ACADÉMIE
DES SCIENCES
INSTITUT DE FRANCE

# Comptes Rendus
# Biologies

Jerome Chave

**Species abundance, urn models, and neutrality**

Volume 347 (2024), p. 119-135

Online since: 2 October 2024

https://doi.org/10.5802/crbiol.162

This article is licensed under the
Creative Commons Attribution 4.0 International License.
http://creativecommons.org/licenses/by/4.0/

*The* Comptes Rendus. Biologies *are a member of the Mersenne Center for open scientific publishing*
www.centre-mersenne.org — e-ISSN : 1768-3238

ACADÉMIE DES SCIENCES
INSTITUT DE FRANCE

*Comptes Rendus. Biologies*
**2024**, Vol. 347, p. 119-135
https://doi.org/10.5802/crbiol.162

Review article

# Species abundance, urn models, and neutrality

**Jerome Chave** [a]

[a] CRBE, Toulouse, France
*E-mail:* jerome.chave@cnrs.fr

**Abstract.** The neutral theory of biodiversity and biogeography has stimulated much research in community ecology. Here, exact results are used to apply neutral model predictions to large regional samples. Three complementary neutral models are presented: the Ewens canonical neutral model, a model of subdivided ecological communities, and a "diversity begets diversity" neutral model. For all three models, an exact sampling formula is provided, and a new R package `neutr`, is presented. This package is used to fit species abundances from regional inventories of tropical forest trees in the Amazon, tropical Africa and Southeast Asia. It is shown that the neutral models fit well empirical data for all but the few most abundant species (from 6 to 40 depending on the continent). When the parameter $\theta$ is taken as an index or regional diversity, the Amazonia and Southeast Asia emerge with similar regional diversities ($\theta = 654$ for Amazonia, versus $\theta = 726$ for Southeast Asia), with a less diverse tropical African tree flora ($\theta = 219$). The model infers 10,141 tree species with at least 50 individuals in Amazonia, 3477 in tropical Africa and 9915 in Southeast Asia. The spatially subdivided neutral model provides clear evidence for a spatial substructure in all three regional floras. These results show how neutral models are useful to explore regional patterns of species abundance and to provide insights about regional species pools.

**Keywords.** Biodiversity, Ecology, Species abundances.

**Funding.** Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-25-01; TULIP, ref. ANR-10-LABX-0041; ANAEE-France: ANR-11-INBS-0001), CNES long-term funding, European Space Agency CCI-BIOMASS and FRM4BIOMASS projects, French Foundation for Research on Biodiversity (FRB).

*Manuscript received 13 May 2024, revised 28 August 2024, accepted 30 August 2024.*

## 1. Introduction

Patterns of species abundance and rarity are an important dimension of biological diversity. A species can be rare because it is limited by its physiology, by local biotic constraints on its abundance, or by its limited geographical distribution [1–3]. These diverse causes for rarity also reflect the fundamentally different reasons why species are threatened. Species that are locally abundant but geographically limited may be threatened with extinction if their habitat is transformed, as is the case with certain species of forest tree in the tropics. Highly specialized sap-sucking insects may be threatened because the trees on which they feed are removed. Finally, the degradation of restricted habitats can lead to the disappearance of species, as for example on the island of Saint Helens [4]. Although rare species exert great fascination, the question of why some species are abundant locally or regionally is no less interesting. The most abundant species are most relevant to many ecosystem processes. In a study of Amazonian rain forest trees, ter Steege et al. have shown that no more than 227 tree species make up half of the trees in this biome, out of more than 6700 species [5]. These were called hyperdominant species. Patterns of abundance for Amazonian rain forest trees [5] are represented in Figure 1 for illustration.

The motivation for this contribution is to understand the processes underlying regional patterns of species abundance. One way to contribute to this research is to ask to what extent empirical species abundance distributions deviate from those of regional species pools generated purely from random

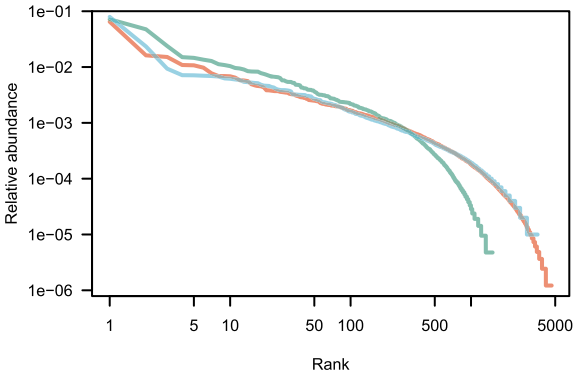


**Figure 1.** Rank abundance distribution of tropical tree species in three continents. These empirical distributions are scaled such that the sum of the relative abundances is equal to one. Red: Amazon; green: tropical Africa; blue: Southeast Asia. Data from [6].

processes. Here, I provide a self-contained treatment of neutral models of relevance in the study of species abundance distributions, together with a code for the algorithms in the R statistical language. Then I illustrate this method with regional species abundance data for three tree flora (Amazonia, Africa, Southeast Asia).

The neutral theory of biodiversity and biogeography [7] has emerged from the mathematics inspired by population genetics theory of the 1970s, and it has generated much debate in ecology. The theory has shed light on questions such as: how are the sampled individuals distributed between species? how many species remain to be discovered after a given number of individuals have been sampled? How representative are samples of larger ecological communities? When I was invited to contribute to the pages of this journal, I took opportunity to return to a debate on neutral models of species abundance that have animated the scientific community in ecology. Neutral models are now in the toolkit of ecologists for the analysis of species abundance [5]. Yet, many important results on neutral models tend to be overlooked in the modern literature and regularly rediscovered [8]. In addition, since the 1970s, important research in probability theory has been developed [9–14], some of which is relevant to quantitative research on biodiversity.

This study explores models of species abundance that mimic the process of species discovery in a real situation, where individual organisms are identified to the species one at a time. The first organism is always a new species in the sample, the second may be a representative of the first species or of a new species, and so on. For such a model to make sense, it is assumed that organisms are identified one by one, through a classical taxonomic study. Bulk identification methods, using mass sequencing of environmental DNA for microbial species provide a different context to the study of biological diversity in that species are substituted for molecular operational taxonomic units, and individuals are not always observable or even clearly defined [15]. With that limitation in mind, it is still helpful to explore the patterns of species abundance in discrete assemblages of organisms [16]. The goal here is to return to known mathematical formulation of neutrality, provide several representations of the model and show that these representations are quite flexible.

Here, I first review the mathematical foundations of two standard models of species abundance, the canonical neutral model, and a spatially subdivided version of this model. I also discuss a model that has not been explored in the context of species abundance distributions, which models the hypothesis that the addition of species to a community may increase the resources and biotic interactions, making that community hospitable to a greater number of species, or in short "diversity begets diversity", as proposed by Whittaker [17][1]. It is curious that this "diversity begets diversity" model has never received proper quantitative mathematical treatment in the ecological literature and one aim of this contribution is to promote this discussion [9, 11].

