observed and predicted relative abundances using the four-step solution as described in the main text. The values generated are described below. The R^2_{KL} value is a generalization of the classic R^2 index of maximum likelihood estimation using the Kullback-Leibler index for a non-linear regression including a multinomial error structure (323, 324). In essence, it is a way of measuring the proportion of total deviance accounted for by that specific model from one of the four steps:

$\bar{R}^2_{KL}(u)$: fit of model bias, the model null hypotheses given a uniform prior
$R^2_{KL}(u, t)$: fit using a uniform prior but including traits as constraints
$\bar{R}^2_{KL}(m)$: fit using the metacommunity prior but excluding traits as constraints
$R^2_{KL}(m, t)$: fit using the metacommunity prior and including traits as constraints

1) The increase in the explained deviance due to traits can be calculated either by

$\Lambda R^2_{KL}(t|\varphi) = R^2_{KL}(u, t) - \bar{R}^2_{KL}(u)$
Increase in explained deviance due to traits beyond that due solely to model bias
or $\Lambda R^2_{KL}(t|m) = R^2_{KL}(m, t) - \bar{R}^2_{KL}(m)$
Increase in explained deviance due to traits beyond contributions made by the meta-community

2) The increase in explained deviance due dispersal mass effects via the metacommunity can be calculated by either:

$\Lambda R^2_{KL}(m|\ \varphi) = \bar{R}^2_{KL}(m) - \bar{R}^2_{KL}(u)$
Increase in explained deviance (if any) due to the metacommunity beyond that due to model bias
or $\Lambda R^2_{KL}(m|t) = \bar{R}^2_{KL}(m, t) - \bar{R}^2_{KL}(u, t)$
Increase in explained deviance due to the meta-community given traits, relative to the explained deviance due only to the traits: i.e. information unique to neutral prior

3) And finally the joint information and the biologically unexplained information:
$\Lambda R^2_{KL}(m+t) = \Lambda R^2_{KL}(m|\varphi) - \Lambda R^2_{KL}(m|t) = \Lambda R^2_{KL}(t|\varphi) - \Lambda R^2_{KL}(t|m)$
Joint information gain, or increase in explained deviance due to both the metacommunity prior and the constraints based on the traits

$1- \Lambda R^2_{KL}(m,t)$
Biologically unexplained variation

From these values the pure trait, pure metacommunity, joint effect and biologically unexplained variation can be calculated by the following calculations:
Pure trait effects: $\Lambda R^2_{KL}(t|m)\ /\ (1-\bar{R}^2_{KL}(u))$
Pure metacommunity effects: $\Lambda R^2_{KL}(m|t)\ /\ (1-\bar{R}^2_{KL}(u))$
Joint metacommunity and trait effects: $\Lambda R^2_{KL}(m+t)\ /\ (1-\bar{R}^2_{KL}(u))$
Unexplained effects: $\Lambda R^2_{KL}(m,t)\ /\ (1-\bar{R}^2_{KL}(u))$

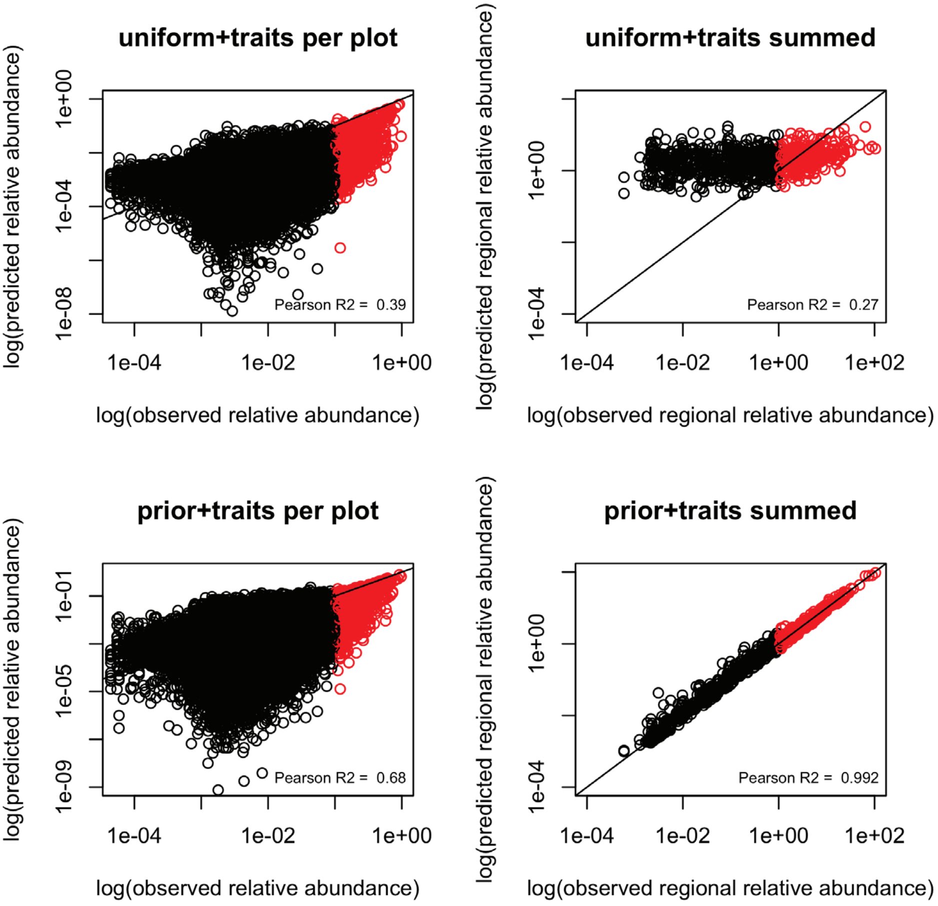


Figure S1. Observed relative abundances plotted against predicted relative abundance per plot (left) and summed (right) using only the traits as constraints in combination with a uniform prior (top) or the hybrid model using both traits and the metacommunity relative abundance as prior (bottom) on a log-log scale. Top figures show predictions for each separate plot and genus, bottom figures show predictions for summed regional abundances. Red points indicate taxa with observed relative abundances over 1e^{-1}. Lines show the x = y prediction and R^2 values correspond to the Pearson's correlation coefficient.

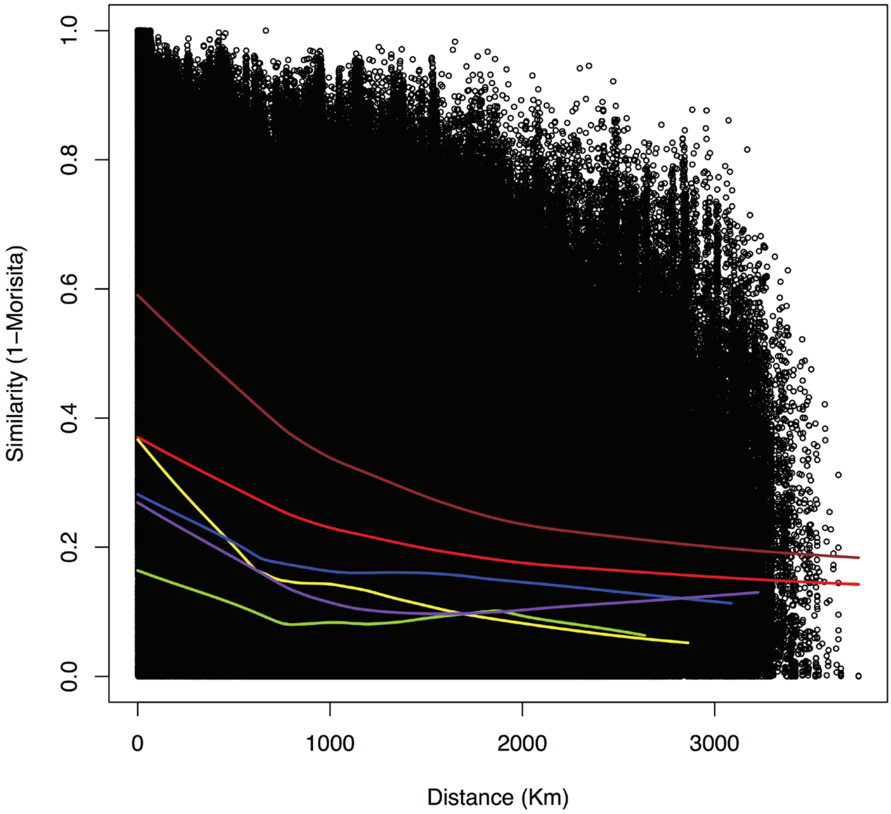


Figure S2. Distance decay of similarity using Morisita index of diversity on genus level for all plots used in the MEF analyses. Points are all plots from the total ATDN forest inventory dataset. Curves indicate LOESS regressions for the different forest types (All combined: red, TF terra firme: brown, PZ white sand: yellow, IG Igapó: blue, VA Várzea: purple, SW Swamp: green).

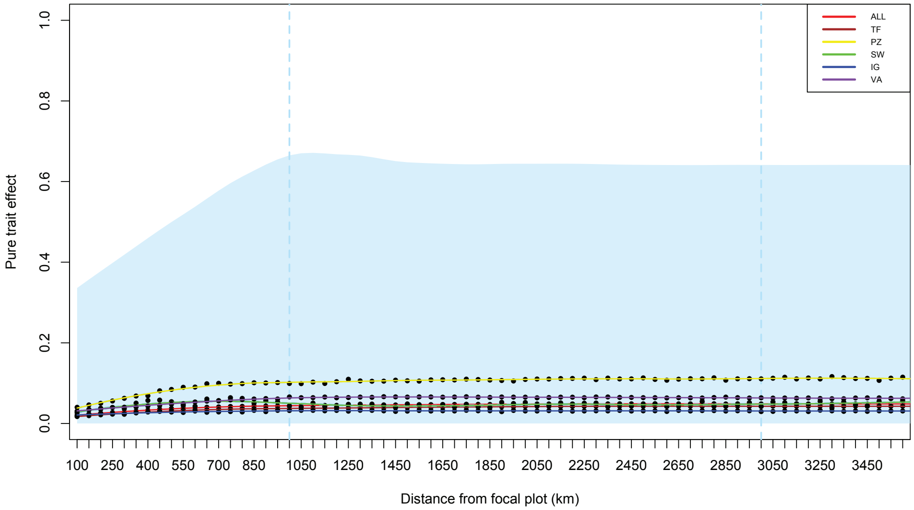


Figure S3. Distance decay of pure trait effect for each forest type separately and the overall dataset. X-axis represents the radius of the metacommunity prior; i.e. the first 100 km consists of just a few plots and at 3800 km all plots are taken into account. Colors indicate the different forest types with abbreviations as in main text. Lines indicate the predictions following from the loess regression based on all points. Blue vertical lines indicate the 1000 and 2500 km boundary points. Blue shading reflects maximum values for that distance of the whole dataset.

Table S1. Decomposition of results from the various maximum entropy models, combined and separated by forest type (PZ podzol, IG igapó, VA várzea, SW swamp, TF terra firme). Top rows indicate the estimated proportions (R^2_{KL} values) of the total information reflective of variation in local relative abundance explained for by the various maximum entropy models. Middle rows indicate the specific information gain from any one of the used models relative to the model bias. Bottom rows show the actual effects of traits, the metacommunity and the joint information relative to the model bias.

	Forest types					
Explained proportions	PZ	VA	IG	SW	TF	Combined
$\bar{R}^2_{KL}(u)$ *model bias fit*	0.1705	0.1284	0.1255	0.2127	0.0878	0.1030
$\bar{R}^2_{KL}(m)$ *pure neutral model fit*	0.5126	0.5013	0.5330	0.5256	0.5810	0.5613
$R^2_{KL}(u,t)$ *pure trait model fit*	0.3495	0.2289	0.2064	0.3279	0.1904	0.2078
$R^2_{KL}(m,t)$ *hybrid model fit*	0.5967	0.5538	0.5587	0.5635	0.6192	0.6029
Increase in explained deviance						
$\Lambda R^2_{KL}(m\|\varphi)$ *metacommunity effect beyond model bias*	0.3420	0.3729	0.4075	0.3130	0.4933	0.4583
$\Lambda R^2_{KL}(t\|\varphi)$ *trait effect beyond model bias*	0.1790	0.1005	0.0809	0.1152	0.1026	0.1048
$\Lambda R^2_{KL}(t\|m)$ *trait effect beyond metacommunity effect*	0.0842	0.0525	0.0257	0.0379	0.0381	0.0416
$\Lambda R^2_{KL}(m\|t)$ *metacommunity effect relative to given trait effects*	0.2472	0.3250	0.3523	0.2357	0.4288	0.3951
$\Lambda R^2_{KL}(m+t)$ *joint contribution of metacommunity and traits*	0.0948	0.0479	0.0552	0.0773	0.0645	0.0632
$1- \Lambda R^2_{KL}(m+t)$ *unexplained effects*	0.4033	0.4462	0.4413	0.4365	0.3808	0.3971
Biologically relevant information						
Pure trait effect *Information from traits, relative to bias*	0.1086	0.0616	0.0313	0.0543	0.0428	0.0482
Pure metacommunity effect *Information from metacommunity, relative to bias*	0.2944	0.3703	0.3983	0.2871	0.4687	0.4368
Joint effect *Information from joint effect, relative to bias*	0.1159	0.0543	0.0650	0.1019	0.0701	0.0705
Unexplained information *Left over information not explained, relative to bias*	0.4810	0.5138	0.5054	0.5567	0.4183	0.4445

Although some emphasize the danger,
I see mainly beauty in the forest.

Hans ter Steege

Chapter Eight

Synthesis

Rolling the dice or struggling for survival
Cheating in life's casino

Imagine yourself in a tropical rain forest, surrounded by numerous trees of different species. Some species you will see almost everywhere, while others you only see occasionally. But what determines if a species is common or rare? Now picture all of these species playing in a casino, at a game called community assembly. The goal of the game is to increase your own abundance, at the expense of others and try to avoid losing individuals yourself. You can either play the game fair and let chance decide your fate or you can cheat at the game, using loaded dice and changing the odds in your favour (40). These loaded dice are an ecological analogy for a deterministic hypothesis, e.g. differences in competitive ability in resource acquirements or a differential ability to escape from pests, pathogens or predators increasing the fitness of individuals of some over those of others. The fair dice on the other hand represent stochastic (neutral) processes, i.e. chance events resulting in ecological drift. Although these are distinct from each other, they are often assumed working simultaneously in some manner (280, 281). Their relative importance and how to determine this has remained unresolved up to this day.

Understanding these rules of community assembly remains important (13, 282, 283), stretching much further than a purely fundamental need towards a practical application in nature management and restoration of degraded systems (284, 285). In light of recent predictions of climate change and the continuous land-use change resulting in severe biodiversity loss (286, 287), this has become even more important and will likely remain so in the coming future. In this dissertation I have tried to further our grasp on community assembly, by developing and studying neutral models at different spatial scales and by applying principles from information theory to quantify signals of selection and stochastic interplay.

My dissertation started with a light primer on neutral theory and the principles of Maximum Entropy (**chapter 2**), to set the stage for the following chapters. I then moved on to validating the use of large-scale data, including unidentified species, showing that large-scale patterns of diversity are extremely robust (**chapter 3**) and providing estimates of diversity in hyper-diverse communities (**chapter 4**). This allowed me to study neutral models in a detailed manner at different spatial scales. My work on necessary input parameters of neutral models revealed these often are aggregated parameters, encompassing much more than just dispersal of individuals, but incorporate many processes influencing betadiversity (**chapter 5**). Following this and using different parameter estimation methods, I worked on adding biological reality to neutral theory and identified scaling issues of neutral theory. It appeared that regional predictions do not necessarily follow from accurate local predictions and vice versa, questioning the interpretation that neutral models provide an accurate reflection of community dynamics (**chapter 6**). However, the main goal of this dissertation was to advance the long-standing debate between niche and neutral proponents, to be able to quantitatively define the realm of selection versus stochasticity. Using the Maximum Entropy principle from information theory and statistical mechanics, I provided this advancement by unequivocally showing just how strong and in which direction selection has influenced composition on genus level taxonomy relative to dispersal and stochastic influences (**chapter 7**). This chapter also studied spatial patterns of dispersal across the Amazon in terms of metacommunity relative to trait importance, and showed that the metapopulation size affecting the community composition of a plot is in the order of $38e^{11}$ km^2, much larger than previously assumed.

The results presented in this dissertation would argue against the evidence produced over the years for communities following neutral dynamics (34). I showed that rather than being accurate predictions, they follow from emergent properties of the models themselves. But what does this mean? Should we discard any prediction made by neutral theory? Should we develop more complex models of neutrality? Are we then no further in understanding community assembly? In this final chapter I synthesize the results of my dissertation to answer these questions and show that although we should tread carefully with respect to earlier made predictions, neutral theory in itself remains a valuable null model similar to the Hardy Weinberg theorem from population genetics (222, 223). We absolutely have gained more understanding of community dynamics by developing such models. I, however, will also show that perhaps it is time to look more with an evolutionary view of community ecology, across trophic levels, if we are to find a single unifying concept of diversity. My work shows that species are in fact playing in the world's casino, but are using loaded dice on crooked tables, resulting from generations of natural selection, differential migration, geographical history and co-evolutionary processes shaping diversity. In a sense, species are indeed cheating an otherwise fair (stochastic) game of life and are also being hindered by other players.

8.1 The robustness of large-scale patterns. In testing hypotheses regarding community dynamics we often focus on large-scale patterns of diversity. In chapter three, I showed that when using a number of different analyses studying patterns of community composition at large spatial scales, these patterns are actually extremely robust when using either a complete dataset or a truncated subset from which unidentified species are removed. Even when simulating a larger fraction of unidentified species by removing a percentage of identified species on top of the unidentified species, major ecological patterns such as the distance decay of similarity still remained identifiable. In other words, the spatial patterns of such community characteristics do not change, even if we know just a little about which species are present. This points to an important conclusion, validating the use of such often-limited datasets. As ecologists, both empirical and theoretical, we inevitably make use of field inventory data whenever we wish to infer explanations regarding community dynamics. There are, however, two important issues to be addressed. First, in many systems, especially the hyper-diverse systems such as the Amazonian rainforests, it is impossible to identify all individuals. Many collections are sterile or either difficult to obtain (e.g. Fig. 8.1). The second issue is that given the sheer size

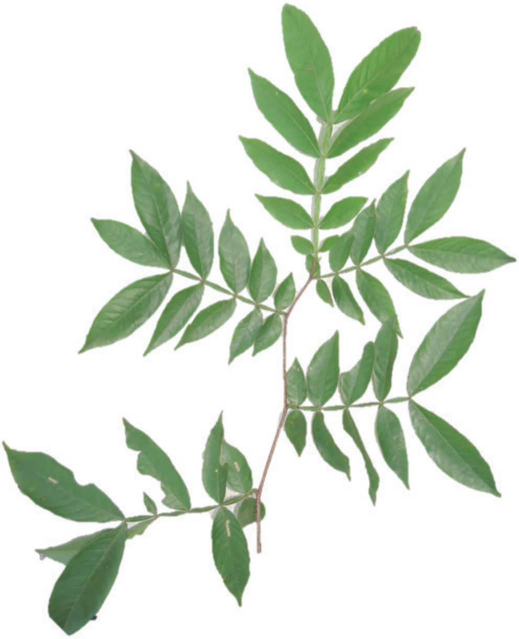


Fig. 8.1 Collection of a species of Inga from the tropical rainforest of Guyana. A clear example of a vegetative collection of which no fruits, flowers or seeds were available.