In Section 2, I present the fascinating mathematical results associated with three neutral models. In Section 3, I illustrate the application of these models

[1]The hypothesis that diversity begets diversity has been extremely popular in the recent literature, yet the history of this catchy term is quite obscure. It is generally traced back to Robert H. Whittaker (1972, p. 216) [17] who writes that "facilitation of increase in species number in interacting trophic levels is reciprocal. We should thus expect diversity to increase in parallel on any adjacent trophic levels and, in fact, throughout the various groups of interacting species that the community comprises." In fact similar ideas are already present in Allee et al. (1949) textbook on animal ecology [18, p. 695]. Vane-Wright (1978) referred explicitly to the diversity begets diversity mechanism in an evolutionary context [19].

to practical examples of parameter inference for empirical samples of tropical forest trees. Finally, I discuss the possible ramifications of the theory of random partitions for the study of empirical patterns of biodiversity.

## 2. Three neutral models of biodiversity

In ecology, models have been referred to as *neutral* in the sense that individuals all have the same prospects of reproduction and mortality, whatever the species to which they belong [20]. In probability theory, the more general notion of *exchangeability* is defined: a model is exchangeable if the probability distribution of the class abundances $\{n_1, n_2, \ldots, n_k\}$, where $n_i$ is the abundance of species $i$, does not depend on the labels of the classes [21,22]. This property is essential in order to obtain exact mathematical results on the probability distribution of the model. This probability distribution, when known, can then be used as a likelihood function to compare the model to empirical species abundance distributions. In Section 2.1, an intuitive construction is provided in the form of urn models.

Model 1 is the canonical neutral model, for which the probability distribution of species abundances is the Ewens sampling formula [23], explained in Section 2.2. I also present the species individual curve for Model 1, and the Griffiths–Engen–McCloskey formula, which allows samples conforming to Model 1 to be drawn in a time proportional to the number of species $k$ rather than the number of individuals $n$. Models 2 and 3 are two possible generalizations of Model 1. Model 2, presented in Section 2.3, assumes a spatially subdivided system, with limited dispersal between local sites. Under general assumptions, this Model 2 is also associated with a probability distribution for species abundances, and the model parameters can be estimated by maximal likelihood estimation. The second generalization, Model 3 (Section 2.4), implements the "diversity begets diversity" model. It turns out to be a equivalent to a model first developed in [9], with the same properties as the canonical neutral model but with the addition of one parameter.

### 2.1. *Urn model representation*

The process of species discovery can be summarized in generic terms using so-called urn models [24]. The "urn" represents the system (here, a sample, or an ecological community), and it is populated by "balls", which represent objects (here, individuals). The objects may belong to two or more classes, which are usually represented by the color of a ball. The urn representation is relevant when an operator picks balls and performs a number of operations based on this sampling. A lottery is an example of a game that can be represented as an urn model, other examples including election systems [25,26] or sports [27].

In the Pólya urn model [24,28], an urn is initially filled with $n_i$ balls of color $i$, and the balls are drawn one by one, being replaced in the urn after its color has been observed. The construction process is as follows. When a ball is drawn, it is replaced in the urn together with a new ball of the same color. The most abundant color tends to be selected more often, so its abundance increases more rapidly. This Pólya urn model resembles the species sampling process, where a color symbolizes a species. Note that even if the process is stochastic, the abundance of each species depends on the initial condition of the system, i.e., the initial abundance of the species.

A slight variant of the Pólya urn model, due to Hoppe [29], is directly related to the problem of species sampling. It assumes that initially the urn contains a single black ball. The construction process is as follows. When the black ball is picked, it is replaced in the urn together with a ball of a completely new color. When any other ball is picked, it is replaced in the urn together with another ball of the same color. All the colored balls have the same chance of being chosen, which we define as a unit "weight". In contrast, the black ball has a weight $\theta$, which is a positive real number. The parameter $\theta$ is proportional to the probability of adding a new color in the urn per iteration. This method yields a partition of the colored balls, and this partition depends on the single parameter $\theta$. This is the basis of Model 1 presented below, also called the "canonical neutral model". It is a drastic simplification of reality, but is amenable to exact probabilistic results.

In large biological assemblages of organisms, geographical structure can become an important factor, and can invalidate the assumption of perfect mixing, i.e., the assumption that there is a single urn from which one samples the balls. For example, the Amazon is a large collection of trees (on the order of $4 \times 10^{11}$, see [5]), covering more than 6 million square

kilometers, and assemblages of tree species differ west and east of the Amazon. Spatially subdivided models have been developed to account for this effect, where local sites are considered as a dispersal-limited sample of the regional pool [7, 30, 31]. The goal is to predict the local distribution of individuals between species in a local community, as a function of immigration rates and regional species diversity [7, 32, 33]. The influence of space on the canonical neutral model is discussed in Model 2 below.

In Model 3, the diversity begets diversity model, the rate of species appearance depends on the number of species in the system, which is denoted $k_n$ (the subset $n$ is because the number of species depends on $n$, the number of individuals). As will be explained below (Section 2.4), one definition of this model is through an urn scheme, sometimes referred to as the Blackwell–MacQueen model [9, 11, 34]. After $n-1$ individuals have been sampled, the probability of sampling an altogether new species (i.e., the probability of sampling the black ball) is equal to $(\theta+\sigma k_n)/(\theta+n)$, so for any value of $\sigma>0$, the probability to pick a new species is proportional to the number of existing species $k_n$. In the special case $\sigma=0$, this model is equivalent to the Hoppe urn model (Model 1). The probability of adding one individual to species $i$ (i.e., the probability of sampling a colored ball) is equal to $(n_i-\sigma)/(\theta+n)$, so each of the colored balls is picked slightly less often than expected by chance, and the rarest colors, represented by a single ball in the urn, are counter-selected. Biologically, rare species are likely to be less viable than expected by chance due to reproductive difficulties, both pre- and postzygotic [35–37]. Crucially, the complete sampling theory is known for this third urn model, as described below, so it lends itself to statistical inference [9, 11].

## 2.2. *Canonical neutral model (model 1)*

The first model describes a natural partition of $n$ objects into $k$ classes. This partitioning could concern many real-life situations, and has been applied in population genetics (number of allelic copies in a population, [23]), linguistics (number of word occurrences in a book [38]), and ecology (species abundance in a sample [39]). The question of how to partition discrete collections is also relevant to many other fascinating problems in mathematics [13, 14]. All the results in this section are classic but they are still provided as they provide essential context to Models 2 and 3. An excellent introduction is found in Ewens' textbook (2004 edition, Chapters 3, 9, and 11, [8]).

In a sample of organisms, let us assume that the species have been labeled: the species differ in detectable ways, such as a taxonomic feature. We denote $n_i$ the abundance of species $i$, and $n=\sum_{i=1}^{k} n_i$ the total number of organisms sampled, where $k$ the total number of distinct species. The sample is fully described by the vector $\{n_1, n_2, \ldots, n_k\}$, and the system is described by a probability distribution $p(n_1, n_2, \ldots, n_k)$ to find the system in a given state. A different description of the state of the system is as follows. Let $a_r$ be the number of species with exactly $r$ individuals in the sample (which can be zero). The total number of individuals is $n=\sum_{r=1}^{\infty} r a_r$. The difference with the above description is that $a_r$ are random numbers, and that the total number of species in the sample $k=\sum_{r=1}^{\infty} a_r$ is the sum of random numbers, and is therefore also a random number. In this second representation, species are unlabeled.