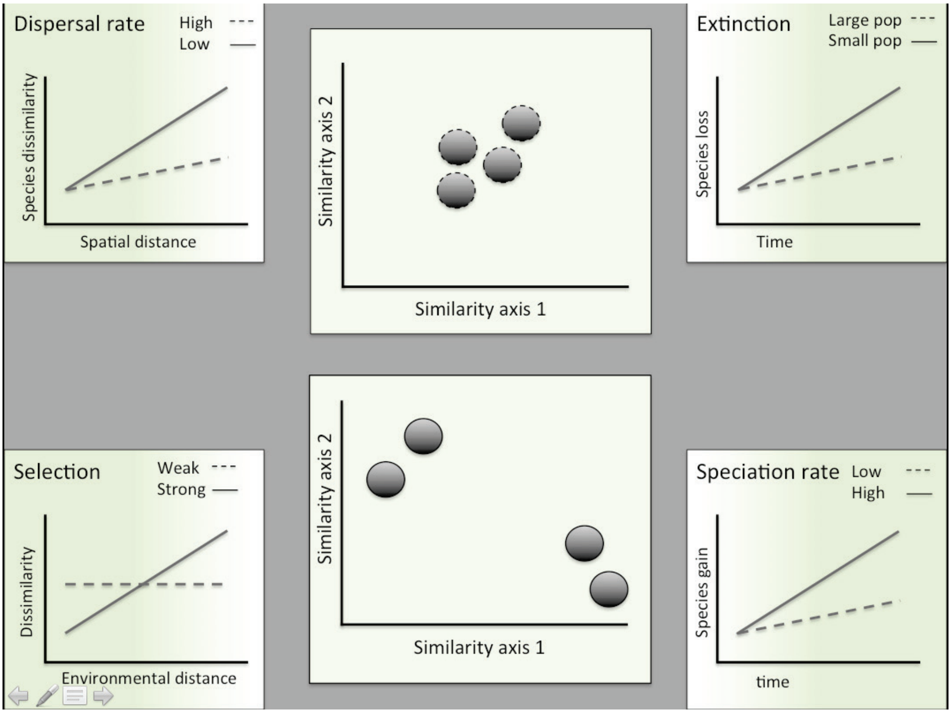


Fig. 8.2 Schematic overview of processes determining similarity between samples. Processes at ecological timescales such as dispersal (top left) and selection (bottom left) or evolutionary timescales such as extinction (top right) and speciation (bottom right) all determine similarity in terms of composition between samples. Centre figures indicate plots that are very similar to each other due to high dispersal: low selection, low extinction and low speciation (separate or in combination) (dashed lines) or the opposite (solid lines) causing low similarity and hence higher betadiversity.

of some communities, such as the 5.6 million square kilometers of the Amazon, it is impossible to approximate this amount of surface area by sampling. Again, using the ATDN as example, after approximately 75 years of plot forest inventory we have only covered little over 2000 hectares, which is approximately 0.00036% of the total area of the Amazon rainforest. Even with botanical collections, given the past and current efforts and the estimated number of tree species in the Amazon it would take at least another 300 years of collecting before we find all estimated species (149). By this time communities might have changed so much we should start over again with sampling the earliest made collections, creating a never-ending cycle just to keep up with all changes. The fact that we find some patterns of community dynamics being extremely robust saves ecologists from having to wait these hundreds of years before doing any analyses. It also means that we can safely use our limited coverage and make inference or interpolate over larger areas, at least for these large-scale patterns of diversity.

However, as chapter three showed, truncating the data either due to actual sampling issues or low quality of data (i.e. having many unidentified collections; see Fig. 3.3) does shroud much information regarding these communities. Those that are not identifiable are not distributed uniformly across the rank abundance distributions, instead it are usually the rare species that are not identifiable or missed in limited datasets. In essence, the truncation transforms the logseries of community structure into a lognormal without the binning trick first used by Preston (122). To play the devil's advocate and to taunt Preston's argument that the lognormal is conceived as the unveiling of a logseries we might even say that the lognormal as imposed on truncated data in some cases actually might be a logseries without knowledge, at least with regard to these details of community structure and composition. Of course, Preston worked with a fully identified dataset so not all lognormal distributions suffer from this effect but this would imply that sometimes much information is still hidden in plain sight when looking only at such large-scale patterns using truncated data and might confound our conclusions regarding local community dynamics. For instance, when neutral theory provides good fits to such regional scale patterns one might think neutrality is the main driver of community structure. If, however, the data was truncated prior to running the model, detailed information that might contain signals of quantitative selection that could have been found at local scales or with these unidentified species is missing. Hence, the interpretation of fits to model output becomes spurious. With the emphasis of neutral theory primarily at regional scales of diversity this does not mean it is incorrect per definition. Nor is it true that previous predictions made by neutral theory, perhaps even using only valid species names, have no value at all. In fact, in my opinion, using robust patterns with limited information available still allows neutral theory to be useful, for instance as a null model. Knowing, however, that neutral theory in many cases provides good fits to empirically observed patterns and can be used even with this limited information still leaves the question whether it can also provide a mechanistic insight into community dynamics.

8.2 Does neutral theory provide insight? One of the major issues in the debate between neutral and niche views of ecology has been that neutral theory, despite its intuitive simplicity, provides such good fits for empirically observed patterns. However, knowing a theory or model provides good fits to empirically observed data does not necessarily mean it also provides mechanistical insight, or as Rosindell and colleagues stated: "*pattern does not imply process*" (288). In other words, one must be careful with interpretation of models depending on the underlying assumptions and input parameters. Looking at neutral theory, in addition to the fundamental biodiversity number theta, migration also forms one its core parameters. Because migration is such a fundamental parameter of any neutral theory model there have been many methods developed for estimating it from empirical data. Strangely enough, estimation of migration even for the spatially explicit models often is done using methods based on the spatially implicit neutral models and has a strong link with methods used to estimate gene flow between populations as measured by genetic differentiation and fixation of specific alleles, e.g. the fixation index (52, 220, 289). Chapter five studied the effect and ability of such methods to accurately estimate migration both from spatially implicit and semi-explicit neutral models. In a way it can be considered a practical review that was necessary before we could implement our own model at larger scales. By introducing a second level of migration from adjacent plots in addition to that from the metacommunity as imposed in the classical UNTB, we studied the behaviour of a number of different estimation methods. We showed that estimation methods based on a spatial implicit model have difficulties to infer migration when this is larger from adjacent plots than from the metacommunity. In other words, estimation was only accurate when migration from the metacommunity outweighed that of adjacent plots. When migration from adjacent plots outweighed that of the metacommunity (which is the case for many taxa) it was in general an underestimate. Only when the spatially semi-explicit model approached a spatially implicit world, with the overarching connection to the metacommunity being the main supply of migrating individuals, these methods proved to be accurate. This can be explained by looking into how these methods actually estimate migration. Many of them are based on compositional differences among all samples to infer the amount of migration, much like the earlier mentioned fixation index from population genetics. In Wright's finite island model (52, 289) this same Fst index can be used to estimate migration rates with fixation of alleles in samples indicating little to no migration whereas no fixation among samples indicating much migration (i.e. panmixis). This can be translated directly to the neutral model of ecology. If all samples are more or less similar in composition and thus are a direct sample from either the whole summed metacommunity or from a hypothetical metacommunity connected to each local community separately, this would imply there is much migration between samples, allowing for homogenization of diversity and lowering betadiversity. However, if there is little migration each local sample differentiates from the whole and betadiversity increases. It should be noted though that as in neutral models each plot will have random species being dominant due to genetic drift, there will be no clear pattern of distance decay in similarity as observed in empirical data, even with little migration. We showed that estimations should be viewed more as an approximation of the homogenization among local communities over time. In other words they approximate how each separate sample reflects the total diversity patterns rather than being an explicit measure of migration and thus they have a direct relationship with beta diversity. But as betadiversity, or the amount of dissimilarity between samples, is the result of many (non)-neutral processes (20 and Fig 8.2), we have to admit that migration, as estimated in a spatial implicit world, encompasses not only direct dispersal but is an ecological aggregate of all of these processes. In other words, this core parameter of the UNTB effectively takes into account everything that might influence betadiversity and cannot possibly reflect only migration of individuals. The parameter *m* of neutral models implemented in such a way then appears more as an emerging property revealed by neutral theory instead of being an effective mechanistic parameter. With this in mind, neutral models should adapt a more direct way of implementing a dispersal mechanism between local communities and set out the use similar sampling schemes to mimic the real world it tries to describe. Chapter six set out to do this by implementing a modified version of Chisholm and Lichstein plot geometry method (138, 139) that was tested earlier in chapter five. With this method, dispersal is measured by a mathematical relation between plot geometry and actual average dispersal distance of individuals instead of estimating dispersal from species composition data (Fig. S4 and Supporting Information chapter five: S3). This allows for a more objective null model approach testing only the direct effect of dispersal of individuals instead of indirectly introducing other (potentially niche-based) processes accounting for patterns in species composition. We used measured species characteristics to infer most likely dispersal distances and implemented these in a newly developed semi-spatially explicit neutral model. Our results showed something rather peculiar, namely that simultaneous prediction of patterns in diversity at different spatial scales were impossible to attain, raising doubts on the interpretability of neutral models, which is the subject of the next paragraph.

8.3 Disagreement between regional and local predictions of neutral models. Many neutral models, either spatially implicit or explicit, have primarily focused on regional scales of diversity and neglected local level contribution, without much attention to scaling up or down. Scaling properties of neutral theory have been addressed earlier (291) but were focused more on how niche dynamics could be masked, depending on scale, and seemingly would make communities act neutrally, while in reality they were not. As earlier results from this dissertation already showed, some large scale patterns are extremely robust (chapter three) and earlier neutral models used a process-aggregated migration parameter (chapter five), we are left wondering whether predictions from neutral models follow from accurate local dynamics. Chapter six combined regional and local results of a newly developed semi-spatially explicit neutral model to test this more fundamental aspect of neutral theory: do regional patterns reflect accurate local dynamics? To this end, we created a model that actually mimics not only accurate dispersal but also sampling schemes often encountered in empirical data. By adding such biological reality to predictions of neutral theory we showed a severe shortcoming in the scaling and hence interpretation of such predictions: no matter how well parameterized they are or how exact their output fits with regional patterns, an accurate simultaneous prediction of both regional and local diversity patterns was impossible to attain. Specifically the dominance of species at local levels could not be predicted accurately, while maintaining regional diversity. This pattern of some species being able to attain high dominance over others is not something that has gone unnoticed (14). Many have proposed several hypotheses explaining this pattern, ranging from resource competition environmental filtering and niche partitioning (292, 293), ecological drift (29) and predator-prey interactions (294). Neutral theory would advocate for severe ecological drift as the main driver of this dominance; continuous local replacement with limited input from outside leading to dominance of some over others as the communities obey zero-sum dynamics. If neutral theory would be an accurate reflection of community dynamics, we should expect such ecological drift at local scales would still result in regional patterns also having good fits to for instance rank abundance distributions. However, I showed in chapter six that the strength of ecological drift necessary to approach patterns of maximum dominance resulted in a severe loss of diversity at regional scales (Fig. 6.3). This disagreement indicates non-neutral processes other than dispersal limitation must be at work allowing for such dominance of species, indicating violation of neutral theory assumptions. Here we can again draw a direct analogy to the Hardy Weinberg theorem from population genetics (222, 223). As a model of the evolution of populations, this theorem puts forward clear assumptions such as no migration, no mutation, no natural nor sexual selection and an infinite population size. It is considered a valuable null model allowing identification of assumption violation and furthering understanding of evolving populations, although a full understanding of interpretation remains difficult (295). Neutral theory has similar clear assumptions such as equal per capita probabilities of birth and death, zero-sum dynamics and recruitment proportional to the relative abundance of species. If all assumptions hold, predictions at both local and regional scales should not deviate from empirically observed patterns. Given the results of chapter six as outlined above it would seem that the assumptions of neutral theory are violated in some manner in the hyper diverse communities of the Amazon rainforests.

8.4 Violation of neutrality assumptions. One of the key assumptions of neutral theory is to treat organisms identical in their probabilities of birth, death, migration and speciation. These probabilities are defined at the individual level, or more specifically to quote Hubbell: "*Neutrality […] is defined as per capita ecological equivalence among all individuals of every species in a given trophically defined community*" (29). Neutrality follows from this ecological equivalence, with all individuals obeying the same rules of community assembly. Other key assumptions of Hubbell's

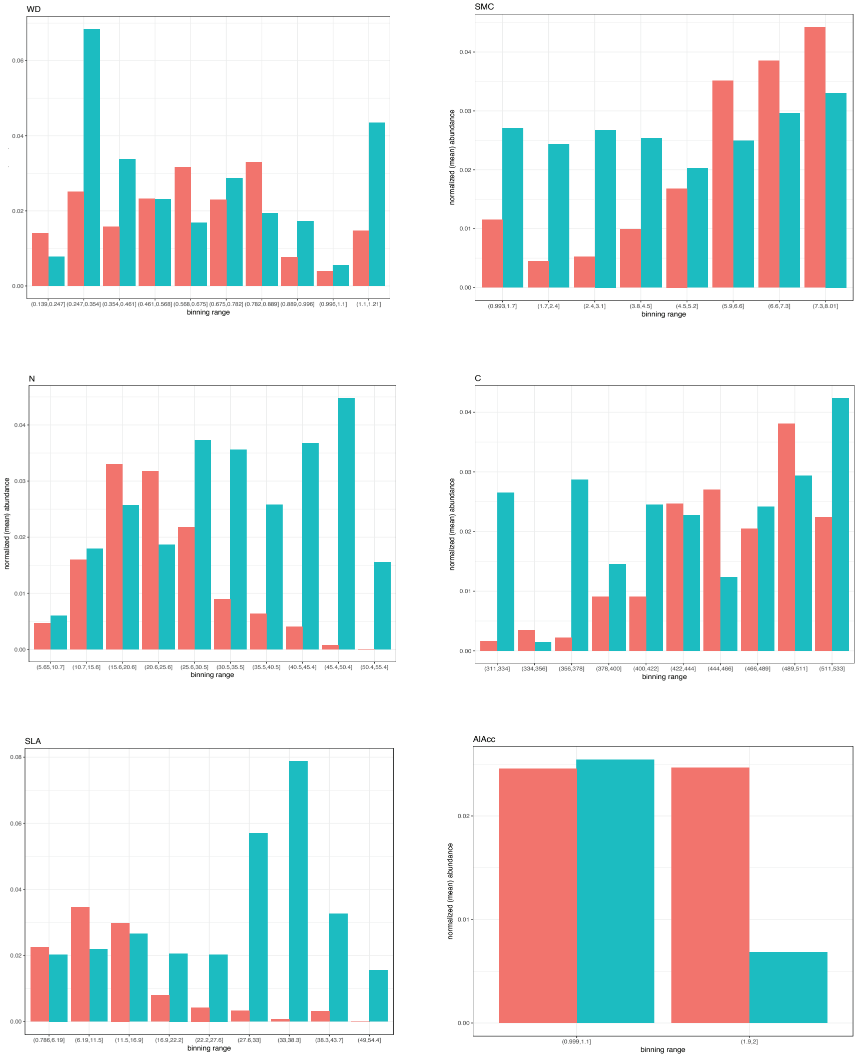


Fig. 8.3 Barplots showing distribution of functional traits in binned classes for either mean abundance on plot level (orange) or overall abundance (blue). WD (wood density), SMC (seed mass class), N leaf nitrogen content, C leaf carbon content, SLA specific leaf area, AlAcc ability to accumulate aluminium (discrete). Distributions show a clear trade-off between life history strategies allowing for high local or high regional abundance depending on functional traits.

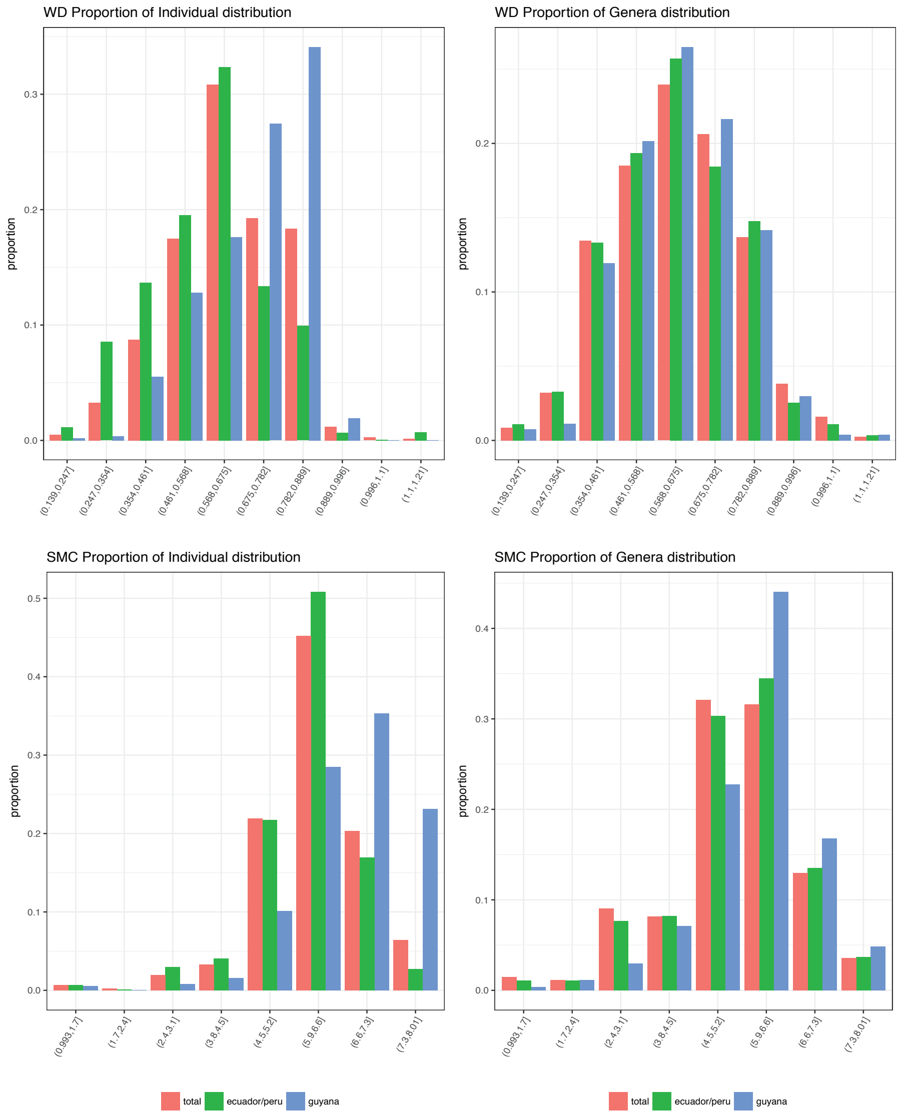


Fig. 8.4 Barplots showing distribution of functional traits per country and on either individual or genus level. For WD wood density, SMC seed mass class, distributions are shown for Guyana (blue), Ecuador/Peru (green) or the total dataset (orange). Patterns show a shift between individual and genus level distribution in relation to country as expected if there is selection for traits depending on environmental context.

neutral theory are zero-sum dynamics, meaning a filled and finite community size, a spatially implicit structure and a point mutation process as speciation. Some of these assumptions have been studied and relaxed in various attempts to reconcile the neutral and niche paradigms (32, 74, 180, 296–298). The first, however, of ecological equivalence remains a key feature of any neutral model, either spatially implicit, explicit with or without zero-sum dynamics and regardless of the speciation process. This assumption of ecological equivalence predicts that patterns in species diversity and abundance should be random. There should be no intrinsic fitness differences in relation to the environment or ecological strategy. In other words, there should be no relation between species identity in rank abundances and the environmental context. Let us focus on this assumption specifically and identify to what extend it is being violated in ecological communities, accounting for the results found in this dissertation.