In the Hoppe urn model, step one draws the black ball with probability one, generating one new species. At the $n$th draw, $n-1$ individuals have been sampled, and the probability of sampling an altogether new class (or color) is equal to $\theta/(\theta+n)$, while the probability of adding one object to class $i$ is equal to $n_i/(\theta+n)$. The Hoppe urn model generates a population of $n$ objects, and patterns of class abundance are parameterized only by $\theta$. This construction turns out to be equivalent with Fisher–Wright model of population genetics [8, 29], or the dispersal-unlimited neutral model in ecology [40]. Consider a time-dependent system of $N$ individuals such that all individuals die exactly at the end of the time step and are replaced by a multinomial draw of their offspring. In addition, with probability $\nu$ they are replaced by a totally new species. This means that new species can arise at a rate $N\nu=\theta$ (new species per time step). This system is assumed to be large in the sense that any sample $n$ verifies $n \ll N$. This process soon reaches a dynamic equilibrium where the appearance of new species is balanced by the extinction of rare species. Hoppe [29] has shown that the urn model generates a typical configuration of the above model at dynamic equilibrium.

In the Fisher–Wright model, the probability $F_2(t)$ that two randomly chosen individuals belong to the same species at time $t$ is computed as follows [8, 41, 42]. For two individuals to belong to the same species, none of the two individuals can be a new species (probability $(1-\nu)^2$), and they can descend from the same parent (probability $1/N$) or from two different parents already of the same species at the previous time step (probability $F_2(t-1)$). This reasoning is summarized in the equation: $F_2(t) = (1-\nu)^2(1/N + (1-1/N)F_2(t-1))$. At dynamic equilibrium, $F_2 = F_2(t) = F_2(t-1)$, and substituting in the equation, one finds $F_2 = 1/(1-N+N(1-\nu)^{-2})$. For $N$ large and $\nu$ small, this results in: $F_2 = 1/(1+\theta)$. This result is equivalent with the Hoppe urn model, since the probability of picking any one ball in the urn is $1/(1+\theta)$. The same reasoning can be applied one step further to the probability of picking three individuals of the same species $F_3(t)$. Three situations can arise: (i) all three could descend from the same parent (probability $1/N^2$), (ii) they could descend from two parents (probability $3(N-1)$) already of the same species at time $t-1$ (probability $F_2(t-1)$), or (iii) they could descend from three different parents (probability $(N-1)(N-2)$) already of the same species at time $t-1$ (probability $F_3(t-1)$). At dynamic equilibrium, the formula reads: $F_3 \sim 2/(2+\theta)F_2$. This reasoning for two and three individuals can be extended to compute the probability of picking $n$ individuals of the same species [23], which is $F_n = (n-1)!/(\theta(\theta+1)\cdots(\theta+n-1))$.

The denominator of this expression, $\theta(\theta+1)\cdots(\theta+n-1)$, is an important mathematical quantity in this theory, and for this reason it deserves a specific notation: $\theta^{(n)} = \theta(\theta+1)\cdots(\theta+n-1)$, which is called the *increasing factorial power*, or sometimes the "Pochhammer symbol". $\theta^{(n)}$ may be expressed in terms of the usual Gamma function as follows: $\theta^{(n)} = \Gamma(\theta+n)/\Gamma(\theta)$. An expansion of $\theta^{(n)}$ as a polynomial is known and it turns out to be useful below:

$$\theta^{(n)} = \sum_{k=0}^{n} S(n,k)\theta^k \tag{1}$$

where the coefficients $S(n,k)$ are called the *absolute value of the Stirling number of the first kind.*

The above calculus on the probabilities $F_n$ suggests that the computation of the complete probability distribution of the state $\{n_1, n_2, \ldots, n_k\}$, denoted $p_\theta(n_1, n_2, \ldots, n_k)$, is possible. Ewens [23] has computed $p_\theta(n_1, n_2, \ldots, n_k)$ as a closed-form expression, and Karlin and McGregor have provided a formal proof for this formula by recurrence [43]. This result is known as the Ewens sampling formula [13, 14]:

$$\begin{aligned} p_\theta(n_1, n_2, \ldots, n_k) &= \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)} \frac{n!}{k!\prod_{j=1}^{k} n_j} \\ &= \frac{\theta^k}{\theta^{(n)}} \frac{n!}{k!\prod_{j=1}^{k} n_j}. \end{aligned} \tag{2}$$

Intuitively, the factor $\theta^k$ reflects the selection of the black ball exactly $k$ times, and the denominator $\theta^{(n)}$ is the product of the successive masses in the urn on each of the first $n$ draws. The coefficient $n!/(k!\prod_{j=1}^{k} n_j)$ is valid here in the case of labeled partitions. Using the vector $\{a_1, a_2, \ldots, a_r\}$ instead, the partitions are unlabeled and the Ewens sampling formula is rewritten as follows:

$$p_\theta(a_1, a_2, \ldots, a_r, \ldots) = \frac{\theta^k}{\theta^{(n)}} \frac{n!}{\prod_{j=1}^{n} j^{a_j} a_j!}. \tag{3}$$

Equations (2) and (3) may look scary but in fact the dependence on $\theta$ of $p_\theta$ is quite simple, and the rest of the formula is the result of a combinatorial exercise. Having these formulas available makes it possible to explore a variety of problems, a few of which I report here.

A first natural question is how many classes, or species, are in a sample of $n$ individuals given the parameter $\theta$. From the Ewens sampling formula, the probability $P_{\theta,n}(k)$ of finding $k$ species in the sample of size $n$ must be proportional to $\theta^k/\theta^{(n)}$, and since $P_{\theta,n}(k)$ is a probability distribution, then $\sum_{k=1}^{n} P_{\theta,n}(k) = 1$. So it follows that $\sum_{k=1}^{n} c_k\theta^k = \theta^{(n)}$, where $c_k$ is the proportionality constant (i.e., $P_{\theta,n}(k) = c_k(\theta^k/\theta^{(n)})$). It now becomes clear why Equation (1) was reported above: comparing the two formulas it appears that $c_k$ must be equal to $S(n,k)$, therefore:

$$P_{\theta,n}(k) = S(n,k)\frac{\theta^k}{\theta^{(n)}} \tag{4}$$

which is the first remarkable exact result from Model 1. Next, from the construction of Hoppe urn model, the expected number of species in the sample, $\overline{k}_n = \sum_{k=1}^{n} kP_{\theta,n}(k)$ increases with $n$ as

$$\overline{k}_n = \sum_{j=0}^{n-1} \frac{\theta}{\theta+j}. \tag{5}$$

For large sample sizes $n$, it is useful to replace the summation term by a closed-form expression. It turns out that the above formula is equal to:

$$\overline{k}_n = \theta\psi^{(0)}(\theta+n) - \theta\psi^{(0)}(\theta) \quad (6)$$

where $\psi^{(0)}(\theta) = (\mathrm{d}/\mathrm{d}\theta)\ln(\Gamma(\theta))$ is the digamma function (the first derivative of the log-Gamma function). This formula must respect the initial condition that there must be exactly one species in the sample if $n = 1$, or: $\overline{k}_1 = 1$. It is true, although not immediately obvious, that $\overline{k}_1 = \theta\psi^{(0)}(\theta+1) - \theta\psi^{(0)}(\theta) = 1$. This is because the Gamma function verifies the following condition: $\Gamma(\theta+1) = \theta\Gamma(\theta)$, which itself implies: $\psi^{(0)}(\theta+1) = (\mathrm{d}/\mathrm{d}\theta)\ln(\Gamma(\theta+1)) = (\mathrm{d}/\mathrm{d}\theta)\ln(\theta\Gamma(\theta)) = 1/\theta + \psi^{(0)}(\theta)$. This demonstrates that $\overline{k}_1 = 1$ for this model.

It may also be shown that the variance of $k$ is equal to

$$\mathrm{var}(k)_n = \sum_{k=1}^{n} k^2 P_{\theta,n}(k) - (\overline{k}_n)^2 = \sum_{j=0}^{n-1} \frac{\theta j}{(\theta+j)^2}. \quad (7)$$

Now, defining the log-likelihood function for Model 1 as $L_k(\theta) = \ln P_\theta(k)$ from Equation (4), the maximum likelihood estimate of $\theta$, $\overline{\theta}$, verifies the following equation: $(\mathrm{d}/\mathrm{d}\theta)L_k(\overline{\theta}) = 0$ which turns out to be exactly Equation (5). So Equation (6) can be used to calculate the maximum likelihood estimate of parameter $\theta$ given $n$ and $k$ in a sample. This approach also gives access to the variance of $\theta$, $\sigma_\theta^2$, through the formula $\mathrm{d}^2L(\theta)/\mathrm{d}\theta^2 = -1/\sigma_\theta^2$:

$$\frac{1}{\sigma_\theta^2} = \frac{\overline{k}_n}{\overline{\theta}^2} - \sum_{j=0}^{n-1} \frac{1}{(\overline{\theta}+j)^2}. \quad (8)$$

In the literature on species diversity estimation, the estimated number of species in a sample $\overline{k}_n$ has been explored in the limit of a large sample sizes $n$. Using the first order approximation of the digamma function $\psi^{(0)}(x) \sim \ln(x)$ for $x$ large, and Equation (6), it is apparent that:

$$\overline{k}_n \sim \theta\ln\left(1+\frac{n}{\theta}\right). \quad (9)$$

Thus the expected number of species $\overline{k}_n$ in a sample increases roughly as the logarithm of sample size for large samples, a scaling relationship first proposed in the form of (9) by Fisher (1943) [44]. Note that Fisher (1943) used the notation $\theta = \alpha$, which was later called Fisher's $\alpha$ in the biological diversity literature [45,46]. The depth of Fisher's intuition concerning this model has already been pointed out by Watterson [47] in the context of population genetics, see also Tavaré's recent review [14].

Finally, the following result turns out to be useful. A second generative process, the residual allocation model, has been shown to converge to the Ewens sampling formula. This construction has been popularized under the name Griffiths–Engen–McCloskey model by Ewens [8], in light of the pioneering works of [39, 48]. Let $\{W_1, W_2, \ldots, W_k\}$ be a vector of independently identically distributed random numbers drawn from the beta probability distribution with parameters $(1,\theta)$: $\mathrm{Beta}(1,\theta) = \theta(1-W)^{\theta-1}$. Let us define the variables $\{V_1, V_2, \ldots, V_k\}$ as follows:

$$V_1 = W_1, V_k = W_k \prod_{i=1}^{k-1}(1-W_i). \quad (10)$$

This model has an intuitive interpretation as a broken stick model, and it has been explored in the species abundance literature [39, 49, 50]: a random fraction $W_1$ of a stick of unit length is labeled 1. The random fraction $W_2$ of the unlabeled portion of the stick, of length $1-W_1$, is then labeled with species 2, with a length $W_2(1-W_1)$, and so forth. The sequence $\{V_1, V_2, \ldots, V_k\}$ can be used to generate a multinomial sample $\{n_1, n_2, \ldots, n_k\}$ with weights $\{V_1, V_2, \ldots, V_k\}$ it was shown that this construction verifies the Ewens sampling formula Equation (2) [11]. The Griffiths–Engen–McCloskey is extremely helpful computationally because it allows to generate a partition structure of $n$ objects, with $n$ possibly very large, while drawing only on the order of $k$ random samples, with typically $k \ll n$.

### 2.3. *Neutral model with dispersal limitation (model 2)*

Spatial extensions of Model 1 have been developed early on in the context of population genetics [30, 41, 51], with subsequent applications in ecology and biogeography [7, 52]. One possible framework is as follows: a region is considered as a collection of $K$ local sites ("demes" in the parlance of population genetics), and each local site has a total size $\{N_1, \ldots, N_K\}$. Within a local site, individuals interact directly, whereas local sites are only connected through migration. This framework is the multideme model of population genetics [30, 53] and the metapopulation model of population dynamics [52]. Within-site processes (local reproduction,

interactions) are the dominant processes over small time scales compared with the genealogical processes occurring across local sites and over much longer time scale [30]. A similar setting arises in spatial interacting particle systems, such as chemical reactions [10,54].

An historically important special case, due to Hubbell [7], is one where a single site is sampled. Conceptually, this model is parameterized by the same parameter $\theta$ as above, which describes the regional species pool. An additional parameter $m$ ($0 \leq m \leq 1$) represents the probability that a new individual appears in the focal community through immigration rather than due to local reproduction. If $m \sim 1$, all the local recruits are immigrants, and the local structure becomes irrelevant, in which case the Ewens sampling formula Equation (2) applies, together with the results of the previous section. In the more general case of an arbitrary parameter $m$, a closed-form solution of the general sampling formula does exist and it generalizes Equation (2) [32, 40]. With $I = m/(1-m)(n-1)$ a rescaled migration parameter, the generalized version of the sampling formula is (see Equation (6) in [32]):

$$p_{\theta,I}(n_1, n_2, \ldots, n_k) = \frac{\theta^k}{\theta^{(n)}} \frac{n!}{k! \prod_{j=1}^{k} n_j} \sum_{j=k}^{n} \left( K(j) \frac{\theta^{(n)}}{\theta^{(j)}} \frac{I^j}{I^{(n)}} \right). \quad (11)$$

This sampling formula involves a summation term and series of numbers $K(j)$ which are defined as the coefficients of the polynomial

$$\sum_{j=k}^{n} K(j) x^j = \prod_{i=1}^{k} \sum_{a_i=1}^{n_i} \frac{S(n_i, a_i) S(a_i, 1)}{S(n_i, 1)} x^{a_i} \quad (12)$$

where $S(n,a)$ are again the unsigned Stirling numbers of the first kind (Equation (1)). However, for relatively large values of $n$, an exact calculation of the coefficients $K(j)$ is difficult (but see [33,55]), and is impossible for very large sample sizes. For this reason Hubbell's dispersal-limited neutral model is of limited use in many practical cases.