Categorization of monodominance. By studying patterns of dominance in relation to forest type we found that although there were some species able to reach high dominance in different forest types, there were also species who adhered more to the classic ecological notion of being the best competitor in a specific niche, i.e. they were dominant only on a single forest type (Fig. 6.4). The first could be interpreted as potentially the result of ecological equivalence, with resistance to frequency dependent mortality (FDM) and high dominance acquired regardless of habitat. The latter, however, would argue against ecological equivalence, a collection of species not only able to cope with FDM allowing for high local abundance but also restricted in outcompeting others only within a specific habitat. If species truly could be considered ecologically equivalent, such a categorization of dominance acquisition depending on environmental context would not be expected. This indicates a first violation of neutral theory assumptions. But what could account for such a categorization of dominance?

Distribution of traits and relation with abundance. Ecological equivalence predicts there should be no relationship between specific functional traits and the relative abundance of taxa per plot or total regional abundance across the Amazon in various habitats. In addition, there also should not be any difference between the proportional distributions of how many individuals or genera have specific trait values. These traits should confer no fitness advantage or disadvantage, and according to the central limit theorem (299) we would expect a more or less normal distribution, regardless of habitat, taxonomic level or ecological strategy. However, if there were a relationship between traits and fitness we would expect to see some pattern other then a normal distribution. For instance, taxa with opposite life history strategies can also be expected to show opposite patterns in terms of traits that are correlated with higher abundances. This trade-off should also be reflected in opposite patterns of high or low local versus regional abundances. Being a fast grower and rapid disperser (pioneers) would potentially allow taxa to gain high overall regional abundance while maintaining low local abundances. In contrast, taxa specialized in defense or slow growth (climax species) would show the opposite with high local abundances and not necessarily a high regional abundance. The distribution of traits associated with either high or low local or regional abundances should then reflect this differential life history strategy, which according to neutral theory should not be there. Plotting either the overall (regional) or mean abundance per plot (local) on genus level against binned values of functional traits (Fig. 8.3) shows for some continuous traits there is indeed a normal-like distribution, such as for wood density. However, for others it resembles more a Gaussian with exact opposite distributions. Seed Mass Class for example shows higher seed masses (limiting dispersal ability) are related with high local abundances yet lower SMC are related with higher regional abundances. Specific Leaf Area shows the exact opposite: lower SLA values are associated with higher local abundances whereas higher SLA values are associated with low local but high regional abundances. Finally, leaf N and C content show similar but opposite patterns, with low N and high C associated with high local abundances and high N but low C associated with low local but high regional abundances. The ability to accumulate aluminium is also related to high local abundance, yet as this is a specific trait related to relatively rare environmental conditions it does not confer high regional abundance. For some discrete traits there also is a clear relation between having a trait and an associated higher abundance (resin and ectomycorrhiza) or a lower abundance (winged fruits) without any difference in local and regional abundance (not shown). The above clearly shows relationships between relative abundance and traits, opposing the view of ecological equivalence.

Evolutionary ecology of trait distributions. Looking at proportional distributions of these same traits but on different levels of organization (e.g. genus and individuals), neutral theory also predicts there should be no difference in distribution as there should be no selective regime. If, however, the assumption of ecological equivalence is in fact violated we might expect a shift in trait distribution depending on the direction of selection. Such a shift follows from John Endlers first expectation of evolutionary change: "*trait frequency distributions will differ among age classes or life-stages… if there is indeed selection*" (300). Within regions of similar conditions we could expect to see a shift between genus level trait distribution and individual level trait distribution. Genus level distributions represent the slower dynamics on higher taxonomic levels whereas individuals are portraying the shift that follows in the coming generations if selection is constant. If this selection pressure were dependent on certain environmental characteristics we would also expect that there is a similar shift from trait distribution of the total dataset in comparison with those from regions on different ends of environmental gradients, both on individual and genus levels. When this is done for the whole ATDN with Ecuador/Peru and the Guiana Shield representing either side of a known gradient (209) we see exactly this pattern in some of the same functional traits as mentioned above. For example, there is a clear shift in the distribution of wood density and SMC over the individuals for higher values in Guyana in comparison with both the total and Ecuador/Peru data. On genus level this shift is also visible, although less apparent (Fig. 8.4). Also within each region this shift for wood density and SMC becomes apparent when comparing distributions on individual or genus level, complying with Endlers first expectation.

The patterns discussed above in trait distribution indicate a clear signal of differential selection not only on individual but also higher levels or organization dependent on life history strategy and environmental context. Clearly the assumption of ecological equivalence is not only violated on species or individual level but also on this overarching level of life history strategy and higher taxonomic ranks. This, in combination with results from previous chapters showing disagreement between local and regional predictions of neutral theory, supports the view that neither neutral nor niche processes are solely responsible for determining the governing dynamics of biological communities and that there are clear identifiable violations of neutral theory assumptions to be found. In order to complete our understanding of community ecology, we need to quantify these signals of selection and see if we can identify not only the strength but also the direction of selection relative to stochastic influences.

8.5 Quantifying natural selection and chance. Chapters three to six suggested that inferences from neutral models should be taken with a grain of salt. However, as discussed earlier, neutral theory can still function as a proper null-model. After identifying the violation of fundamental assumptions of neutral theory as explained in the previous paragraphs, the next step was to quantitatively disentangle deterministic versus stochastic processes. As previous paragraphs have shown, this can be done indirectly by studying differences in for instance trait distribution among genera either at local or regional scales of abundance. It would, however, be preferred if this can be linked directly with variation in abundance across spatial scales. The Maximum Entropy Formalism (MEF) from information theory and statistical mechanics provides such a way when applied on ecological systems (40). In chapter seven I set out to apply this principle to the entire Amazon rainforest using forest inventory data of over 2000 hectares. The principle of maximum entropy is a mathematical approach without any a priori assumptions regarding community dynamics. In other words, it does not predict relative abundances based on iterations from mechanistic models such as neutral theory but instead mathematically derives predicted relative abundances in the form of Bayesian probabilities (Fig. 7.1 and Supporting Information chapter 7: boxes S1-S3 provide more detailed information on the MEF and the necessary calculations). Results showed that using over a dozen traits and inventory data on more than 800 genera across over 2000 hectares yielded clear quantifiable results in terms of selection versus stochastic processes. Overall, we showed there were strong correlative relationships between functional traits, regional abundances and relative abundances of taxa. However, less than 10% of the variation in composition on genus level could be explained by pure trait based filtering using the traits when corrected for model bias whereas dispersal limitation as approximated by pure metacommunity effects was as high as 30% for the entire dataset. In line with the above outlined explanations for violations of ecological equivalence, the amount of variation explained by selection was strongly dependent on environmental context. Plotting the ratio of metacommunity relative to pure trait effects for specific forest types shows a clear trend (Fig. 7.3). White sand and swamp forests on one end of the spectrum, indicating low metacommunity but high trait effects and terra firme forests on the other end showing the opposite. In other words, white sand forests appear to experience much more selection relative to the effects of migration from a regional species pool, even when data was rarefied to accommodate differences in sample size. This was also apparent when looking at the specific strengths and direction of selection with traits similar as described above strongly related to either high or low abundances dependent on forest type. For example, traits such as SMC and leaf C content showed a clear positive relation between high trait values and higher local abundances as indicated by the positive lambda values from the MEF for white sand forests (Fig. 7.5). In contrast, high N and P leaf content showed exact opposite patterns, which was expected, considering the severe nutrient limitation in white sand forests (301).

The metacommunity relative to pure trait effects ratios become even more interesting when looked at from a spatial perspective for each plot. Mapping this across the Amazon shows a clear trend with the interior of the Amazon having much higher ratios, indicating higher metacommunity effects relative to trait effects, in comparison with the edges of the Amazon (Fig. 7.4). Furthermore, when plotting the distance decay of metacommunity importance when the size of the metacommunity is increased incrementally by a radius of 50 kilometers, two important observations can be made: first that different forest types experience similar decays of metacommunity importance over distance and second that even at so much as 3500 kilometers there is still a relatively high importance of the metacommunity. Again, the spatial pattern of this distance decay of metacommunity importance per plot shows a peculiar pattern (Fig. 7.7). Taking the ratio of metacommunity importance at regional species pools of 100 km and 2500 km in diameter shows that the interior of the Amazon by far has the highest ratios, indicating the shallowest declines whereas the edges of the Amazon show the opposite pattern. This in itself could account for the higher metacommunity effects found per plot.

The interior part of the Amazon has been shown to be one of the most diverse areas of the world (115, 209) containing many hotspots of diversity. Many theories have been proposed to explain its diversity ranging from niche overlap to being refugia in dire times (302). We, however, propose a much simpler explanation. As shown by our calculations using the maximum entropy formalism, the relative importance of the regional species pool in comparison with trait based filtering is much higher for the interior then along the edges. We hypothesize this simply means that the potential source pool of species is also much larger for the interior parts of the Amazon akin to the hypothesis of the mid-domain effect (265, 266). Regardless if community dynamics lean more towards a neutral or niche perspective, this in itself would allow for a higher diversity in comparison with the edges of the Amazonian rainforest. Of course, we only have data of tropical trees of the Amazon so any input of species from for instance the Cerrado (savannah) to the South or Andes (mountain) to the West along the edges of the Amazon are not taken into account. In other words, our results do not indicate interior parts of the Amazon are more diverse per se but do contain more of the diversity found across the Amazonian rain forest. In fact, the high diversity of Western Amazonian forests would indicate much influx from the species rich Andes, whereas the less diverse Eastern Amazonian forests potentially receive input from the Cerrado, which is not as rich as the Andes. This theory of a larger potential species pool, however, needs to be studied further to provide solid support but could have far reaching implications for nature conservation and restoration. Knowing this metacommunity is the main source of diversity could ask for different strategies of conservation.

Maximum entropy and neutral theory. Our results regarding spatial patterns of metacommunity importance can also provide insight into why neutral theory performs as well as it does. It showed that the proposed panmictic metacommunity in neutral theory might actually be true in some cases (at least on genus level taxonomy) with influences of the regional species pool being as high as they are at even 3500 kilometers. It should be noted, however, that this panmictic community in neutral theory is viewed at ecological time-scales with actual dispersal coming from a hypothetical metacommunity. However, my argument would be to look at this more from an evolutionary time-scale perspective as the MEF does not estimate dispersal directly but its influence on species composition. From a genetic point of view, vicariance and limited dispersal would predict a certain amount of genetic divergence between disjunct populations. When this is much less than expected under the timescale implied for the specific situation, then long-distance dispersal could be inferred as the potential cause. From this principle, such an evolutionary metacommunity in ecology has already been supported by studying DNA sequence phylogeography (303, 304). More recently similar methods have been used to show a distinct lack of geographically based phylogenetic structure for tree genera in the Amazon, also supporting such an evolutionary metacommunity (264). Neutral theory then should perhaps not so much be regarded as an ecological theory per se, but perhaps more an evolutionary ecological theory encompassing vast spatial and temporal scales that should also be accounted for in the interpretation of results.

Conclusions

In this dissertation I have attempted to put together the proverbial puzzle of community dynamics. Explaining not only why neutral theory sometimes performs as good as it does but also determined how, and if so, how much species are cheating in life's casino. It is clear that although much still needs to be discovered regarding Amazonian tree flora we can work with what we have in terms of inventory data and move on to further our understanding. Analyses have revealed a clear violation of the assumptions of neutral theory and revealed important caveats in the foundation of the theory itself to take into account with respect to scaling predictions up or down from local to regional scales and vice versa. We show that adding biological reality to the general approach of neutral theory revealed an inability to reconcile these local and regional predictions indicating that other processes must be operating in addition to stochasticity. Linking local and regional patterns of abundance to trait distributions further identified violations of its fundamental assumptions, similar to the use of the Hardy Weinberg equation from population genetics. The world clearly is not solely neutral and long-term natural selection has changed the distribution of traits among geographically separated regions and among genera. In addition, this process of selection has had different directions and strength dependent on habitat making sole neutral dynamics even on these smaller scales unlikely. However, low pure trait based filtering in general as shown by our calculations suggest hyper-diverse communities such as the Amazonian forest may experience much overlap in niche differentiation with regard to functional traits in relation with interspecific competition, also accounting for good fits of neutral theory. But even dispersal from the regional species pool accounted for only less than half of the variation in species composition, leaving more than half of variation in composition still unexplained. Using the largest known tree inventory database and over a dozen functional traits we have identified and established the ground rules of life's casino

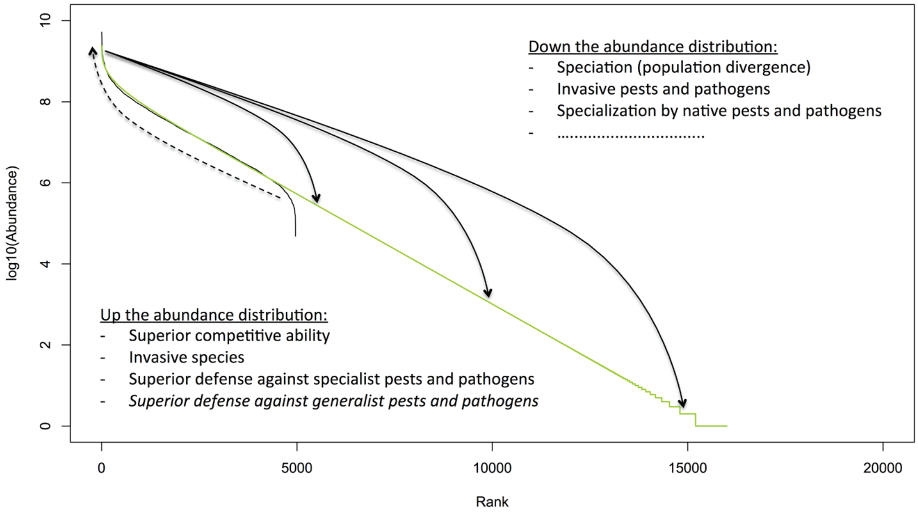


Fig. 8.5 Rank Abundance Distribution and hypothetical causes of dynamics. Distribution is shown for empirical tree data (black) of 4962 species from (14) and in green the analytical expansion of the logseries ($\Phi n = (\alpha/n)*xn$) for S = 16,000 and N = $3.9e^{11}$. Solid arrows indicate taxa losing individuals whereas dashed arrow indicates taxa gaining individuals. Loss can be 1) a consequence of speciation, where populations are divided and each separate species takes a new position in the rank abundance distribution or 2) due to severe loss of individuals due to invasive or specialized native pests or pathogens. Likewise, increasing in abundance can be due to 1) superior competitive ability allowing for slow increase, 2) rapid increase if species is invasive and lacks any species specific pests or pathogens or 3) as described in the main text defence against specialized or generalist native pests or pathogens.

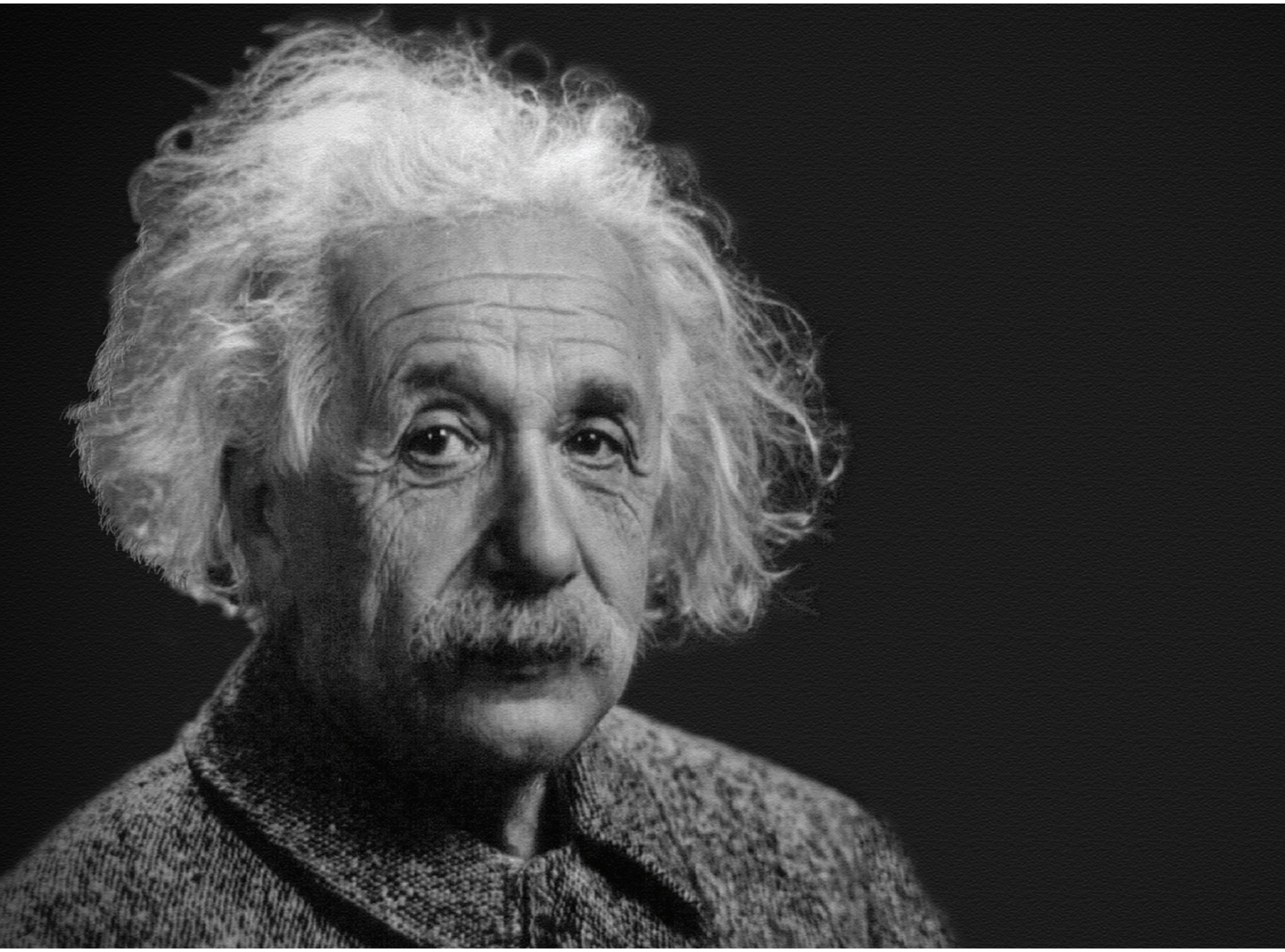
in Amazonian rainforests but there is more than meets the eye to community assembly. I believe this large part still left unexplained, even on genus level, can be due to two main reasons: either more than half of this variation in composition is determined by large scale random events such as environmental stochasticity or there is something else lurking in the shadows of community assembly. Although it is clear that such stochasticity could play a large part from time to time, I doubt whether it could explain so much of variation in community composition over such large spatial and temporal scales. There must be more to ecological communities we have not yet discovered. My work and discussions with colleagues, friends and my students have urged me to look more with an evolutionary and integrated view of ecosystem dynamics. Just as theoretical physicists have been attempting for more than a hundred years, my ambition is to find a theory of everything in evolutionary ecology. Explaining the ultimate foundations of life, diversity and dynamics in one elegant solution. In a way, I am looking for the biologists' equivalent of string theory, using various fundamental principles of life from different dimensions (or in biological terms, different levels of organization) into a single unifying concept. This dissertation is the first step towards such unification, starting by identifying the direction we should seek this solution. I believe it is time for a theory that remains simple in its foundation, one could say even neutral, but connects multiple trophic levels interacting in explaining dynamics of community structure. To remain in the analogy of life's casino, I think that species do roll the dice of life and from time to time struggle for survival using loaded dice but I also firmly believe that there are hidden players in the casino representing interactions between trophic levels that have a large impact on the structure and dynamics of communities. In a way, there are sleeping armies of pests and pathogens, co-evolving not only with each other but also with communities on other trophic levels, ever changing the rules of the game with all interactions intertwined in some manner. My final words of this thesis put forward this hypothesis, providing suggestions for future research.