Let us now turn to Model 2. One far more interesting model of spatially subdivided populations is the $K$-deme model, with $K$ local samples (or demes), of size $\{N_1, \ldots, N_K\}$. Assume that the regional relative species abundance distribution is given by $\{x_1, \ldots, x_k\}$, where $x_i$ is the regional relative abundance of species $i$, and $\sum_i x_i = 1$. Denote $\{n_{1j}, \ldots, n_{kj}\}$ the species abundances in deme $j$, such that $\sum_i n_{ij} = N_j$. Finally, assume that each local deme $j$ has an immigration rate $m_j$, with $m_j \in [0,1]$ (or equivalently the rescaled immigration rate $I_j = m_j/(1-m_j)(N_j - 1)$). This model could describe an archipelago with $K$ islands, some far away from the continent ($m \ll 1$) and others closer, as in the insular theory of biogeography [56]. In this case, the sampling formula $p_{\mathbf{x},\mathbf{I}}(n_{ij})$, $i \in \{1, \ldots, k\}$, $j \in \{1, \ldots, K\}$ can be written as [33]:

$$p_{\mathbf{x},\mathbf{I}}(n_{i,j}) = \prod_{j=1}^{K} \frac{N_j!}{\prod_{i=1}^{k} n_{ij}!} \frac{\prod_{i=1}^{k} (I_j x_i)^{(n_{ij})}}{I_j^{(N_j)}}. \quad (13)$$

In Model 2, the regional species abundance $\mathbf{x}$ is assumed known, rather than resulting from a neutral regional dynamics as in Model 1, so the parameter $\theta$ is absent. It may also be assumed that the local species abundances $n_{ij}$ are a representative and unbiased sample of the regional species pool, implying that the vector $\mathbf{x}$ can be approximated by $x_i = \sum_{j=1}^{K} n_{ij} / \sum_{j=1}^{K} N_j$ [33]. A less straightforward alternative consists in assuming that the region follows Model 1, compute $\theta$ and deduce $\mathbf{x}$ [57]. From Equation (13), the likelihood function for this problem is:

$$L_{\mathbf{x},n_{i,j}}(\mathbf{I}) = C + \sum_{j=1}^{K} \left( \sum_{i=1}^{k} \ln \frac{\Gamma(I_j x_i + n_{ij})}{\Gamma(I_j x_i)} - \ln \frac{\Gamma(I_j + N_j)}{\Gamma(I_j)} \right) \quad (14)$$

with $C$ a constant, and using again the equality $x^{(n)} = \Gamma(x+n)/\Gamma(x)$.

Note also that $p_{\mathbf{x},\mathbf{I}}(n_{i,j})$ is the product of the probabilities for each of the $K$ demes. The maximum likelihood estimate of $I_j$ is obtained for the condition: $\forall j$, $\partial L_{\mathbf{x},n_{i,j}}(\mathbf{I})/\partial I_j = 0$, and this implies that the migration parameter $I_j$ can be estimated independently of all other model parameters through the following equation, for all $j$:

$$\sum_{i=1}^{k} x_i [\psi^{(0)}(I_j x_i + n_{ij}) - \psi^{(0)}(I_j x_i)] = \psi^{(0)}(I_j + N_j) - \psi^{(0)}(I_j) \quad (15)$$

$\psi^{(0)}(\theta) = (\mathrm{d}/\mathrm{d}\theta) \ln(\Gamma(\theta))$. From Equation (15), or by maximization of Equation (14), the parameters $I_j$ of Model 2 can be simply inferred at each site $j$.

### 2.4. *Neutral "diversity begets diversity" model (model 3)*

A second natural extension of Model 1 is one where the probability of creating new species increases with

the number of species in the sample. The idea of this model is that species entering a community generate the conditions for the establishment of more species than originally possible. Formally, the rate of species appearance is not strictly equal to $\theta$ but increases with the number of extant species $k$ as $\theta + \sigma k$, where $\sigma$ is a new parameter in the model compared with Model 1.

As outlined above, let us first represent this model as an urn scheme. After $n-1$ individuals have been sampled, the probability of adding one individual to species $i$ (i.e., sampling a colored ball) is equal to $(n_i - \sigma)/(\theta + n)$, while the probability of sampling an altogether new species (i.e., sampling the black ball) is equal to $(\theta + \sigma k)/(\theta + n)$. In the special case $\sigma = 0$, this model is equivalent to the Hoppe urn model (Model 1). Values $1 \geq \sigma \geq 0$ imply that rare species tend to be picked less, and that more new species arise. As $\sigma \to 1$, the probability of picking a singleton species vanishes, and at $\sigma = 1$, species cannot increase in abundance and each new individual represents a different species. This model was introduced by Blackwell and MacQueen [34] in the early 1970s, then was formally studied by Pitman and colleagues [9, 58, 59].

In Model 3, the expected number of species $k$ is given by the summed probability of picking the black ball at each step:

$$\begin{aligned}\overline{k}_{n+1} &= \sum_{j=0}^{n} \frac{\theta + \overline{k}_j \sigma}{\theta + j} = \overline{k}_n + \frac{\theta + \overline{k}_n \sigma}{\theta + n} \\ &= \frac{\theta}{\theta + n} + \left(1 + \frac{\sigma}{\theta + n}\right) \overline{k}_n. \end{aligned} \tag{16}$$

This equation has an exact solution, with the boundary condition $\overline{k}_1 = 1$:

$$\sigma \overline{k}_n = \frac{\Gamma(\theta + 1)}{\Gamma(\theta + \sigma)} \frac{\Gamma(n + \theta + \sigma)}{\Gamma(n + \theta)} - \theta. \tag{17}$$

The validity of the boundary condition $\overline{k}_1 = 1$ is verified immediately from the equality: $\Gamma(1 + \theta + \sigma) = (\theta + \sigma)\Gamma(\theta + \sigma)$. The asymptotic regime at large sample size $n$ is obtained with the Stirling formula $\Gamma(x) \approx \sqrt{2\pi} x^{x-1/2} \mathrm{e}^{-x}$, valid for large $x$ and applied on Equation (17):

$$\overline{k}_n \approx \frac{\Gamma(\theta + 1)\mathrm{e}^{-\sigma}}{\Gamma(\theta + \sigma)\sigma} n^{\sigma} - \frac{\theta}{\sigma}. \tag{18}$$

This complements the asymptotic regime of Equation (9) in the case $\sigma > 0$. Interestingly, the power-law species accumulation curve emerges from this simple generalization of Model 1. The discussion of whether the species accumulation curve should follow a logarithmic or a power-law shape has been much discussed in the ecological literature for at least a century [60].

Let us now turn to the existence of a sampling formula for Model 3. Pitman has shown that a sampling formula analogous to that of Ewens can also be derived in this case [58] and that it has the following form:

$$\begin{aligned} &p_{\theta,\sigma}(n_1, n_2, \ldots, n_k) \\ &\quad = \frac{\theta(\theta + \sigma) \cdots (\theta + (k-1)\sigma)}{\theta^{(n)}} \prod_{j=1}^{k} (1 - \sigma)^{(n_j - 1)}. \end{aligned} \tag{19}$$

This formula is called the two-parameter Pitman sampling formula [9, 11, 58, 59]. Noticing that the numerator in the above equation can be rewritten $\sigma^k (\theta/\sigma)^{(k)}$, it follows that:

$$p_{\theta,\sigma}(n_1, n_2, \ldots, n_k) = \frac{\sigma^k (\theta/\sigma)^{(k)}}{\theta^{(n)}} \prod_{j=1}^{k} (1 - \sigma)^{(n_j - 1)}. \tag{20}$$

Using again the fact that the increasing factorial obeys the following relationship: $x^{(n)} = \Gamma(x+n)/\Gamma(x)$, Equation (19) can be rewritten in terms of the Gamma function:

$$\begin{aligned} &p_{\theta,\sigma}(n_1, n_2, \ldots, n_k) \\ &\quad = \sigma^k \frac{\Gamma(\theta)}{\Gamma(\theta + n)} \frac{\Gamma(\theta/\sigma + k)}{\Gamma(\theta/\sigma)} \prod_{j=1}^{k} \frac{\Gamma(n_j - \sigma)}{\Gamma(1 - \sigma)}. \end{aligned} \tag{21}$$

In empirical species abundance studies, one objective is to infer the values of model parameters $\theta$, $\sigma$ given the observed vector $n_1, n_2, \ldots, n_k$. In the case of the Ewens sampling formula, the number of species $k$ and the sampling size $n$ are jointly sufficient to estimate the parameter $\theta$. It is no longer the case in this two-parameter Model 3. However, it remains true that Equation (21) can be used to define a log-likelihood function $L_n(\theta, \sigma) = \ln p_{\theta,\sigma}(n_1, n_2, \ldots, n_k)$, which takes the form:

$$\begin{aligned} &L_{\mathbf{n}}(\theta, \sigma) \\ &\quad = \ln \left[ \frac{\sigma^k}{\Gamma(1 - \sigma)^k} \frac{\Gamma(\theta)}{\Gamma(\theta + n)} \frac{\Gamma(\theta/\sigma + k)}{\Gamma(\theta/\sigma)} \prod_{j=1}^{k} \Gamma(n_j - \sigma) \right]. \end{aligned} \tag{22}$$

The values of $\theta$, $\sigma$ such that the partial derivatives of $L_n(\theta, \sigma)$ vanish yield the necessary conditions for the existence of maximum likelihood estimates $\overline{\theta}$, $\overline{\sigma}$:

$$\frac{\partial L_{\mathbf{n}}}{\partial \theta}(\overline{\theta}, \overline{\sigma}) = 0, \quad \frac{\partial L_{\mathbf{n}}}{\partial \sigma}(\overline{\theta}, \overline{\sigma}) = 0. \tag{23}$$

Finding best-fit parameters $\theta$, $\sigma$ can be obtained by solving these two equations jointly, but it is simpler to maximize the function $L_{\mathbf{n}}(\theta,\sigma)$ (Equation (22)). The numerical problem of finding $\theta$, $\sigma$ given the vector $\mathbf{n}$ is therefore easily resolved (see next section for a numerical implementation).

Importantly, species abundance distributions for Model 3 can be generated through a random allocation model (Griffiths–Engen–McCloskey construction) similar to that in Model 1. Let $\{W_1, W_2, \ldots, W_k\}$ be a vector of i.i.d. random draws with $W_i$ from the probability distribution Beta$(1-\sigma, \theta + k\sigma)$, with $0 \le \sigma \le 1$. Define the variables $\{V_1, V_2, \ldots, V_k\}$ as follows:

$$V_1 = W_1, \quad V_k = W_k \prod_{i=1}^{k-1} (1 - W_i). \tag{24}$$

This model generates variables $\{V_1, V_2, \ldots, V_k\}$. A sample of $n$ individuals from a multinomial distribution with weights $\{V_1, V_2, \ldots, V_k\}$ is denoted $\{n_1, n_2, \ldots, n_k\}$ and it was shown [11, 59] that this sample verifies the Pitman sampling formula Equation (19).

### 2.5. *Numerical analyses*

To perform empirical analyses, I have written the package `neutr` in the R Statistical Language [61], available at https://github.com/jeromechave/neutr.

Parameters can be fit against the empirically observed species abundance data. For Model 1, function `optim.ewens()` optimizes Equation (2). It has the empirical species abundance as an argument and returns parameter $\theta$ and the maximal log-likelihood value. For Model 2, the function `optim.multideme()` takes as argument a matrix with entry the abundance $n_{ij}$ (species $i$ in deme $j$) and it optimizes Equation (13). This function returns a vector of parameters $I_j = m_j/(1-m_j)(n_j-1)$, one per deme, and $m_j$. Importantly, the `optim.multideme()` function assumes that the regional species abundance distribution is the sum of all local species abundances. For Model 3, the function `optim.pitman()` takes the empirical species abundance as an argument, plus initial values of $\theta$, $\sigma$ and returns the best-fit parameters $\theta$, $\sigma$ and the maximal log-likelihood value based on Equation (22).

Another set of functions return typical species abundance distribution generated by Models 1 and 3. Package `neutr` includes the function `generate.hoppe.urn0()` that generates a species abundance distribution according to the Hoppe urn model (Model 1). It takes parameter value $\theta$ and sampling size $n$ as an argument, and returns one possible species abundance distribution and the species number $\overline{k}_n$ inferred from Equation (6). There are at least two ways to code this process. Drawing balls one at a time results in a relatively inefficient procedure (but function `generate.hoppe.urn0()` does this in an efficient way). The alternative is to generate random variables according to the residual allocation model described in Equation (10), and to perform a single multinomial sampling of $n$ individuals with weights $\{V_1, V_2, \ldots, V_k\}$. This second ultrafast procedure is coded in function `generate.hoppe.urn()`. Package `neutr` also includes the function `generate.pitman.urn()` that generates a species abundance distribution according to the Pitman model (Model 3). These two last functions have been coded to run with sample sizes of more than $10^{12}$.

The three models can also be compared: Model 1 is nested within both Models 2 and 3, but the number of $k_p$ parameters vary ($k_p = 1$ for Model 1, $k_p = K$ for Model 2, and $k_p = 2$ for Model 3). It is possible to compare the models based on some form of the Akaike Information Criterion [62].

Package `neutr` bears resemblance with package `untb` [63], which implements Model 1, but not its Griffiths–Engen–McCloskey construction. Also, packages `ecolottery` [64] and package GUILDS [55] both implement Models 1 and some forms of Model 2 based on the coalescent, as first proposed by [40]. To my knowledge, Model 3 and its Griffiths–Engen–McCloskey construction have never been implemented in a R package.

## 3. Application to the tropical tree flora

### 3.1. *Datasets*

An empirical application of the above theory is now provided for three large empirical data sets taken from numerous tropical forest inventories around the world and reproduced as a Supplementary Information of [6]. It contains a total of over a million sampled trees all identified to the species, for a total area of forest sampled of 2324 ha (23.24 km$^2$; summary

in Table 1). This is a huge sampling size, although it is very small compared with the ca. 10.7 million km$^2$ of tropical forests [65]. Since the exact species name is not reported in this data set, it is impossible to estimate the exact number of species in total, although the species overlap across continents is small, and the species total is likely to be close to the sum (9765 species). The raw data is plotted as a rank-abundance curve in Figure 1.

Briefly, the data have been obtained using a standard method in tropical forest science: standard areas, usually squares of one hectare (100 × 100 m), are positioned on the ground, and all trees with at least 10 cm in trunk diameter are mapped, tagged using a permanent tag (in plastic or metal), and identified to the species [66]. Establishing a permanent forest inventory plot requires several days of work in the field, and complete botanical identification is usually much more time consuming. The above data set is therefore the result of a long term vision, and hard work of a large scientific community from across the tropics.