8.6 The sleeping army hypothesis. As stated earlier in this dissertation, the origin and maintenance of the relative abundance distribution of taxa in communities is still shrouded in much mystery. However, regarding the dynamics of rank abundance distributions, there is ample evidence that, as Rosenzweig already stated, "*no species is safe*" (120). There are numerous possibilities for (hyper-) dominant taxa to go down the proverbial drain of the rank abundance distribution; examples such as the decimation of the North American chestnut by an invading blight (305) or the Elms by the Dutch elm disease (306). The question still remains, however, how does a taxon move up the ladder of abundance, i.e. how to become the top tree by becoming more common than others in the community. In this dissertation I have shown that the dominance of species at local scales cannot be approximated by neutral theory alone (chapter 6). How then do we explain the excessive dominance seen in the field?

Taxa that reach this high dominance must be good at some thing or the other, allowing for higher abundances. In general we could state they should be able to better defend against pests, pathogens and predators escaping effects of frequency dependent mortality as hypothesized earlier (14), or (perhaps in combination) should be able to outcompete other taxa in terms of resource competition (307). Such niche differentiation has received much attention, for instance with the broken-stick hypothesis by McArthur (308, 309), a never-ending division of niches and resources allowing for hyper diversity. However, in highly diverse systems such as Amazonian rainforests it is my belief that there has to be much niche overlap so that the latter argument becomes less likely. One reason for this is that as a taxon you simply cannot reach every site, which is especially true for trees, and a more profitable life history strategy would be to have a wide niche to be able to cope with a variety of environmental conditions. In addition, taxa in general also have to deal with the physical limitation of the adaptive potential, i.e. much but not everything is possible immediately. Second, such strong evolutionary niche differentiation would require many interactions over time with other species. As a tree, however, you mainly compete with individuals around you and interspecific competition seems less likely as you do not move about all the time and individuals have relatively long generation times. Thus for the major part an individual interacts with only a few different species and not every species in the whole community. It is for these two reasons I believe scientific efforts should be focused at better understanding the role of predator-prey interactions in shaping community structure of these diverse ecological communities across trophic levels. Pests and pathogens have already been shown to have the capabilities of significantly changing composition and dynamics of communities (230, 310, 311) and were already hypothesized to be able to structure communities and influence life history strategy (58, 312). Such dynamics between predators and their specific prey obviously could account for the downward movement along the rank abundance distribution which has also been shown both theoretically (230, 313, 314) and empirically (229, 230, 315). The more interesting question, however, is if such dynamics could also be shown to account for the excessive dominance of prey species, i.e. the upward movement along the rank abundance distribution. Most studies have focused on the influence of predator specialization and the associated negative feedback loops on abundance, but what if we look at the potentially larger population of generalists? In addition to pathogen specialists, generalist pathogens also confer some fitness disadvantage, albeit less than a specific predator (316). Perhaps then running away from these generalists before they specialize on you could confer some fitness advantage to the prey over others, accounting for higher dominance in the population. So in an extension of the Red Queen hypothesis (317), what if you could increase your fitness by running away from a sleeping army of generalists that have not yet specialized, i.e. have not woken up yet. Perhaps this could account for the excessive dominance of prey species without invoking any resource competitive abilities. Such fitness advantage, however, is most likely temporary, for no one sleeps forever and when the army does wake (i.e. starts specializing on abundant prey species) you most likely end up down in the abundance distribution again (see also Fig. 8.5).

Although it is far for complete, the inclusion of such a mechanism, in addition to the negative feedback of predators on prey abundance, could provide the necessary explanations for not only the downward movement of species but also the upward movement along the rank abundance distribution and provide testable hypotheses (318). I believe such an integrated approach can bring us a step further in understanding community dynamics as it unifies concepts from ecology and evolution in an elegant way. Red queen dynamics have already been shown to emerge from adaptive dynamic approaches (319–321). Such an approach allows linking dynamics at ecological time scales to those at evolutionary time scales and generalizes fundamental ideas from game theory to an eco-evolutionary application (322). It is my belief we can extend this principle towards explaining community dynamics as a whole using the theory described above, which I shall endeavour to accomplish in the coming years.

Although not a biologist, Einstein has meant much to me personally from a young age - reading Relativity sparked my enthusiasm of understanding the world and so he too has a place amongst the giants that this thesis is build upon.



The important thing is not to stop questioning. Curiosity has its own reason for existence. One cannot help but be in awe when he contemplates the mysteries of eternity, of life, of the marvelous structure of reality. It is enough if one tries merely to comprehend a little of this mystery each day.

Albert E. Einstein, LIFE Magazine (1955)

References

1. C. Darwin, On the origin of the species by means of natural selection or, the preservation of favoured races in the struggle for life (John Murray, London, UK, 1869).

2. E. Baldridge, D. J. Harris, X. Xiao, E. P. White, An extensive comparison of species-abundance distribution models. PeerJ. 4, e2823 (2016).

3. A. C. Quaresma, M. T. F. Piedade, Y. O. Feitosa, F. Wittmann, H. ter Steege, Composition, diversity and structure of vascular epiphytes in two contrasting Central Amazonian floodplain ecosystems. Acta Bot. Brasilica. 31, 686–697 (2017).

4. J. Alroy, The shape of terrestrial abundance distributions. Sci. Adv. 1, 1–9 (2015).

5. M. Delgado-baquerizo et al., Bacteria Found in Soil. 325, 320–325 (2018).

6. N. Fierer, R. B. Jackson, The diversity and biogeography of soil bacterial communities. Proc. Natl. Acad. Sci. U. S. A. 103, 626–31 (2006).

7. L. Tedersoo et al., Global diversity and geography of soil fungi. Science. 346, 1256688 (2014).

8. S. R. Connolly et al., Commonness and rarity in the marine biosphere. Proc. Natl. Acad. Sci. 111, 8524–8529 (2014).

9. K. J. Gaston, J. I. Spicer, Biodiversity: an introduction (John Wiley & Sons, 2013).

10. A. D. Barnosky et al., Has the Earth's sixth mass extinction already arrived? Nature. 471, 51–57 (2011).

11. E. Pennisi, What determines species diversity? Science. 309, 90 (2005).

12. D. Kennedy, C. Norman, What don't we know? Science. 309, 75 (2005).

13. W. J. Sutherland et al., Identification of 100 fundamental ecological ques tions. J. Ecol. 101, 58–67 (2013).

14. H. ter Steege et al., Hyperdominance in the Amazonian tree flora. Science. 342,1243092-1–9 (2013).

15. A. Freestone, B. Inouye, Dispersal limitation and environmental heterogeneity shape scale-dependent diversity patterns in plant communities. Ecology. 87, 2425–2432 (2006).

16. V. Volterra, Fluctuations in the abundance of a species considered mathematically. Nature. 118, 558–560 (1926).

17. A. J. Lotka, The growth of mixed populations: two species competing for a common food supply. J. Washingt. Acad. Sci. 22, 461–469 (1932).

18. R. H. MacArthur, Geographical ecology: patterns in the distribution of species (Princeton University Press, 1972).

19. R. B. McKane et al., Resource-based niches provide a basis for plant species diversity and dominance in arctic tundra. Nature. 415, 68–71 (2002).

20. J. M. Chase, M. a Leibold, Spatial scale dictates the productivity-biodiversity relationship. Nature. 416, 427–30 (2002).

21. J. M. Chase, M. A. Leibold, Ecological niches: linking classical and con temporary approaches (University of Chicago Press, 2003).

22. P. Chesson, Mechanisms of maintenance of species diversity. Annu. Rev. Ecol. Syst. 31, 343–366 (2000).

23. D. Tilman, S. W. Pacala, The maintenance of species richness in plant communities. Species Divers. Ecol. Communities (1993), pp. 13–25.

24. D. Tilman, Resource Competition between Plankton Algae: An Experimental and Theoretical Approach. Ecology. 58, 338–348 (1977).

25. D. Tilman, Resource competition and community structure (Princeton university press, 1982).

26. P. R. Grant, Ecology and evolution of Darwin's finches (Princeton University Press, 1999).

27. E. Weiher, P. A. Keddy, Assembly Rules, Null Models, and Trait Dispersion: New Questions from Old Patterns. Oikos. 74, 159 (1995).

28. P. A. Keddy, Assembly and response rules: two goals for predictive community ecology. J. Veg. Sci. 3, 157–164 (1992).

29. S. P. Hubbell, The unified neutral theory of biodiversity and biogeography. (Princeton Monographs in Population Biology. Princeton University Press, Princeton, New Jersey, USA, 2001).

30. G. Bell, Neutral macroecology. Science. 293, 2413–8 (2001).

31. Y. Malhi et al., The above-ground coarse wood productivity of 104 Neotropical forest plots. Glob. Chang. Biol. 10, 563–591 (2004).

32. D. Alonso, R. S. Etienne, A. J. McKane, The merits of neutral theory. Trends Ecol. Evol. 21, 451–7 (2006).

33. M. Holyoak, M. Loreau, Reconciling empirical ecology with neutral community models. Ecology. 87, 1370–1377 (2006).

34. B. McGill, B. Maurer, M. Weiser, Empirical evaluation of neutral theory. Ecology. 87, 1411–1423 (2006).

35. R. Condit et al., Beta-diversity in tropical forest trees. Science. 295, 666–9 (2002).

36. J. Chave, Neutral theory and community ecology. Ecol. Lett. 7, 241–253 (2004).

37. T. Bell et al., Larger islands house more bacterial taxa. Science. 308, 1884 (2005).

38. I. Volkov, J. R. Banavar, F. He, S. P. Hubbell, A. Maritan, Density dependence explains tree species abundance and diversity in tropical forests. Nature. 438, 658–61 (2005).

39. D. W. Purves, L. a Turnbull, Different but equal: the implausible assumption at the heart of neutral theory. J. Anim. Ecol. 79, 1215–25 (2010).

40. B. Shipley, From Plant Traits to Vegetation Structure. Chance and selection in the assembly of ecological communities (Cambridge University Press, 2010).

41. H. ter Steege et al., http://atdn.myspecies.info/.

42. R. C. Stauffer, Haeckel, Darwin, and ecology. Q. Rev. Biol. 32, 138–144 (1957).

43. R. Dawkins, Climbing mount improbable (WW Norton & Company, 1997).

44. J. Grinnell, The Niche-Relationships of the California Thrasher. Auk. 34, 427–433 (1917).

45. J. Vandermeer, Niche theory. Annu. Rev. Ecol. Syst. 3, 107–132 (1972).

46. C. S. Elton, Animal ecology (University of Chicago Press, 1927).

47. F. Gause, The Struggle for Existence. Yale J. Biol. Med. 7, 609 (1934).

48. G. E. Hutchinson, The niche: an abstractly inhabited hypervolume. Ecol. Theatr. Evol. Play, 26–78 (1965).

49. H. Väre, R. Ohtonen, J. Oksanen, Effects of reindeer grazing on understorey vegetation in dry Pinus sylvestris forests. J. Veg. Sci. 6, 523–530 (1995).

50. K. Soetaert, plot3D: Plotting Multi-Dimensional Data (2017), (available at https://cran.r-project.org/package=plot3D).

51. R Core Team, R: A Language and Environment for Statistical Computing (2016), (available at https://www.r-project.org/).

52. S. Wright, Isolation by distance. Genetics. 28, 114–138 (1943).

53. R. H. MacArthur, E. O. Wilson, The theory of island biogeography (Princeton university press, 1967), vol. 1.

54. J. M. Chase, Towards a really unified theory for metacommunities. Funct. Ecol. 19, 182–186 (2005).

55. G. Hutchinson, The Paradox of the Plankton. Am. Nat. 95, 137–145 (1961).

56. P. J. Grubb, The maintenance of species-richness in plant communities: the importance of the regeneration niche. Biol. Rev. 52, 107–145 (1977).

57. P. S. Ashton, Speciation among tropical forest trees: some deductions in the light of recent evidence. Biol. J. Linn. Soc. 1, 155–196 (1969).

58. D. H. Janzen, Herbivores and the number of tree species in tropical forest. Am. Nat. 104, 501–6528 (1970).

59. J. H. Connell, On the role of natural enemies in preventing competitive exclusion in some marine animals and in rain forest trees. Dyn. Popul. (1971).

60. H. Caswell, Community Structure: A neutral model Analysis. Ecol. Monogr. 46, 327–354 (1976).

61. S. P. Hubbell, Tree dispersion, abundance, and diversity in a tropical dry forest. Science. 203, 1299–309 (1979).

62. P. L. Chesson, R. R. Warner, Environmental Variability Promotes Coexistence in Lottery Competitive Systems. Am. Nat. 117, 923–943 (1981).

63. M. Kimura, G. H. Weiss, The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. Genetics. 49, 561–576 (1964).

64. M. Kimura, The neutral theory of molecular evolution. (Cambridge University Press, Cambridge, 1983).

65. J. Nekola, P. White, The distance decay of similarity in biogeography and ecology. J. Biogeogr., 867–878 (1999).

66. G. Malécot, others, The mathematics of heredity. Math. Hered. (1948).

67. T. Maruyama, A simple proof that certain quantities are independent of the geographical structure of population. Theor. Popul. Biol. 5, 148–154 (1974).

68. T. Nagylaki, The decay of genetic variability in geographically structured populations. Proc. Natl. Acad. Sci. 71, 2932–2936 (1974).

69. T. Nagylaki, The Decay of Genetic Structured in Geographically Populations II. Theor. Popul. Biol. 10, 70–82 (1976).

70. J. Chave, E. G. Leigh, A Spatially Explicit Neutral Model of β-Diversity in Tropical Forests. Theor. Popul. Biol. 62, 153–168 (2002).

71. M. Bramson, J. T. Cox, R. Durrett, A spatial Model for the Abundance of Species. Ann. Probab. 26, 658–709 (1998).

72. J. Chave, H. Muller-Landau, S. Levin, Comparing Classical Community Models : Theoretical consequences for patterns of diversity. Am. Nat. 159 (2002)

73. T. Zillio, I. Volkov, J. Banavar, S. Hubbell, A. Maritan, Spatial Scaling in Model Plant Communities. Phys. Rev. Lett. 95, 98101 (2005).

74. I. Volkov, J. R. Banavar, S. P. Hubbell, A. Maritan, Neutral theory and relative species abundance in ecology. Nature. 424, 1035–7 (2003).

75. F. He, Deriving a neutral model of species abundance from fundamental mechanism of population dynamics. Funct. Ecol. 19, 187–193 (2005).

76. E. P. Economo, T. H. Keitt, Species diversity in neutral metacommunities: A network approach. Ecol. Lett. 11, 52–62 (2008).

77. E. T. Jaynes, Information theory and statistical mechanics. Phys. Rev. 106, 620–630 (1957).

78. E. T. Jaynes, Information theory and statistical mechanics. II. Phys. Rev. 108, 171–190 (1957).

79. B. Shipley, D. Vile, E. Garnier, From Plant Traits to Plant Communities: A Statistical Mechanistic Approach to Biodiversity. Science. 314, 812–814 (2006).

80. J. Harte, Maximum entropy and ecology: a theory of abundance, distribution, and energetics (OUP Oxford, 2011).

81. B. Haegeman, M. Loreau, Limitations of entropy maximization in ecology. Oikos. 117, 1700–1710 (2008).

82. B. Shipley, Trivial and non-trivial applications of entropy maximization in ecology: Shipley's reply. Oikos. 118, 1279–1280 (2009).

83. B. Shipley, Limitations of entropy maximization in ecology: A reply to Haegeman and Loreau. Oikos. 118, 152–159 (2009).

84. J. Mallet, Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 363, 2971–86 (2008).

85. N. Pitman, J. Terborgh, M. Silman, P. N. V, Tree species distributions in an upper Amazonian forest. Ecology. 80, 2651–2661 (1999).

86. K. Ruokolainen, H. Tuomisto, J. Vormisto, N. Pitman, Two biases in estimating range sizes of Amazonian plant species. J. Trop. Ecol. 18 (2002), doi:10.1017/S0266467402002614.

87. A. Terlizzi, S. Bevilacqua, S. Fraschetti, F. Boero, Taxonomic sufficiency and the increasing insufficiency of taxonomic expertise. Mar. Pollut. Bull. 46, 556–561 (2003).

88. L. Cayuela, M. de la Cruz, K. Ruokolainen, A method to incorporate the effect of taxonomic uncertainty on multivariate analyses of ecological data. Ecography (Cop.). 34, 94–102 (2011).

89. M. A. Vanderklift, T. J. Ward, J. C. Phillips, Use of assemblages derived from different taxonomic levels to select areas for conserving marine bio diversity. Biol. Conserv. 86, 307–315 (1998).

90. A. J. Pik, I. A. N. Oliver, A. J. Beattie, Taxonomic sufficiency in ecological studies of terrestrial invertebrates. Aust. J. Ecol., 555–562 (1999).