### 3.2. *Results*

For Model 1, the estimate of parameter $\theta$ for all three data sets is provided using the empirical values of $n$ and $k$ with Equation (6). I used the empirical values of $\theta$ and Equation (10) to produce 1000 neutral distributions for each of the three empirical distributions. Figure 1 reveals that the fit to the data is not bad, except for a small number of the most abundant species (see [67] for a similar pattern). It is therefore interesting to explore if the goodness of fit improves when the top species are removed.

To estimate the goodness of fit, one method is to define a distance between the observed $P_{\mathrm{obs}}$ and the theoretical $P_{\mathrm{theor}}$ distributions. I use the Kullback–Leibler distance, defined by $\mathrm{KL} = \sum_{i=1}^{k1} P_{\mathrm{theor}} \ln(P_{\mathrm{theor}}/P_{\mathrm{obs}})$, with $k1$ the minimum of the non-null values of both observed and theoretical distributions. Successively removing $1,2,\ldots$ of the top species, a new value of $\theta$ was computed, and the Kullback–Leibler distance was calculated. Figure 2 shows the shape of the Kullback–Leibler distance against successive removals of top species. The removal of 13, 5, and 40 top species, for the Amazon, tropical Africa and Southeast Asia respectively, resulted in a massive improvement of the model's goodness of fit as see in Figure 2 and in Table 2. Removing the ultradominant species results in a much improved fit (compare Figures 1 and 3), even though the neutral model tends to underestimate slightly the abundance of the rare species (right panel of Figure 3).

The multideme model (Model 2) can be fitted with the maximization of Equation (15). The distribution of $m$ values, which represent the fraction of individuals drawn from the regional pool rather than from the local site, peaks at $m = 0.122$ for Amazonia, $m = 0.094$ for Africa and $m = 0.127$ for Southeast Asia (Figure 4). This shows that local sites are dispersal-limited in similar ways across continents on average. Sites-specific parameters $m$ could be interpreted as an environmental filtering effect, since $m$ measures how dissimilar the local assemblage is from the regional one [33].

A fit of the data set against the two-parameter Model 3 is illustrated in Figure 5. The fit is not improved for the Amazonia and tropical Africa data ($\sigma$ values $<10^{-7}$), and barely so for Southeast Asia ($\sigma = 0.034$). Thus, for these data sets, Model 3 does not result in a better fit of the data than Model 1.

The neutral model represents well tropical tree species abundances at regional scale, and this finding is used to extrapolate species numbers [5]. Assuming that the Amazonian tropical forest covers about 6.3 million km$^2$, with an average 500 trees per hectare, the estimated number of trees is $N \approx 3.15 \times 10^{11}$. Using the value of $\theta$ reported in Table 2 and $N$ in Equation (6) yields $\overline{k}_N$ = 13,081 species. Using the improved estimate of $\theta$ after the removal of ultra-dominant species yields $\overline{k}_N$ = 13,440 species. The values for the two other continents are reported in Table 3. It is also possible to estimate the number of species with at least 50 individuals overall (a lower bound of the minimum viable population size [68]) to avoid the pitfall of predicting the occurrence of species represented by a handful of individuals in a region [35] (Table 3). For Amazonia, the resulting number is $\overline{k}_{N,n>50}$ = 10,141 species, very close to the latest estimate of 10,071 species reported in [69].

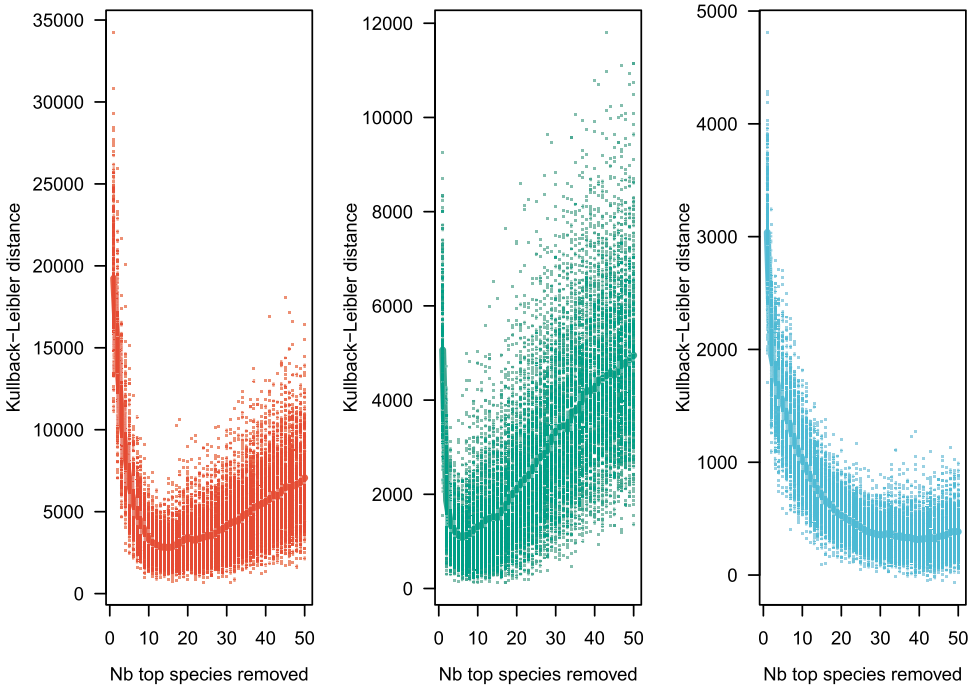


**Figure 2.** Kullback–Leibler distance between observed and simulated distributions replicated 1000 times, after sequentially removing 1 to 50 of the most abundant species. The minimal Kullback–Leibler distance, number of species to remove to minimize the distance, and $\theta$ are reported in Table 2. Color codes as in Figure 1.

**Table 1.** Statistics of the data used in this study

| Region | Nb trees | Nb species | Cumul. area (ha) | Nb plots |
|---|---|---|---|---|
| Amazon | 821,670 | 4670 | 1590 | 1417 |
| Tropical Africa | 210,313 | 1509 | 504 | 483 |
| Southeast Asia | 100,152 | 3586 | 201 | 230 |
| Total | 1,021,974 | NA | 2324 | 2130 |

The table reports the total number of trees sampled, total number of species, cumulative sampled area (in hectares, ha), and total number of permanent sampling plots in the biome. Data from Ref. [6].

## 4. Discussion

### 4.1. *Three neutral models*

The three models presented here are only a few examples of the many systems that can be framed as urn models [24]. The reader may be surprised to read no mention of the coalescent theory, even if Model 1 is the key building block of this theory [12, 70]. The coalescent is a powerful approach, but the choice was made here to focus solely on species abundance distribution and efficient numerical analyses can be performed without resorting to coalescent models. There is no doubt that exploring further applications of the coalescent in ecology should be a rewarding effort.

The three models presented here are neutral in the following sense. All three describe a partition of the collection into disjoint subsets (classes), such that the objects within each class are interchangeable, and the abundance $n_i \geq 1$ within class $i$ is

**Table 2.** Best fit of the canonical neutral model after removing a number of the top species (see Figure 2 for an illustration)

| Region | Initial $\theta$ | Min. Kullback–Leibler | Nb. of ultradominant species | $\theta$ of best model |
|---|---|---|---|---|
| Amazon | 654.3 | 2748.6 | 13 | 673.19 |
| Tropical Africa | 219.7 | 1077.0 | 6 | 226.69 |
| Southeast Asia | 726.8 | 299.9 | 40 | 778.23 |

Empirical values of parameter $\theta$ are reported before (first column) and after (last column) the ultradominant species have been removed. Column "Min. Kullback–Leibler" reports the mean value of the minimal Kullback–Leibler distance between model and observations (average across 1000 values).