91. A. J. Pik, J. M. Dangerfield, R. A. Bramble, C. Angus, D. A. Nipperess, The use of Invertebrates to Detect Small-scale Habitat Heterogeneity and its Application to Restoration. Environ. Monit. Assess. 75, 179–199 (2002).

92. N. Swenson, B. Enquist, Ecological and Evolutionary determinants of a key plant functional trait: wood density and its community-wide variation across latitude and elevation. Am. J. Bot. 94, 451–459 (2007).

93. M. a Higgins et al., Geological control of floristic composition in Amazonian forests. J. Biogeogr. 38, 2136–2149 (2011).

94. B. Boyle et al., The taxonomic name resolution service: an online tool for automated standardization of plant names. BMC Bioinformatics. 14, 16 (2013).

95. S. Chamberlain et al., taxize: Taxonomic information from around the web (2018), (available at https://github.com/ropensci/taxize).

96. L. Cayuela, A. Stein, J. Oksanen, Taxonstand: Taxonomic Standardization of Plant Species Names (2017), (available at https://cran.r-project.org/package=Taxonstand).

97. M. Vellend, P. L. Lilley, B. M. Starzomski, Using subsets of species in biodiversity surveys. J. Appl. Ecol. 45, 161–169 (2008).

98. R. A. Fisher, A. S. Corbet, C. B. Williams, the Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. J. Anim. Ecol. 1, 42–58 (1943).

99. J. Oksanen et al., Vegan: Community Ecology Package. R package version 2.0-7. https://CRAN.R-project.org/package=vegan (2013), (available at http://cran.r-project.org/package=vegan).

100. B. Wheeler, lmPerm: Permutation tests for linear models (2010), (available at http://cran.r-project.org/package=lmPerm).

101. L. Taylor, R. Kempton, I. Woiwod, Diversity Statistics and the Log Series Model. J. Anim. Ecol. (1976) (available at http://www.jstor.org/stable/10.2307/3778).

102. R. Kempton, The Structure of Species Abundance Measurement of Diversity. Biometrics. 35, 307–321 (1979).

103. R. Condit et al., in Forest biodiversity research, monitoring and modeling. Conceptual background and Old World case studies. (1998), pp. 247–268.

104. N. Mantel, The detection of disease clustering and a generalized regression approach. Cancer Res. 27, 209–20 (1967).

105. P. Legendre, M.-J. Fortin, Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. Mol. Ecol. Resour. 10, 831–44 (2010).

106. J. Bray, J. Curtis, An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecol. Monogr. 27, 325–349 (1957).

107. P. Jaccard, Etude comparative de la distribution florale dans une portion des Alpes et du Jura.
Bull. la Société vaudoise des Sci. Nat. 37, 547–579 (1901).

108. T. Sørensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol. Skr. 5, 1–34 (1948).

109. D. M. Raup, R. E. Crick, Measurement of faunal similarity in paleontology. J. Paleontol., 1213–1227 (1979).

110. M. Anderson, T. Crist, J. Chase, Navigating the multiple meanings of biodiversity : a roadmap for the practicing ecologist. Ecol. Lett., 19–28 (2011).

111. J. M. Chase, N. J. B. Kraft, K. G. Smith, M. Vellend, B. D. Inouye, Using null models to disentangle variation in community dissimilarity from variation in a-diversity. Ecosphere. 2, 1–11 (2011).

112. M. J. R. Fasham, A comparison of nonmetric multidimensional scaling, principal components and reciprocal averaging for the ordination of simulated coenoclines, and coenoplanes. Ecology. 58, 551–561 (1977).

113. V. K. Salako, A. Adebanji, R. Glèlè Kakaï, On the empirical performance of non-metric multidimensional scaling in vegetation studies. Int. J. Appl. Math. Stat. 36, 54–67 (2013).

114. P. R. Minchin, An evaluation of the relative robustness of techniques for ecological ordination. Vegetatio. 69, 89–107 (1987).

115. H. ter Steege et al., A spatial model of tree a -diversity and tree density for the Amazon. Biodivers. Conserv. 12, 2255–2277 (2003).

116. J. J. Lennon, P. Koleff, J. J. D. GreenwooD, K. J. Gaston, The geographical structure of British bird distributions: diversity, spatial turnover and scale. J. Anim. Ecol. 70, 966–979 (2001).

117. J. J. Lennon, P. Koleff, J. J. D. Greenwood, K. J. Gaston, Contribution of rarity and commonness to patterns of species richness. Ecol. Lett. 7, 81–87 (2004).

118. D. Mouillot et al., Rare species support vulnerable functions in high-diversity ecosystems. PLoS Biol. 11, e1001569 (2013).

119. J. Izsák, S. Pavoine, Links between the species abundance distribution and the shape of the corresponding rank abundance curve. Ecol. Indic. 14, 1–6 (2012).

120. M. L. Rosenzweig, Species diversity in space and time (Cambridge University Press, 1995).

121. N. J. Gotelli, R. K. Colwell, Quantifyinf Biodiversity: Procedures and Pitfalls in the Measurement and Comparison of Species Richness. Ecol. Lett. 4, 379–391 (2001).

122. F. W. Preston, The Commonness, And Rarity, of Species. Ecology. 29, 254–283 (1948).

123. B. J. McGill et al., Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. Ecol. Lett. 10, 995–1015 (2007).

124. U. Brose, N. D. Martinez, R. J. Williams, Estimating species richness: Sensitivity to sample coverage and insensitivity to spatial patterns. Ecology. 84, 2364–2377 (2003).

125. J.-P. Z. Wang, B. G. Lindsay, A Penalized Nonparametric Maximum Likelihood Approach to Species Richness Estimation. J. Am. Stat. Assoc. 100, 942–959 (2005).

126. H. Xu, S. Liu, Y. Li, R. Zang, F. He, Assessing non-parametric and area-based methods for estimating regional species richness. J. Veg. Sci. 23, 1006–1012 (2012).

127. J. Harte, J. Kitzes, Inferring regional-scale species diversity from small-plot censuses. PLoS One. 10, 1–12 (2015).

128. A. Chiarucci, N. J. Enright, G. L. W. Perry, B. P. Miller, B. B. Lamont, Performance species richness estimators in a high diversity plant community. 9, 283–295 (2003).

129. B. A. Walther, J. L. Moore, The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. Ecography (Cop.). 28, 815–829 (2005).

130. J. Hortal, P. A. V Borges, C. Gaspar, Evaluating the performance of species richness estimators: Sensitivity to sample grain size. J. Anim. Ecol. 75, 274–287 (2006).

131. A. Chao, Estimating Population Size for Sparse Data in Capture-Recapture Experiments. Biometrics. 45, 427 (1989).

132. A. Chao, R. K. Colwell, C. W. Lin, N. J. Gotelli, Sufficient sampling for asymptotic minimum species richness estimators. Ecology. 90, 1125–1133 (2009).

133. A. Chao, J. Bunge, Estimating the Number of Species in a Stochastic Abundance Model. Biometrics. 58, 531–539 (2002).

134. A. Chao, S.-M. Lee, Estimating the Number of Classes via Sanple Coverage. J. Am. Stat. Assoc. 87, 210–217 (1992).

135. K. P. Burnham, W. S. Overton, Robust Estimation of Population Size When Capture Probabilities Vary Among Animals. Ecology. 60, 927–936 (1979).

136. S. P. Hubbell, Estimating the global number of tropical tree species, and Fisher's paradox: Fig. 1. Proc. Natl. Acad. Sci. 112, 7343–7344 (2015).

137. S. P. Hubbell et al., How many tree species are there in the Amazon and how many of them will go extinct? Proc. Natl. Acad. Sci. 105, 11498–11504 (2008).

138. R. A. Chisholm, J. W. Lichstein, Linking dispersal, immigration and scale in the neutral theory of biodiversity. Ecol. Lett. 12, 1385–1393 (2009).

139. E. Pos et al., Estimating and interpreting migration of Amazonian forests using spatially implicit and semi-explicit neutral models. Ecol. Evol. 7, 4254–4265 (2017).

140. C. V. de Castilho, thesis (2004).

141. D. Sabatier et al., The influence of soil cover organization on the floristic and structural heterogeneity of a Guianan rain forest. Plant Ecol. 131, 81–108 (1997).

142. R. P. Salomão, Restauraçao Florestal de precisão: dinâmica e espécies estruturantes. Evolução de áreas restauradas em uma Unidade de Conservação na Amazônia - Porto Trombetas, Pará. (OmniScriptum GmbH & Co. KG, Saarbrücken, FRG., 2015).

143. J. Oksanen et al., vegan: Community Ecology Package (2013), (available at http://cran.r-project.org/package=vegan).

144. J.-P. Wang, others, SPECIES: an R package for species richness estimation. J. Stat. Softw. 40, 1–15 (2011).

145. J. Harte, T. Zillio, E. Conlisk, A. B. Smith, Maximum entropy and the state-variable approach to macroecology. Ecology. 89, 2700–2711 (2008).

146. J. Bunge et al., Estimating population diversity with CatchAll. Bioinformatics. 28, 1045–1047 (2012).

147. P. Fine, S. Kembel, Phylogenetic community structure and phylogenetic turnover across space and edaphic gradients in western Amazonian tree communities. Ecography (Cop.). (2011) doi:10.1111/j.1600-0587.2010.06548.x.

148. I. Rocchetti, J. Bunge, D. Böhning, Population size estimation based upon ratios of recapture probabilities. Ann. Appl. Stat. 5, 1512–1533 (2011).

149. H. Ter Steege et al., The discovery of the Amazonian tree flora with an updated checklist of all known tree taxa. Sci. Rep. 6, 1–15 (2016).

150. J. B. Plotkin et al., Predicting species diversity in tropical forests. P Natl Acad Sci Usa. 97, 10850–10854 (2000).

151. R. Krishnamani, A. Kumar, J. Harte, Estimating species richness at large spatial scales using data from small discrete plots. Ecography (Cop.). 27, 637–642 (2004).

152. S. P. Hubbell, R. B. Foster, Diversity of canopy trees in a neotropical forest and implications for conservation (Blackwell Scientific Publications, 1983), vol. 2.

153. A. Chiarucci, Estimating species richness: Still a long way off! J. Veg. Sci. 23, 1003–1005 (2012).

154. J. Riberiro et al., Flora da Reserva Ducke: Guia de Identificação das Plantas Vasculares de uma Floresta de Terra Firme na Amazônica Central (1999).

155. J. Bunge, K. Barger, Parametric models for estimating the number of classes. Biometrical J. 50, 971–982 (2008).

156. A. E. Magurran, P. a Henderson, Explaining the excess of rare species in natural species abundance distributions. Nature. 422, 714–716 (2003).

157. A. Zizka, H. ter Steege, M. do C. R. Pessoa, A. Antonelli, Finding needles in the haystack: where to look for rare species in the American tropics. Ecography (Cop.). 41, 321–330 (2018).

158. D. L. DeAngelis, J. C. Waterhouse, Equilibrium and Nonequilibrium Ecological Concepts in Ecological Models. Ecol. Monogr. 57, 1–21 (1987).

159. E. G. Leigh, Neutral theory: a historical perspective. J. Evol. Biol. 20, 2075–91 (2007).

160. J. S. Clark, Beyond neutral science. Trends Ecol. Evol. 24, 8–15 (2009).

161. P. B. Adler, J. HilleRislambers, J. M. Levine, A niche for neutrality. Ecol. Lett. 10, 95–104 (2007).

162. G. Hardin, The competitive exclusion principle. Science. 131, 1292–1297 (1960).

163. J. Grinnell, Geography and Evolution. Ecology. 5, 225–229 (1924).

164. B. Patten, G. Auble, System theory of the ecological niche. Am. Nat. 117, 893–922 (1981).

165. G. Hutchinson, Homage to Santa Rosalia or why are there so many kinds of animals? Am. Nat. 93, 145–159 (1959).

166. J. Terborgh, R. B. Foster, P. N. V, Tropical Tree Communities: A Test of the Nonequilibrium Hypothesis. Ecology. 77, 561–567 (1996).

167. N. C. A. Pitman et al., Dominance and Distribution of Tree Species in Upper Amazonian Terra Firme Forests. Ecology. 82, 2101–2117 (2001).

168. N. Pitman, J. Terborgh, M. Silman, A comparison of tree species diversity in two upper Amazonian forests. Ecology. 83, 3210–3224 (2002).

169. J. F. Duivenvoorden, J. C. Svenning, S. J. Wright, beta diversity in tropical forests. Science. 295, 636–637 (2002).

170. H. Tuomisto, K. Ruokolainen, M. Yli-Halla, Dispersal, environment, and floristic variation of western Amazonian forests. Science. 299, 241–4 (2003).

171. F. Valladares, S. Wright, E. Lasso, Plastic Phenotypic Response to Light of 16 Congeneric Shrubs from a Panamanian Rainforest. Ecology. 81, 1925–1936 (2000).

172. G. Bell, The Distribution of Abundance in Neutral Communities. Am. Nat. 155, 606–617 (2000).

173. M. A. M. de Aguiar, M. Baranger, E. M. Baptestini, L. Kaufman, Y. Bar-Yam, Global patterns of speciation and diversity. Nature. 460, 384–7 (2009).

174. S. Barot, J. Gignoux, Mechanisms promoting plant coexistence: Can all the proposed processes be reconciled? Oikos. 106, 185–192 (2004).

175. D. Gravel, C. D. Canham, M. Beaudet, C. Messier, Reconciling niche and neutrality: The continuum hypothesis. Ecol. Lett. 9, 399–409 (2006).

176. B. J. McGill, Towards a unification of unified theories of biodiversity. Ecol. Lett. 13, 627–642 (2010).

177. B. J. McGill, J. C. Nekola, Mechanisms in macroecology: AWOL or purloined letter? Towards a pragmatic view of mechanism. Oikos. 119, 591–603 (2010).

178. C. Beeravolu, P. Couteron, R. Pélissier, F. Munoz, Studying ecological communities from a neutral standpoint: A review of models' structure and parameter estimation. Ecol. Modell. 220, 2603–2610 (2009).

179. S. Horvát, a Derzsi, Z. Néda, a Balog, A spatially explicit model for tropical tree diversity patterns. J. Theor. Biol. 265, 517–23 (2010).

180. J. P. O'Dwyer, J. L. Green, Field theory for biogeography: A spatially explicit model for predicting patterns of biodiversity. Ecol. Lett. 13, 87–95 (2010).

181. R. Durrett, S. Levin, Spatial Models for Species-Area Curves. J. Theor. Biol. 179, 119–127 (1996).

182. R. S. Etienne, J. Rosindell, The spatial limitations of current neutral models of biodiversity. PLoS One. 6, e14717 (2011).

183. J. Rosindell, S. P. Hubbell, R. S. Etienne, The unified neutral theory of biodiversity and biogeography at age ten. Trends Ecol. Evol. 26, 340–8 (2011).

184. R. S. Etienne, A new sampling formula for neutral biodiversity. Ecol. Lett. 8, 253–260 (2005).

185. F. Jabot, R. S. Etienne, J. Chave, Reconciling neutral community models and environmental filtering: theory and an empirical test. Oikos. 117, 1308–1320 (2008).

186. F. Munoz, P. Couteron, B. R. Ramesh, Beta diversity in spatially implicit neutral models: a new way to assess species migration. Am. Nat. 172, 116–27 (2008).

187. R. Etienne, Improved estimation of neutral model parameters for multiple samples with different degrees of dispersal limitation. Ecology. 90, 847–852 (2009).

188. F. Munoz, P. Couteron, B. Ramesh, R. Etienne, Estimating parameters of neutral communities: from one single to several small samples. Ecology. 88, 2482–2488 (2007).

189. Y. Malhi et al., The regional variation of aboveground live biomass in old-growth Amazonian forests. Glob. Chang. Biol. 12, 1107–1138 (2006).

190. PARI/GP version 2.4.3 (2008).

191. MATLAB, version 7.0.0 (R14) (The MathWorks Inc., Natick, Massachusetts, 2004).

192. R. K. S. Hankin, Introducing untb, an R package for simulating ecological drift under the unified neutral theory of biodiversity. J. Stat. Softw. 22, 1–15 (2007).

193. R. Koenker, Quantreg: quantile regression. R package version 5.24. https://CRAN.R-project.org/package=quantreg (2013), (available at http://cran.r-project.org/package=quantreg).

194. D. W. Roberts, labdsv: Ordination and Multivariate Analysis for Ecology. R package version 1.8-0. https://CRAN.R-project.org/package=labdsv (2013), (available at http://cran.r-project.org/package=labdsv).

195. D. Wuertz, fAsianOptions: EBM and Asian Option Valuation. R package version 3010.79. https://CRAN.R-project.org/package=fAsianOptions. (2013), (available at http://cran.r-project.org/package=fAsianOptions).

196. H. ter Steege et al., Estimating species richness in hyper-diverse large tree communities. Ecology. 98 (2017), doi:10.1002/ecy.1813.

197. R. Nathan, H. C. M. Muller-Landau, Spatial patterns of seed dispersal, their determinants and consequences for recruitment. Trends Ecol. Evol. 15, 278–285 (2000).

198. G. Bohrer, G. G. Katul, R. Nathan, R. L. Walko, R. Avissar, Effects of canopy heterogeneity, seed abscission and inertia on wind-driven dispersal kernels of tree seeds. J. Ecol. 96, 569–580 (2008).

199. K. D. Maurer, G. Bohrer, D. Medvigy, S. J. Wright, The timing of abscission affects dispersal distance in a wind-dispersed tropical tree. Funct. Ecol. 27, 208–218 (2013).

200. Pos, E., J. E. Guevara Andino, D. Sabatier, J.-F. Molino, N. Pitman, H. Mogollón, D. Neill, C. Cerón, G. Rivas, A. Di Fiore, R. Thomas, M. Tirado, K. R. Young, O. Wang, R. Sierra, R. García-Villacorta, R. Zagt, W. Palacios, M. Aulestia and H. ter Steege. Are all species necessary to reveal ecologically important patterns? Ecol. Evol. 4, 4626–4636 (2014).

201. F. Jabot, J. Chave, Inferring the parameters of the neutral theory of biodiversity using phylogenetic information and implications for tropical forests. Ecol. Lett. 12, 239–248 (2008).

202. R. S. Etienne, Maximum likelihood estimation of neutral model parameters for multiple samples with different degrees of dispersal limitation. J. Theor. Biol. 257, 510–4 (2009).

203. M. D. Swaine, T. C. Whitmore, On the definition of ecological species groups in tropical rain forests. Vegetatio. 75, 81–86 (1988).

204. M. Westoby, A leaf-height-seed (LHS) plant ecology strategy scheme. Plant Soil. 199, 213–227 (1998).

205. H. Gitay, I. R. Noble, J. H. Connell, Deriving functional types for rain-forest trees. J. Veg. Sci. 10, 641–650 (1999).

206. S. Mota de Oliveira, H. ter Steege, J. H. C. Cornelissen, S. Robbert Gradstein, Niche assembly of epiphytic bryophyte communities in the Guianas: a regional approach. J. Biogeogr. 36, 2076–2084 (2009).

207. D. S. Hammond, V. K. Brown, Seed Size of Woody Plants in Relation to Disturbance , Dispersal , Soil Type in Wet Neotropical Forests. Ecology. 76, 2544–2561 (2012).

208. H. ter Steege, D. Hammond, Character convergence, diversity, and disturbance in tropical rain forest in Guyana. Ecology. 82, 3197–3212 (2001).

209. H. ter Steege et al., Continental-scale patterns of canopy tree composition and function across Amazonia. Nature. 443, 444–7 (2006).

210. T. G. Seidler, J. B. Plotkin, Seed dispersal and spatial pattern in tropical trees. PLoS Biol. 4, e344 (2006).

211. O. L. Phillips, P. Hall, a H. Gentry, S. a Sawyer, R. Vásquez, Dynamics and species richness of tropical rain forests.
Proc. Natl. Acad. Sci. U. S. A. 91, 2805–9 (1994).

212. O. L. Phillips et al., Pattern and process in Amazon tree turnover, 1976-2001. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 359, 381–407 (2004).

213. J. M. Marzluff, K. P. Dial, Life-history correlates of taxonomic diversity. Ecology. 72, 428–439 (1991).

214. E. T. Pos et al., Are all species necessary to reveal ecologically important patterns? Ecol. Evol. 4 (2014), doi:10.1002/ece3.1246.

215. E. T. Pos et al., Adding biological reality to general predictions of neutral theory reveals scaling issues between local and regional patterns of diversity - *currently under review.*

216. R. Etienne, D. Alonso, A dispersal-limited sampling theory for species and alleles. Ecol. Lett., 1147–1156 (2005).

217. R. Etienne, H. Olff, A novel genealogical approach to neutral biodiversity theory. Ecol. Lett., 170–175 (2004).

218. W. Ewens, The Sampling Theory of Selectively neutral alleles.
Theor. Popul. Biol. 112, 87–112 (1972).

219. M. Nei, F-statistics and analysis of gene diversity in subdivided populations. Ann. Hum. Genet. 41, 225–233 (1977).

220. N. Takahata, M. Nei, FST and GST statistics in the finite island model. Genetics. 11, 349–352 (1984).

221. E. Simpson, Measurement of diversity. Nature. 163, 1949 (1949).

222. G. H. Hardy, Mendelian Proportions in a Mixed Population.
Science. 28, 49–50 (1908).

223. W. Weinberg, Über den Nachweis der Vererbung beim Menschen. Jahresh Wuertt Ver vaterl Natkd. 64, 368–382 (1908).

224. T. Zillio, R. Condit, The impact of neutrality, niche differentiation and species input on diversity and abundance distributions. Oikos. 116, 931–940 (2007).

225. J. Rosindell, S. Cornell, Species–area curves, neutral models, and long-distance dispersal. Ecology. 90, 1743–1750 (2009).

226. H. ter Steege et al., Estimating species richness in hyper-diverse large tree communities. Ecology. 98, 1444–1454 (2017).

227. N. J. B. Kraft, R. Valencia, D. D. Ackerly, Functional traits and niche-based tree community assembly in an Amazonian forest. Science. 322, 580–2 (2008).

228. T. Dobzhansky, N. P. Spassky, Genetic Drift and Natural Selection in Experimental Populations of Drosophila Pseudoobscura. Proc. Natl. Acad. Sci. U. S. A. 48, 148–156 (1962).

229. L. S. Comita, H. C. Muller-Landau, S. Aguilar, S. P. Hubbell, Asymmetric density dependence shapes species abundances in a tropical tree community. Science. 329, 330–332 (2010).

230. S. A. Mangan et al., Negative plant-soil feedback predicts tree-species relative abundance in a tropical forest. Nature. 466, 752–5 (2010).

231. S. D. Torti, P. D. Coley, T. A. Kursar, Causes and Consequences of Monodominance in Tropical Lowland Forests. Am. Nat. 157, 141–153 (2001).

232. S. P. Ribeiro, V. K. Brown, Prevalence of monodominant vigorous tree populations in the tropics: Herbivory pressure on Tabebuia species in very different habitats. J. Ecol. 94, 932–941 (2006).

233. K. S. H. Peh, S. L. Lewis, J. Lloyd, Mechanisms of monodominance in diverse tropical tree-dominated systems. J. Ecol. 99, 891–898 (2011).

234. R. E. Ricklefs, S. S. Renner, Global correlations in tropical tree species richness and abundance reject neutrality. Science. 335, 464–467 (2012).

235. R. E. Ricklefs, S. S. Renner, Response to comments on "Global correlations in tropical tree species richness and abundance reject neutrality." Science. 336, 1639 (2012).

236. E. Pos et al., Estimating and interpreting migration of Amazonian forests using spatially implicit and semi-explicit neutral models. Ecol. Evol. 7 (2017), doi:10.1002/ece3.2930.

237. R. Analytics, S. Weston, doParallel: Foreach Parallel Adaptor for the "parallel" Package (2015), (available at https://cran.r-project.org/package=doParallel).

238. S. Weston, doSNOW: Foreach Parallel Adaptor for the "snow" Package (2015), (available at https://cran.r-project.org/package=doSNOW).

239. R. a Chisholm, S. W. Pacala, Niche and neutral models predict asymptotically equivalent species abundance distributions in high-diversity ecological communities. Proc. Natl. Acad. Sci. U. S. A. 107, 15821–5 (2010).

240. T. Yumoto, Seed Dispersal by Salvin ' s Curassow , Mitu salvini (Cracidae), in a Tropical Forest of Colombia : Direct Measurements of Dispersal. Biotropica. 31, 654–660 (1999).

241. H. C. Muller-Landau, S. J. Wright, O. Calderón, R. Condit, S. P. Hubbell, Interspecific variation in primary seed dispersal in a tropical forest. J. Ecol. 96, 653–667 (2008).

242. F. J. Massey, The Kolmogorov-Smirnov Test for Goodness of Fit. J. Am. Stat. Assoc. 46, 68–78 (1959).

243. F. Wilcoxon, Individual Comparisons by Ranking Methods. Biometrics Bull. 1, 80 (1945).

244. M. Morisita, Measuring of Interspecific Association and Similarity Between Communities. Mem. Fac. Sci. Kyushu Univ. Ser. E. 3 (1959).

245. H. ter Steege et al., SOM Hyperdominance in the Amazonian tree flora. Science. 342, 1243092 (2013).

246. B. J. Mcgill, A test of the unified neutral theory of biodiversity. Nature. 422, 881–885 (2003).

247. J. Rosindell, Y. Wong, R. S. Etienne, A coalescence approach to spatial neutral ecology. Ecol. Inform. 3, 259–271 (2008).

248. J. Rosindell, S. J. Cornell, Universal scaling of species-abundance distributions across multiple scales. Oikos. 122, 1101–1111 (2012).

249. D. Tilman, Niche tradeoffs, neutrality, and community structure: a stochastic theory of resource competition, invasion, and community assembly. Proc. Natl. Acad. Sci. U. S. A. 101, 10854–61 (2004).

250. R. D. Holt, Emergent neutrality. Trends Ecol. Evol. 21, 531–3 (2006).

251. B. Shipley, C. E. T. Paine, C. Baraloto, Quantifying the importance of local niche-based and stochastic processes to tropical tree community assembly. Ecology. 93, 760–769 (2012).

252. B. Shipley, D. C. Laughlin, G. Sonnier, R. Otfinowski, A strong test of a maximum entropy model of trait-based community assembly. Ecology. 92, 507–517 (2011).

253. A. A. Agrawal, Macroevolution of plant defense strategies. Trends Ecol. Evol. 22, 103–109 (2007).

254. J. C. Ordoñez et al., A global study of relationships between leaf traits, climate and soil measures of nutrient fertility. Glob. Ecol. Biogeogr. 18, 137–149 (2009).

255. L. H. M. Cosme, J. Schietti, F. R. C. Costa, R. S. Oliveira, The importance of hydraulic architecture to the distribution patterns of trees in a central Amazonian forest. New Phytol. 215, 113–125 (2017).

256. P. B. Adler et al., Functional traits explain variation in plant life history strategies. Proc. Natl. Acad. Sci. 111, 10019–10019 (2014).

257. G. Kunstler et al., Plant functional traits have globally consistent effects on competition. Nature. 529, 204–207 (2016).

258. C. O. Marks, M. J. Lechowicz, Alternative Designs and the Evolution of Functional Diversity. Am. Nat. 167, 55–66 (2006).

259. L. Poorter, C. V. Castilho, J. Schietti, R. S. Oliveira, F. R. C. Costa, Can traits predict individual growth performance? A test in a hyperdiverse tropical forest. New Phytol. 219, 109–121 (2018).

260. J. Yang, M. Cao, N. G. Swenson, Why Functional Traits Do Not Predict Tree Demographic Rates. Trends Ecol. Evol. 33, 326–336 (2018).

261. S. E. Russo, S. J. Davies, D. A. King, S. Tan, Soil-related performance variation and distributions of tree species in a Bornean rain forest. J. Ecol. 93, 879–889 (2005).

262. S. J. Davies, P. A. Palmiotto, P. S. Ashton, H. S. Lee, J. V. Lafrankie, Comparative ecology of 11 sympatric species of Macaranga in Borneo: Tree distribution in relation to horizontal and vertical resource heterogeneity. J. Ecol. 86, 662–673 (1998).

263. R. Nathan, Long-distance dispersal of plants. Science. 313, 786–8 (2006).

264. K. G. Dexter et al., Dispersal assembly of rain forest tree communities across the Amazon basin. Proc. Natl. Acad. Sci. 114, 2645–2650 (2017).

265. R. K. Colwell, C. Rahbek, N. J. Gotelli, The mid-domain effect and species richness patterns:what have we learned so far? Am. Nat. 163, E1-23 (2004).

266. R. K. Colwell, D. C. Lees, The mid-domain effect : geometric constraints on the geography of species richness. 15, 70–76 (2000).

267. T. F. Rangel et al., Modeling the ecology and evolution of biodiversity: Biogeographical cradles, museums, and graves. Science. Science. in press (2018).

268. Tropicos Missouri Botanical Garden. www.tropicos.org.

269. S. van Buuren, K. Groothuis-Oudshoorn, Mice: Multivariate Imputation by Chained Equations in R. J. Stat. Softw. 45, 1–67 (2011).

270. E. Laliberté, B. Shipley, FD: measuring functional diversity from multiple traits, and other tools for functional ecology. R package version 1.0-11. URL http//CRAN. R-project. org/package= FD (2011).

271. B. Shipley, Inferential permutation tests for maximum entropy models in ecology. Ecology. 91, 2794–2805 (2010).

272. C. Baraloto et al., Decoupled leaf and stem economics in rain forest trees. Ecol. Lett. 13, 1338–1347 (2010).

273. C. Fortunel, P. V. A. Fine, C. Baraloto, Leaf, stem and root tissue strategies across 758 Neotropical tree species. Funct. Ecol. 26, 1153–1161 (2012).

274. N. Fyllas, S. Patino, Basin-wide variations in foliar properties of Amazonian forest: phylogeny, soils and climate. Biogeosciences, 2677–2708 (2009).

275. F. M. J. Ohler, Phytomass and mineral content in untouched forest (1980).

276. C. E. T. Paine, C. Baraloto, J. Chave, B. Hérault, Functional traits of individual trees reveal ecological constraints on community assembly in tropical rain forests. Oikos. 120, 720–727 (2011).

277. T. L. Pons, K. Perreijn, C. Van Kessel, M. J. A. Werger, Symbiotic nitrogen fixation in a tropical rainforest: 15N natural abundance measurements supported by experimental isotopic enrichment. New Phytol. 173, 154–167 (2007).

278. J. Thompson et al., Ecological studies on a lowland evergreen rain forest on Maraca Island, Roraima, Brazil. I. Physical environment, forest structure and leaf chemistry. J. Ecol., 689–703 (1992).

279. J. Kattge et al., TRY - a global database of plant traits. Glob. Chang. Biol. 17, 2905–2935 (2011).

280. M. Leibold, M. McPeek, Coexistence of the niche and neutral perspectives in community ecology. Ecology. 87, 1399–1410 (2006).

281. M. A. Leibold, Return of the niche. 454, 3–4 (2008).

282. R. May, Unanswered questions in ecology.
Philos. Trans. R. Soc. B Biol. Sci. 354, 1951–1959 (1999).

283. A. A. Agrawal et al., Filling key gaps in population and community ecology. Front. Ecol. Environ. 5, 145–152 (2007).

284. J. L. Reid et al., Foundations of Restoration Ecology.
Restor. Ecol. 25, 844–845 (2017).

285. J. B. Socolar, J. J. Gilroy, W. E. Kunin, D. P. Edwards, How Should Beta-Diversity Inform Biodiversity Conservation?
Trends Ecol. Evol. 31, 67–80 (2016).

286. R. A. Betts, Y. Malhi, J. T. Roberts, The future of the Amazon: new perspectives from climate, ecosystem and social sciences.
Philos. Trans. R. Soc. Lond. B. Biol. Sci. 363, 1729–35 (2008).

287. O. E. Sala et al., Global Biodiversity Scenarios for the Year 2100 Global Biodiversity Scenarios for the Year 2100. Science. 287, 1770–1774 (2000).

288. J. Rosindell, S. P. Hubbell, F. He, L. J. Harmon, R. S. Etienne, The case for ecological neutral theory. Trends Ecol. Evol. 27, 203–8 (2012).

289. S. Wright, Evolution in mendelian populations.
Genetics. 16, 97–159 (1931).

290. H. ter Steege, R. Zagt, Ecology: density and diversity.
Nature. 417, 698–699 (2002).

291. J. M. Chase, Spatial scale resolves the niche versus neutral theory debate.
J. Veg. Sci. 25, 319–322 (2014).

292. M. Lohbeck et al., Changing drivers of species dominance during tropical forest succession. Funct. Ecol. 28, 1052–1058 (2014).

293. R. D. Holt, Species Coexistence. Encycl. Biodivers. 5, 667–678 (2013).

294. R. D. Holt, J. Grover, D. Tilman, Simple Rules for Interspecific Dominance in Systems with Exploitative and Apparent Competition. Am. Nat. 144, 741–771 (1994).

295. R. S. Waples, F. Allendorf, Testing for hardy-weinberg proportions: Have we lost the plot? J. Hered. 106, 1–19 (2015).

296. R. S. Etienne, D. Alonso, A. J. McKane, The zero-sum assumption in neutral biodiversity theory. J. Theor. Biol. 248, 522–36 (2007).

297. J. Rosindell, S. J. Cornell, S. P. Hubbell, R. S. Etienne, Protracted speciation revitalizes the neutral theory of biodiversity. Ecol. Lett. 13, 716–27 (2010).

298. T. J. Matthews, R. J. Whittaker, Neutral theory and the species abundance distribution: Recent developments and prospects for unifying niche and neutral perspectives. Ecol. Evol. 4, 2263–2277 (2014).

299. H. Fischer, A history of the central limit theorem: From classical to modern probability theory. (Springer Science & Business Media, 2010).

300. J. A. Endler, Natural selection in the wild (Princeton University Press, 1986).

301. C. a. Quesada et al., Soils of Amazonia with particular reference to the RAINFOR sites. Biogeosciences. 8, 1415–1440 (2011).

302. A. A. De Oliveira, S. A. Mori, A central Amazonian terra firme forests. I. High tree species richness on poor soils. Biodivers. Conserv. 8, 1219–1244 (1999).

303. C. W. Dick, E. Bermingham, M. R. Lemes, R. Gribel, Extreme long-distance dispersal of the lowland tropical rainforest tree *Ceiba pentandra L.* (Malvaceae) in Africa and the Neotropics. Mol. Ecol. 16, 3039–3049 (2007).

304. R. T. Pennington, C. W. Dick, The role of immigrants in the assembly of the South American rainforest tree flora. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 359, 1611–22 (2004).

305. S. L. Anagnostakis, Chestnut Blight: The Classical Problem of an Introduced Pathogen. Mycologia. 79, 23 (1987).

306. G. N. Lanier, D. C. Schubert, P. D. Manion, Dutch elm disease and elm yellows in Central New York. Plant Dis. 3) (1988).

307. J. H. Connell, M. D. Lowman, Low-Diversity Tropical Rain Forests: Some Possible Mechanisms for Their Existence. Am. Nat. 134, 88–119 (1989).

308. R. MacArthur, On the relative abundance of species. Am. Nat. 94, 25–36 (1960).

309. R. H. MacArthur, On the relative abundance of bird species. Proc. Natl. Acad. Sci. U. S. A. 43, 293–295 (1957).

310. H. M. Alexander, R. D. Holt, The interaction between plant competition and disease. Perspect. Plant Ecol. Evol. Syst. 1, 206–220 (1998).

311. G. S. Gilbert, Evolutionary ecology of plant diseases in natural ecosystems. Annu. Rev. Phytopathol. 40, 13–43 (2002).

312. J. H. Connel, Diversity in tropical rain forests and coral reefs. Science. 199, 1302–1310 (1978).

313. R. A. Chisholm, H. C. Muller-Landau, A theoretical model linking interspecific variation in density dependence to species abundances. Theor. Ecol. 4, 241–253 (2011).

314. K. M. L. Mack, J. D. Bever, Coexistence and relative abundance in plant communities are determined by feedbacks when the scale of feedback and dispersal is local. J. Ecol. 102, 1195–1201 (2014).

315. J. N. Klironomos, Feedback with soil biota contributes to plant rarity and invasiveness in communities. Nature. 417, 67–70 (2002).

316. J. D. Bever, S. A. Mangan, H. M. Alexander, Maintenance of Plant Species Diversity by Pathogens. Annu. Rev. Ecol. Evol. Syst. 46, 305–325 (2015).

317. L. Van Valen, A new evolutionary law. Evol. theory. 1, 1–30 (1973).

318. G. G. Mcnickle, R. Dybzinski, Game theory and plant ecology. Ecol. Lett. 16, 545–555 (2013).

319. É. Kisdi, F. J. a Jacobs, S. a H. Geritz, Red Queen evolution by cycles of evolutionary branching and extinction. Selection. 2, 161–176 (2001).

320. P. Marrow, U. Dieckmann, R. Law, Evolutionary dynamics of predator-prey systems: an ecological perspective. J. Math. Biol. 34, 556–578 (1996).

321. F. Dercole, R. Ferrie, A. Gragnani, S. Rinaldi, Coevolution of slow-fast populations: Evolutionary sliding, evolutionary pseudo-equilibria and complex Red Queen dynamics. Proc. R. Soc. B Biol. Sci. 273, 983–990 (2006).

322. Å. Brännström, J. Johansson, N. von Festenberg, The Hitchhiker's Guide to Adaptive Dynamics. Games. 4, 304–328 (2013).

323. B. Shipley, Measuring and interpreting trait-based selection versus meta-community effects during local community assembly. J. Veg. Sci. 25, 55–65 (2014).

324. A. Colin Cameron, F. A. G. Windmeijer, An R-squared measure of goodness of fit for some common nonlinear regression models. J. Econom. 77, 329–342 (1997).

The forests surrounding Masakanari, a small village in the South of Guyana - near the Cuyuwini river.

Summary

At every scale of life, from microscopic bacteria to the towering giants of Amazonian trees we see a similar pattern of diversity where not everybody gets to be equally abundant. Instead, there is a pattern in which common species are rare and rare species are common. But how does this seemingly universal pattern of diversity come to be and how is it is maintained? What mechanisms account for the rarity of some and the commonness of others? Its answer frames our fundamental understanding of life and the driving forces of diversity, in the past, present and future. This subject has been tried and tested for generations of scientists, generating a vast body of both theoretical and empirical attempts. The various approaches in solving this conundrum can, however, be broadly categorized in two different perspectives: either processes driven by generations of natural selection that have shaped and altered taxa to outcompete others in specific areas or stochastic events, a never ending game of chance that alters community structure according to fixed laws of probability. In a sense, one might say that any organism is either struggling for life or rolling the dice in life's casino. Quantifying the relative importance of these perspectives has been the main theme of this dissertation.

In **chapters 1 and 2** we start with an introduction into the debate described above, providing readers with the necessary background in the theoretical aspects of evolutionary ecology. We explain that classic niche theory states that species are adapted to certain environmental characteristics and composition is dictated primarily by environmental filtering. In contrast, neutral theory proposes dynamics of competition have neutral outcomes and composition is mainly governed by stochastic demographic events, dispersal and migration. In these first two chapters we present a short overview of both niche and neutral based theories, providing a short historic and theoretical synthesis of both perspectives. Regardless of which of these two perspectives is tested, however, empirical data on diversity is used to test ideas and theories. Data that is meticulously collected and identified. Despite ample efforts, ecologists are often unable to identify all collections, forcing them to either omit these unidentified records entirely, without knowing the effect of this, or pursue very costly and time-consuming efforts of identifying them.

In **chapter three** we therefor investigated the consequence of omitting such unidentified records and provide an explanation for the results. Our results showed a high congruence for all analyses when using only the identified rather than all morphospecies, suggesting that patterns of similarity and composition are very robust. In other words, having a large number of unidentified species in a dataset may not affect our conclusions as much as is often thought, paving the way to make conscious decisions of whether or not to use unidentified records in ecological analyses regarding species composition.

For species diversity and richness, however, it did matter. This brings us to the important aspect of accurately being able to estimate species diversity. For many ecological questions, one of the first difficulties is how many species there presumably are in the total area given any inventory dataset representing a subset of this area. Hence, species richness estimation is one of the most widely used analyses carried out by ecologists, and nonparametric estimators are probably the most used techniques to carry out such estimations. In **chapter four**, we tested the assumptions and results of nonparametric estimators and those of a logseries approach to species richness estimation for simulated tropical forests and five datasets from the field. We concluded that nonparametric estimators are not suitable to estimate species richness in tropical forests, where sampling intensity is usually low and richness is high, because the assumptions of the methods do not meet the sampling strategy used in most studies. The logseries, while also requiring substantial sampling, is much more effective in estimating species richness than commonly used nonparametric estimators, and its assumptions better match the way field data is being collected.

The results of chapters three and four paved the way for developing our theoretical models in which we tested the importance of migration and stochastic events on structuring community composition. The final necessary step before we could actually test our models was to accurately estimate one of the core parameters of neutral models: migration. With many sophisticated methods available for estimating migration, ecologists face the difficult decision of choosing for their specific line of work. In **chapter five** we tested and compared several methods, performing robustness tests, applying these methods to large-scale inventory data. We selected five different methods and compared their ability to estimate migration from a spatially implicit and semi-explicit model. In the former, there is no spatial relationship between local communities and the regional species community while in the latter there is, creating a dependency between the distance and the probability of migration. This spatial relationship is important as it mimics what we see in real life - not everything is able to get everywhere, as many organisms have limited dispersal. Most methods were able to accurately estimate migration from spatially implicit simulations. For spatially semi-explicit simulations, estimation was shown to be the additive effect of migration from adjacent plots and the regional (summed) pool of species, i.e. the metacommunity. Estimation was only accurate when migration from the metacommunity outweighed that of adjacent plots. We concluded that estimated migration is more an approximation of the homogenization among local communities over time rather than a direct measurement of migration and hence has a direct relationship with beta-diversity. But as beta-diversity is the result of many (non)-neutral processes, we also concluded that migration, as estimated from a spatial explicit world, encompasses not only direct migration but is an ecological aggregate of these processes. The migration parameter of neutral models then appears more as an emerging property, revealed by neutral theory instead of being an effective mechanistic parameter. Thus, spatially implicit models should be rejected as an approximation of forest dynamics and chapter six therefor uses the spatially semi-explicit approach to test an important distinction in patterns of diversity: those at local and those at regional scales.

Most of the theoretical body of work focused only on single scales of diversity, either regional or local scales. In **chapter six** we added a level of biological reality to predictions from neutral theory, simultaneously focusing on both local and regional scales of ecosystems. We used the three different datasets introduced in chapter two, with estimates of diversity from chapter four, and migration from chapter five and studied predictions at both local and regional scales to assess the scalability of neutral theory and whether correct regional predictions follow from accurately reflected local dynamics. Our results presented a novel interpretation of neutral models: no matter how well parameterized or how well the output of simulations fitted the regional patterns, an accurate simultaneous prediction on both regional and local diversity patterns was impossible to attain. Specifically, the dominance of species at local levels could not be predicted accurately. Thus, as dispersal limitation is the only mechanism in neutral models, we concluded that other non-neutral processes must be at work, at least at the local level. This, however, does not provide information on the relative importance of both deterministic and neutral process, only that neither can be solely responsible for community assembly.

Our analyses throughout the chapters of this dissertation have revealed a clear violation of the assumptions of neutral theory and revealed important caveats in the foundation of the theory itself to take into account with respect to scaling predictions up or down from local to regional scales and vice versa. In **chapter seven** we therefor set to provide a new perspective by using principles from information theory to solve ecological problems. More specifically, we used the Maximum Entropy Formalism on a large scale by applying it to a large database of tropical tree diversity: the Amazonian Tree Diversity Network. We were able to quantify the relative importance of deterministic and neutral based processes in structuring community composition. We were able, for the first time, to estimate the actual potential size of the regional species pool from which local communities most likely draw recruits and studied the relative importance between relative abundance of genera, their functional traits and the regional species pool abundances, providing a quantification of selection for certain traits versus chance and migration. We showed an overall low, but strong, environmentally dependent effect of specific functional traits on genus level composition. Traits associated with, for instance, nutrient limitation were indeed positively correlated with higher abundances in nutrient-limited habitats. In addition, we showed very strong effects of dispersal from the regional taxonomic pool into each local community relative to functional traits across large distances, accompanied by a strong spatial pattern that depends on geographical distance.

Chapter eight is the final chapter in which I synthesize all results and put them in a wider perspective. It not only provides the closing statement of this thesis, addressing the questions posed at the start but also suggestions for future research. Given the results presented in this thesis I argue that we need a much more integrated view of ecosystem dynamics, encompassing not only those on ecological, but also at evolutionary time-scales. Using various fundamental principles of life from different dimensions (or in biological terms, different levels of organization) into a single unifying concept I believe it is time for a theory that remains simple in its foundation, one could say even neutral, but connects multiple trophic levels interacting in explaining dynamics of community structure. By linking dynamics at ecological time scales to those at evolutionary time scales and generalizing fundamental ideas from game theory to an eco-evolutionary application we can extend multiple principles from both fields such as specialization, competition, stochasticity, predator-prey interactions and adaptive fitness landscapes towards explaining community dynamics as a whole.

In conclusion, we are still a long way off from truly understanding the game of life but at least we are getting closer. And providing an integrated approach unifying concepts from ecology and evolution can bring us a step further in understanding the origin and maintenance of the vast diversity on our planet.

View from the Rupununi Savannah, Guyana.

Samenvatting

Op elke schaal van het leven, van microscopische bacteriën tot de torenhoge reuzen van bomen in het Amazonegebied, zien we een soortgelijk patroon van diversiteit waarin niet alle soorten even veel voor kunnen komen in een bepaald gebied. In plaats daarvan is er een patroon waarin de algemene soorten zeldzaam zijn en de zeldzame soorten juist veel voorkomen. Maar hoe is dit schijnbaar universele patroon van diversiteit ontstaan en hoe wordt het onderhouden? Welke mechanismen verklaren de zeldzaamheid van de ene soort en de algemeenheid van een ander? Het antwoord op deze vraag geeft ons een fundamenteel begrip van het leven en de drijvende krachten van diversiteit, in het verleden, het heden en de toekomst. Dit onderwerp is daarom al generaties lang beproefd en getest door een groot aantal theoretische en empirische studies. De verschillende benaderingen bij het oplossen van dit raadsel kunnen grofweg worden gecategoriseerd in twee verschillende perspectieven: ofwel processen die worden aangestuurd door generaties van natuurlijke selectie die taxa hebben gevormd en veranderd om anderen te overtreffen in specifieke gebieden of stochastische gebeurtenissen, een nooit eindigend kansspel dat de gemeenschapsstructuur verandert volgens vaste wetten van kansberekening. In zekere zin zou je kunnen zeggen dat elk organisme worstelt voor overleving of de dobbelstenen gooit in het casino van het leven. Het kwantificeren van het relatieve belang van deze perspectieven was het hoofdthema van dit proefschrift.

In de **hoofdstukken 1 en 2** beginnen we met een inleiding in het hierboven beschreven debat, waarbij lezers de nodige achtergrondinformatie krijgen over de theoretische aspecten van de evolutionaire ecologie. We leggen uit dat de klassieke nichetheorie stelt dat soorten zijn aangepast aan bepaalde omgevingskenmerken en dat de samenstelling van populaties voornamelijk wordt bepaald door omgevingsfilters. De neutrale theorie stelt daarentegen dat de dynamiek van gemeenschappen een neutrale basis heeft en dat de samenstelling voornamelijk wordt bepaald door stochastische demografische gebeurtenissen en dispersie(limitatie). In deze eerste twee hoofdstukken presenteren we een kort overzicht van zowel niche als neutraal gebaseerde theorieën, een korte historische en theoretische synthese van beide perspectieven. Echter, ongeacht welke van deze twee perspectieven wordt getest, empirische gegevens over diversiteit worden vrijwel altijd gebruikt om ideeën en theorieën te testen. Gegevens die zorgvuldig worden verzameld en geïdentificeerd. Ondanks grote inspanningen zijn ecologen echter vaak niet in staat om alle collecties te identificeren, waardoor ze gedwongen worden om deze niet-geïdentificeerde records volledig weg te laten, zonder het effect hiervan te kennen, of om zeer kostbare en tijdrovende pogingen na te streven om ze te identificeren. In **hoofdstuk drie** hebben we daarom de consequentie onderzocht van het weglaten van dergelijke niet-geïdentificeerde records. Onze resultaten lieten een hoge overeenkomst zien tussen analyses wanneer alleen de geïdentificeerde in plaats van alle morphospecies (inclusief de niet-geïdentificeerde records) werden gebruikt, wat suggereert dat patronen van gelijkenis en samenstelling erg robuust zijn. Met andere woorden, het hebben van een groot aantal ongeïdentificeerde soorten in een dataset hoeft onze conclusies niet altijd zo extreem te beïnvloeden als vaak wordt gedacht. Dit maakt het ook mogelijk om bewuste beslissingen te nemen over het al dan niet gebruiken van niet-geïdentificeerde records in ecologische analyses met betrekking tot soortensamenstelling.

Hoofdstuk drie liet echter zien dat voor soortendiversiteit dit onderscheid tussen volledig en onvolledig geïdentificeerde soorten wel van belang is. Dit brengt ons bij het belangrijke aspect van het accuraat kunnen schatten van soortenrijkdom. Voor veel ecologische vragen is een van de eerste problemen hoeveel soorten er vermoedelijk in het totale gebied voorkomen, gegeven een inventarisatie van een subset van dit gebied. De schatting van de soortenrijkdom is dan ook een van de meest gebruikte analyses door ecologen, en niet-parametrische schatters zijn daarbinnen waarschijnlijk de meest gebruikte technieken om dergelijke schattingen uit te voeren. In **hoofdstuk vier** hebben we de aannames en resultaten van niet-parametrische schattingen en die van een logserie benadering (een mathematische beschrijving van een continue kansverdeling) voor het schatten van soortenrijkdom voor gesimuleerde tropische bossen en vijf datasets uit het veld getest. We concludeerden dat niet-parametrische schatters niet geschikt zijn om de soortenrijkdom in tropische bossen te schatten, waar de bemonsteringsintensiteit meestal laag is en de rijkdom hoog, omdat juist deze aannames van de methoden niet voldoen aan de bemonsteringsstrategie die in de meeste onderzoeken werd gebruikt. We laten ook zien dat de logserie, die ook substantiële bemonstering vereist, veel effectiever is in het schatten van soortenrijkdom dan algemeen gebruikte niet-parametrische schatters. De aannames ervan passen ook beter bij de manier waarop veldgegevens worden verzameld.

De resultaten van de hoofdstukken drie en vier baanden de weg voor de ontwikkeling van onze theoretische modellen waarin we het belang van migratie en stochastische gebeurtenissen bij het structureren van gemeenschapssamenstelling hebben getest. De laatste noodzakelijke stap voordat we onze modellen daadwerkelijk konden testen, was om een van de kernparameters van neutrale modellen nauwkeurig te schatten: migratie. Met veel geavanceerde methoden beschikbaar voor het schatten van migratie, worden ecologen vaak geconfronteerd met de moeilijke beslissing van het kiezen voor hun specifieke lijn van onderzoek. In **hoofdstuk vijf** hebben we verschillende methoden getest en vergeleken, robuustheidstests uitgevoerd en deze methoden toegepast op grootschalige empirische data. We selecteerden vijf verschillende methoden en vergeleken hun vermogen om migratie te schatten vanuit een ruimtelijk impliciet en semi-expliciet model. In het eerste geval is er geen ruimtelijke relatie tussen lokale gemeenschappen en de regionale soortengemeenschap, terwijl in het laatste geval juist wel een afhankelijkheid bestaat tussen de afstand en de waarschijnlijkheid van migratie. Deze ruimtelijke relatie is belangrijk omdat het lijkt op wat we in het echte leven zien - niet alles kan overal komen, omdat veel soorten een beperkte verspreiding hebben (bijvoorbeeld omdat ze grote vruchten of zaden hebben). De meeste methoden waren in staat de migratie van ruimtelijk impliciete simulaties nauwkeurig te schatten. Voor ruimtelijk semi-expliciete simulaties bleken schattingen echter het opgetelde effect van migratie van aangrenzende plots en de regionale (gesommeerde) collectie van soorten, d.w.z. de metagemeenschap. De schatting was alleen accuraat als de migratie uit de metagemeenschap groter was dan die van aangrenzende plots. We concludeerden dat geschatte migratie meer een benadering is van de homogenisatie onder lokale gemeenschappen door de tijd heen dan een directe meting van migratie en heeft daarom een directe relatie met bètadiversiteit. Maar aangezien bètadiversiteit het resultaat is van vele (niet) neutrale processen, hebben we ook geconcludeerd dat migratie, zoals geschat vanuit een ruimtelijk expliciete wereld, niet alleen directe migratie omvat, maar eigenlijk een ecologisch aggregaat is van al deze processen. De migratieparameter van neutrale modellen lijkt dan meer een intrinsieke eigenschap onthuld door neutrale theorie in plaats van een effectieve mechanistische parameter en ruimtelijk impliciete modellen moeten worden afgewezen als een benadering van bosdynamiek. Hoofdstuk zes gebruikt daarom dan ook de ruimtelijk semi-expliciete benadering om een belangrijk onderscheid te testen in diversiteitspatronen: die op lokaal niveau en die op regionale schaal.

Het grootste deel van de theoretische hoeveelheid werk concentreert zich alleen op enkele schalen van diversiteit, ofwel regionaal of lokaal. In **hoofdstuk zes** hebben we daarom een niveau van biologische realiteit toegevoegd aan voorspellingen uit de neutrale theorie, waarbij zowel lokale als regionale schalen van ecosystemen tegelijkertijd worden bekeken. We gebruikten de drie verschillende datasets die in hoofdstuk twee werden geïntroduceerd, met schattingen van diversiteit uit hoofdstuk vier, en migratie uit hoofdstuk vijf en bestudeerde voorspellingen op zowel lokale als regionale schaal om de schaalbaarheid van neutrale theorie te beoordelen en of correcte regionale voorspellingen volgen uit correct geïnterpreteerde lokale dynamiek. Onze resultaten gaven een nieuwe interpretatie van neutrale modellen: hoe goed geparametriseerd of hoe goed de output van simulaties paste bij de regionale patronen, een nauwkeurige gelijktijdige voorspelling van zowel regionale als lokale diversiteitspatronen was onmogelijk te bereiken. Met name de dominantie van soorten op lokale niveaus kon niet nauwkeurig worden voorspeld. En aangezien gelimiteerde verspreiding van individuen het enige mechanisme is in neutrale modellen concludeerden we dat andere (niet-neutrale) processen aan het werk moeten zijn, in ieder geval op lokaal niveau. Dit gaf echter nog geen informatie over het relatieve belang van zowel het deterministische als het neutrale proces, alleen dat geen van beide als enige verantwoordelijk kan zijn voor het structuren van de compositie van gemeenschappen.

Onze analyses in de hoofdstukken van dit proefschrift hebben een duidelijke schending van de veronderstellingen van de neutrale theorie onthuld en wezen op belangrijke kanttekeningen van de theorie zelf om rekening mee te houden met het omhoog of omlaag schalen van voorspellingen, van lokale naar regionale schaalniveaus en vice versa. In **hoofdstuk zeven** hebben we daarom een nieuw perspectief gegeven door principes uit de informatietheorie te gebruiken om ecologische problemen op te lossen. We hebben hiertoe het Maximum Entropy principe op grote schaal toegepast op een grote database van tropische boom diversiteit: het Amazonian Tree Diversity Network (ATDN). Op deze manier konden we het relatieve belang van deterministische en neutraal gebaseerde processen in het structureren van gemeenschapssamenstelling kwantificeren. Ook konden we voor het eerst de werkelijke potentiële omvang schatten van de regionale soortenpool waaruit de lokale gemeenschappen mogelijk hun rekruten trekken. Ook hebben we de relaties tussen de lokaal relatieve abundantie van genera, hun functionele kenmerken en hun abundanties in de regionale soortenpool bestudeerd. Dit leidde uiteindelijk tot een kwantificering van de mate van selectie voor bepaalde kenmerken versus kans en migratie. We toonden een algemeen laag, maar sterk van het milieu afhankelijk, effect van specifieke functionele kenmerken op de samenstelling van compositie op genusniveau. Kenmerken in verband met bijvoorbeeld een lage hoeveelheid voedingsstoffen waren inderdaad positief gecorreleerd met hogere abundanties in habitats met een beperkte hoeveelheid nutriënten. Ook toonden we zeer sterke effecten van verspreiding van de regionale taxonomische pool in elke lokale gemeenschap ten opzichte van functionele kenmerken over grote afstanden, vergezeld van een sterk ruimtelijk patroon dat afhankelijk is van geografische afstand.

Hoofdstuk acht is het laatste hoofdstuk waarin ik alle resultaten samen heb gevoegd en in een breder perspectief heb geplaatst. Het biedt niet alleen de slotverklaring van dit proefschrift, maar behandelt ook de vragen die aan het begin werden gesteld, alsmede suggesties voor toekomstig onderzoek. Gezien de resultaten gepresenteerd in dit proefschrift, betoog ik dat we een veel meer geïntegreerd beeld van de dynamiek van ecosystemen nodig hebben: niet alleen die op ecologische, maar ook op evolutionaire schalen. Door gebruik te maken van verschillende fundamentele principes van het leven vanuit verschillende dimensies (of in biologische termen, verschillende organisatieniveaus) in een enkel verenigend concept, geloof ik dat het tijd is voor een theorie die eenvoudig van opzet blijft, zelfs neutraal zou kunnen zijn, maar meerdere trofische niveaus integreert in het verklaren van de dynamiek van de gemeenschapsstructuur. Door dynamieken op ecologische tijdsschalen te koppelen aan die op evolutionaire tijdschalen en fundamentele ideeën te generaliseren van speltheorie tot een eco-evolutionaire toepassing, kunnen we meerdere principes uit beide gebieden uitbreiden, zoals specialisatie, competitie, stochasticiteit, roofdier-prooi-interacties en adaptieve fitnesslandschappen naar het verklaren van gemeenschapsdynamiek als geheel.

In conclusie, we zijn nog ver verwijderd van het echte begrip van het spel van het leven, maar we komen in ieder geval dichterbij. En door een geïntegreerde aanpak te bieden die concepten uit ecologie en evolutie verenigt, kunnen we een stap verder komen in het begrijpen van de oorsprong en het behoud van de enorme diversiteit op onze planeet.

View from Caxiuanã, Brazil. The place where I first experienced both the beauty and riscs of being in the Amazonian rainforests.

Acknowledgements

The journey towards this thesis has been both challenging and rewarding, but never solitary. A long time ago, Isaac Newton wrote a letter to Robert Hooke saying that "*If I have seen further than others, it is because I've stood on the shoulders of giants*". This also is the opening statement of my dissertation. I felt this was appropriate as my work and as I believe all scientific work only comes forth from previous discoveries. This also is the reason why each chapter starts with a portrait of such a giant in their own respective field. However, I also see this analogy stretching further than just this scientific point of view. My family and friends, in which I have also stood on the shoulders of giants, gave me the discipline, courage and sometimes (un)necessary distraction needed to bring this towards its conclusion. In these final pages I want to look back the past six years, to take you on the short version of this journey and providing a note of thank you to all those giants who made it possible or joined me along the way and without whom this work could not have been finished.

I could not have imagined that almost ten years ago a single meeting would have such a profound change on my life. It was here that my first research internship would also mean the start of my love and interest for Amazonian rainforests and my scientific career, both in research and education. In this single meeting, Hans ter Steege (to become my daily supervisor and promoter) provided me with his trust to successfully go on an expedition to the Brazilian rainforest to climb trees and collect data on vascular epiphytes, which also culminated in my first scientific publication. In addition, he also started me on my path as an educator, beginning as a student-assistant in two courses simultaneously: evolutionary biology and an ecological field course which would also change my life forever. Hans, in the years that followed you were not only my supervisor but also became a good friend and mentor. I owe you much more than I could possible express here in words but I will try in any case, as you would have probably guessed. I thank you for the support, both scientifically and emotionally from time to time, encouragement but foremost for your trust in allowing me to find my own way and to work on my own ideas. This freedom has been essential in my development as a scientist, permitting me to generate new thoughts and ideas to which you always had an open mind. In teaching you taught me to be honest, critical but constructive, as you have always been towards my work.

The past six years you helped me along every step of the way, in science, teaching and in developing my career, helping me making decisions, structuring my thoughts and ideas and sometimes also being the necessary brake to my seemingly unlimited ambitions and sometimes bold decisions. I will never forget the many hours and laughs we had wandering on Terschelling during the field course which I continue to teach to this day, teaching students the joys of fieldwork or still discussing the gravitational forces that generate our tides well into the night. Nor will I forget our Guyana expedition where I learned much from you about being in the rainforest, being a field biologist, having respect for those who live in and with the forest and to see the beauty of the forest. These and many more experiences are forever engrained in my memory.

In these first stages of my career there is another person without whom this could never had been realized and who helped me developed my teaching skills enormously: Roy Erkens, now a scientist and teacher at Maastricht University. Roy was involved in the evolutionary biology courses in which I started as a student assistant but gradually became more a teacher than an assistant, giving lectures when he could not or coordinating assignments and much more. After finishing my masters degree Roy offered me a job in becoming co-coordinator and teacher in his courses to give me time to write a research proposal to gain funding. Although this proposal was not funded, Roy played a pivotal role in another way: he decided to go from Utrecht back to Maastricht and suggested to the Department that I take over his position teaching Evolutionary Biology at Utrecht University. To make a long story a bit shorter, they agreed and a dual appointment was created in which I provided teaching on evolutionary biology to realize my own PhD project. And even though my PhD project is now finished, I still teach Evolutionary Biology and in the mean time have expanded this curriculum and continue to develop this at Utrecht University. In my time with Roy, teaching the many courses, he provided me with much feedback and helped shaped me into the teacher I now am today, telling me I had a natural predisposition towards teaching and that whatever I do I should at least try to incorporate this into my career. Roy, thank you for the unwavering support, many (many) laughs and discussion, I would not be where I am today without your help and support. During my time teaching and writing a research proposal there were also many at Utrecht University who provided me with short-term projects to bridge the time until I found a way to realize this PhD project. One in particular that I would like to mention is Merel Soons. She has been my office-roommate for the past eight years and hopefully many more years to come. She also provided projects that helped me get by, letting me do part of her fieldwork in the Netherlands, organizing and sifting through her years of collected data. She was and still is also part of the great team that teaches the field course mentioned before and I could not have wished for a better roommate. Merel, thank you for all the lively discussions, questions, laughs and support during the years, I hope many more will follow. One other member from the staff of the field course has provided me with years of good memories and lively discussions that I wish to mention: Hans Persoon, whose knowledge and love of plants and taxonomy is beyond comprehension. Hans, thank you for the opportunity to learn from you and the great memories we share. I also want to take this opportunity to thank everybody at the research group of Ecology and Biodiversity throughout the years with whom I have shared many a laugh and discussion during coffee breaks, dinners and outside lunch walks. I am afraid I will forget many, for which I ask forgiveness, but a few, in no particular order, come to mind: Erik, Jelle, Jeroen, Joost, Jan, Rob, Heinjo, Boudewijn, Marinus, Feike (who taught me the appreciation of mathematics and jump-started my journey into theoretical biology), Jos, Marijke, Yann, George, Mariet, Arjan, Bas, Betty, Gerrit, Jasper, Amber, Bjorn, Niels, Pieter and Bertus (who was also always in early like me and I have shared many a coffee and good conversation with), Peter, Peter and many many more people. I also want to thank all the students that I have taught throughout the years, for your questions and curiosity have driven me to always give my best and to be creative in my teaching. To see somebody truly understand something for the first time and to be able to structure their own ideas and thoughts from that understanding is one of the greatest gifts of scientific development. I also want to thank Simon Levin, who gave me the opportunity to spend time in his lab at Princeton University where I learned much and spend a great deal of time with great minds. I hope the future will bring much collaboration.

I also owe a great deal to my friends, both in Utrecht and Nunspeet who provided me with so many laughs and memories that I could write a whole separate book on just this subject. Gijs, Rens, Stefan, Sander and Jelmer, throughout the years we have become as close as friends that can be and I cherish each and every memory (both good and bad). Many of us have known each other now for over ten years, each developing our own unique perspective but always remaining good friends surrounding that one pivotal moment in history: the birth of our underground committee within the Utrechtse Biologen Vereniging (The Utrecht Biologists Association) which we dubbed 2 Punten Pino. Starting on that same field course in 2006 when we participated ourselves it has been an ever-present glue that binds us together and although we are (perhaps finally) growing up, I hope to prolong our friendship indefinitely and still return to being those students from time to time. I also want to thank all of my friends in Nunspeet with whom I grew up and whose company I still enjoy during the weekends when I return to the forests where I grew up. You have provided me with so many good memories and late-night discussions over a few beers that I will always want to return to my old home. Starting over twenty years ago I consider many of you more than just friends, having been through many good and sometimes also tough times but always having each other's backs. I will treasure those moments at two o'clock at night when suddenly I was asked to explain theories of relativity (even though everybody knew I was a biologist and not a physicist) or to talk about me latest results of my project and finally walking home in the forest in the early mornings enjoying the company of wildlife and of the last stars. Your never-ending curiosity, but always with a slight sarcasm, sometimes forced me to rethink my results and many a beer-coaster with scribbling's of good ideas still remain somewhere in my stacks of paper that I will no doubt find again someday.

I also want to thank Irmgard, who has to put up with my never-ending desire to learn more about our natural world, who sometimes needs to force me to stop working and spend more time with her because she deserves it, although I do not tell her this enough. I am also grateful for my dogs: Fenrir, Emin and June who teach me to have patience every day and although it sometimes is hard to combine everything in my life they ground me in a very strong and uncomplicated way, they provide me with an ancient and unconditional bond that makes life that much happier. And last but certainly not least, I owe an enormous part of who I am to my family: my parents, brother and sister, uncles and aunts, grand parents and all that have been included or lost throughout the years. I am grateful for each and every experience we share for it made me who I am today. To my parents, whose unwavering support in all my efforts, both academic and personal, has been pivotal in shaping my personality and character, I owe a great deal. I remember as a small boy, in the library, walking over to the science section and picking up the book that would change the way I looked at life. That book was Charles Darwin's on the Origin of Species and instead of taking it out of my hands and replacing it with a children's book, you encouraged me to read and to understand it (followed by Albert Einstein's Relativity). This signifies how you raised us by giving us strength, discipline, courage and curiosity. You taught us never to be afraid of what is to come, to combine knowledge with wisdom, to always give your best, and to never, ever give up. This has made me who I am today.

Thank you.

Curriculum Vitae

Edwin Pos was born on the 29th of January 1987 in Nunspeet, the Netherlands. Growing up in one of the largest forested areas in his country meant that from an early age he developed a profound fascination for all that grows, crawls and walks among these trees. His parents stimulated this fascination by joining on many walks in the early morning, trying to spot wild boar, deer, badgers and foxes, although the latter always remained quite elusive. He furthered this by joining the scouting as a small boy, a decision that would have a major effect giving him discipline, responsibility and many virtues that played a large part in bringing this dissertation to a success. He still remains with the scouting to this day, albeit less active.

It was also clear that Edwin had a strongly rooted desire to understand all that is around us, reading books like *On the Origin of Species* by Charles Darwin and *Relativity* by Einstein while still being in elementary school. This never ending ambition ultimately lead Edwin to pursue a bachelor degree in Biology at Utrecht University, that was followed by a Masters degree at the same institution. Although he left his home town to be able to fully devote himself to his studies, those roots never truly left and many a times he still visits the forests of the Veluwe, roaming there with his dogs.

During his academic education at Utrecht University it became clear that his main interest lay in the understanding of the diversity of life, the incredible and alluring idea that we all share a common ancestor and that forces still acting today were shaping this life into a myriad of variations. But it was not until his first internship during his Master degree that he was going to be hooked on tropical forests. One single meeting with Hans ter Steege would change his life forever, paving the way towards this doctoral thesis. For this internship Edwin would venture deep into the Amazonian jungle without any supervision and climb trees to collect vascular epiphytes such as orchids and bromeliads. This meant much physical and theoretical preparation but also required a highly independent and self-disciplined attitude as there would be no one to guide them while on Brazilian soil. Edwin was adamant to successfully finish this internship and even an accident that resulted in a severe head-injury some 400 km inside of the Brazilian rainforest did not stop him.

Instead, it reinforced his idea of becoming a scientist and ultimately the results of this internship were published, becoming the first scientific paper of his career. Following this internship, Edwin pursued the study of roe-deer management in the Netherlands but quickly realized that although important, his interests lay more with the fundamental instead of the applied sciences. And although two internships were enough to finish his Masters degree he managed to pursue a third internship studying elephants in Kenya for many months sleeping on the savannah where his endurance would be tested once more due to rapid floods that wiped out the entire base-camp.

After finishing his Masters degree he was asked by Hans ter Steege to pursue a PhD project, creating a proposal that was unfortunately rejected two times. During the writing of this proposal many people from the University helped to keep him around by assigning him many small projects. One single red thread throughout these years was his teaching as he was involved in all evolutionary biology courses and a field ecology course, first as a student assistant, later as junior lecturer and lecturer. It was this endeavour that would ultimately make it possible to start his PhD project, when one of his colleagues left Utrecht University. This left a gap in the education curriculum and it was suggested that Edwin could fill this gap and in return the University would grant him is PhD project. This was six years ago and in the meantime Edwin has finished with the latter but still teaches evolutionary biology and that same field ecology course that set him on this track almost ten years ago. He has been given a permanent position as a University lecturer at Utrecht University and is pursuing his own ideas in finding a unifying concept to explain the patterns in the diversity of life.

Current Position

University Lecturer
Evolutionary Biology and Ecology
Department of Biology, Utrecht University

Publications (incl. submitted/ in preparation)

2018 Gomes, V. et al (including **Pos, E.T.**) (2018). Species Distribution Modelling: Contrasting presence-only models with plot abundance data. Scientific reports, 8(1), 1003.

2017 Pos, Edwin, Guevara Andino, Juan Ernesto, Sabatier, Daniel, Molino, Jean-François, Pitman, Nigel, Mogollón, Hugo, Neill, David, Cerón, Carlos, Rivas, Gonzalo, Di Fiore, Anthony, Thomas, Raquel, Tirado, Milton, Young, Kenneth R, Wang, Ophelia, Sierra, Rodrigo, García-Villacorta, Roosevelt, Zagt, Roderick, Palacios, Walter, Aulestia, Milton & ter Steege, Hans. Estimating and interpreting migration of Amazonian forests using spatially implicit and semi-explicit neutral models. Ecology and Evolution, 7(12) 4254-4265

2017 ter Steege, H., Sabatier, D., de Oliveira, S. M., Magnusson, W. E., Molino, J. F., Gomes, V. F., **Pos, E.T**. & Salomão, R. P.. Estimating species richness in hyper-diverse large tree communities. Ecology, 98(5), 1444-1454.

2016 E. T. Pos et al., Biodiversity Assessment Survey of the South Rupununi Savannah Guyana - Chapter 1 Plants of Southwest Guyana: Rupununi Savannah and Parabara region.

2014 Pos, Edwin, Guevara Andino, Juan Ernesto, Sabatier, Daniel, Molino, Jean-François, Pitman, Nigel, Mogollón, Hugo, Neill, David, Cerón, Carlos, Rivas, Gonzalo, Di Fiore, Anthony, Thomas, Raquel, Tirado, Milton, Young, Kenneth R, Wang, Ophelia, Sierra, Rodrigo, García-Villacorta, Roosevelt, Zagt, Roderick, Palacios, Walter, Aulestia, Milton & ter Steege, Hans. Are all species necessary to reveal ecologically important patterns? Ecology and Evolution, 4 (24), 4626-4636

2010 Pos, E. T., & Sleegers, A. D. M.,Vertical distribution and ecology of vascular epiphytes in a lowland tropical rain forest of Brazil. Boletim do Museu Paraense Emílio Goeldi, Ciências Naturais, 5(3), 335-344.

In review Pos, Edwin, Guevara Andino, Juan Ernesto, Sabatier, Daniel, Molino, Jean-François, Pitman, Nigel, Mogollón, Hugo, Neill, David, Cerón, Carlos, Rivas, Gonzalo, Di Fiore, Anthony, Thomas, Raquel, Tirado, Milton, Young, Kenneth R, Wang, Ophelia, Sierra, Rodrigo, García-Villacorta, Roosevelt, Zagt, Roderick, Palacios, Walter, Aulestia, Milton & ter Steege, Hans. Adding biological reality to general predictions of neutral theory reveals scaling issues between local and regional patterns of diversity. *In review at Ecology Letters*

In review Pos, E.T. et al. Rolling the dice or struggling for survival, using Maximum Entropy to unravel drivers of community composition. *In review at Science Advances*

In prep Kong, J, **Pos, E.T**., Hao, W. Macroscopic epidemiological cycles result from Microscopic bacteria-phage cycles. *In preparation.*

In prep. Kong, J., Chadi, S.M., Cooney, D., **Pos, E.T.**. Dynamics of an indirect cholera transmission model with immunological threshold and temporal immunity. *In preparation.*

In prep. Pos, E.T., Hautier Y.. ExtractR, an R-package and practical guide to extract and visualize informative statistics: practicing what statisticians preach.
In preparation.

Grants and Funding

2013	Miquel Fund, funding for South-Guyana expedition (1.5K)
2013	Alberta Mennega Foundation, funding for South-Guyana expedition (1K)
2008	Alberta Mennega Foundation, funding for first MSc internship (.8K)
2008	Kronendak, funding for first MSc internship (.6K)
2008	Van Eeden Fund, funding for first MSc internship (.625K)
2008	Miquel Fund, funding for first MSc internship (.8K)
2007	Trajectum Scholarship (.23K)

Awards and Nominations

2017 Nominated teacher of the year, department of Biology Utrecht University
2015 Nominated and Awarded teacher of the year
2014 Nominated teacher of the year

Presentations

2017 Speaker on the Netherlands Annual Ecology Meeting, two-day event organised by NERN and NecoV (Dutch - Flemish Ecological Society) supported by the Netherlands Organisation for Scientific Research.

2016 Speaker on the symposium "Current Research in Tropical Ecology in the Netherlands", organised by the University of Amsterdam

2015 Speaker on the launch of the "Academy of Ecosystem Services", organised by the University of Utrecht.

2015 Joint Scientific Thematic Research Programme: China Annonaceae Meeting. Symposium on diverse research topics regarding Annonaceae to plan further research opportunities.

2014 Speaker for the "Childrens University" at Naturalis Biodiversity Center, talks organised to stimulate children to pursue a scientific career.

2013 Speaker on the Tropical Forest Careernight at the University of Utrecht, organised by the VTB ("Association Tropical Forests")

2009 Darwin Symposium, part of the organizing team at the University of Utrecht under the name: "Darwin's legacy, the influence of evolutionary thinking on science".

Other

2016 – Present Owner and founder of Edwin Pos – Dog evolution and Behavior, dog behaviour consultancy and symposia to educate on dog evolution and behaviour.

2012 – Present Part owner and founder at Dactylis VOF
Ecological consultant agency

2012 – 2014 Ministry of Defence, soldier first class, National Reserve 11th Airborne brigade, 20th Bataljon, Echo Company.