**Table 3.** Estimated number of species in tropical forests based on an extrapolation of Model 1

| Region | Area (M km$^2$) | Nb trees (estimated) | Nb species | Nb species with at least 50 ind. |
|---|---|---|---|---|
| Amazon | 6.3 | $3.15 \times 10^{11}$ | 13,081 | 10,141 ± 91 |
| Tropical Africa | 2.9 | $1.45 \times 10^{11}$ | 4,462 | 3,477 ± 51 |
| Southeast Asia | 1.1 | $0.55 \times 10^{11}$ | 13,186 | 9,915 ± 90 |

The estimated number of species is provided together with the estimated number of species with at least 50 individuals.

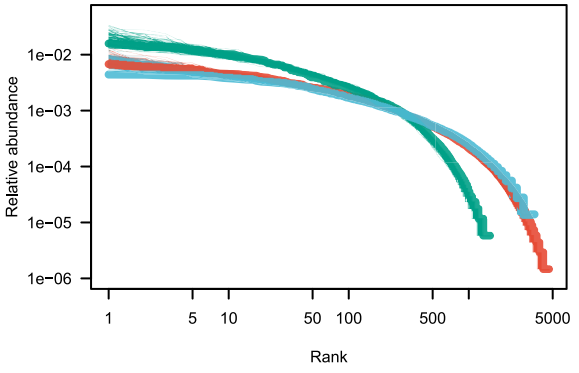


**Figure 3.** Fit of the rank abundance distributions after removing the ultradominant species (see Table 2). Color codes are as in Figure 1.

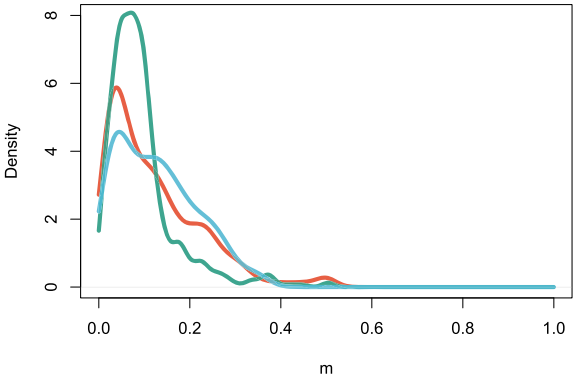


**Figure 4.** Estimation of the local $m$ parameters for all the sites in the three continents. The figure reports the density distribution of $m$ values. Color codes are as in Figure 1.

a random number that fully describes the class. The models are thus fully described by the probability distribution $p(n_1,\ldots,n_k)$ of the sequence of random numbers $\{n_1,\ldots,n_k\}$, where $\sum_{i=1}^{k} = n$, and this probability distribution function is invariant under any permutation $\eta$ of the class labels: $p(n_1,\ldots,n_k) = p(n_{\eta(1)},\ldots,n_{\eta(k)})$. This last property is called *exchangeability*, and the probability distribution $p(n_1,\ldots,n_k)$ is then called *exchangeable* [71]. It turns out that there is a formal equivalence between the statement (1) the random partition has the property of exchangeability and (2) the property that Models 1 and 3 can be constructed as urn processes and as random allocation processes (Proposition 9 in [59]). This is an important result because it provides a rigorous definition of the concept of class equivalence in neutral models in ecology and population genetics.

Another property common to exchangeable models is that a random partition following Equations (10) and (24) define a size-biased partition of

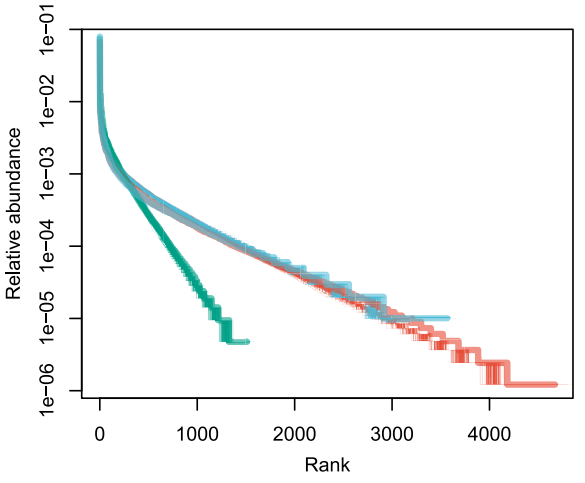


**Figure 5.** Fit of the rank abundance distributions (thick solid lines) against Model 3 ($n$ = 100 simulations, thin solid lines), in linear–log axes, which allow a better display of the tail of the distribution (rare species). Compare with Figure 1 for a similar representation in log–log axes. Color codes are as in Figure 1.

$\{V_1, V_2, \ldots, V_k\}$ that can be used to define an ordered series of relative abundances $P_i$, $i \in \{1, \ldots, k\}$, with $P_1 \geq P_2 \geq \cdots \geq P_k$, such that $\sum_{i=1}^{k} P_i = 1$. The probability distribution of this collection is unchanged upon the removal of the first species $P_1$ and normalization $P_i' = P_i/(1-P_1)$, $i > 1$, since the new sequence has exactly the same formal structure. This provides the opportunity to generate a test of neutrality by sequentially removing the first species, the second species, and so forth, until the empirical data best fits the model. This idea provides an intuitive method to define a notion related to that of species hyperdominance in a species assemblage [5]. Here, a concept related to hyperdominance, ultradominance, is defined: a top species is ultradominant if is removal significantly improves the fit of the neutral model. More precisely, ultradominant species are the first $U$ species such that the series $P_{U+1} \geq P_{U+2} \geq \cdots \geq P_k$ minimizes the distance between the neutral model fit and the empirical observations (cf. [67] for a related discussion). This definition is relative to a choice of distance on probability distributions, and also on the choice of the neutral model (Model 1, 2, 3 or another variant). As shown in the Section 3.2 (Figure 2), the number of ultradominant species is usually far lower than that of hyperdominant species.

The one-parameter ($\theta$) Model 1 is a special case of a more general two-parameter ($\theta, \sigma$) Model 3. When $\sigma = 0$, Models 1 and 3 are equal. Model 3 turns out to have a biological interpretation as a species abundance model where diversity begets diversity: new species tend to be picked with probability $(\theta + k\sigma)/(\theta + n)$ $(0 \leq \sigma \leq 1)$, when $k$ species are already present, so more often than expected by chance. With respect to the asymptotic scaling regime $n \gg 1$, Equations (9) and (18) are two well-known scaling relationships in ecology and there has been much literature on whether the species accumulation curve $\overline{k}_n$ should follow a logarithmic shape $\overline{k}_n \sim \theta \ln(n)$ (as proposed by Fisher 1943) or a power-law shape $\overline{k}_n \sim n^{\sigma}$. Models 1 and 3 show that the two regimes are consistent with a single representation of a model of exchangeable random partitions. While Model 3 did not provide a better fit of empirical data than Model 1 for tropical tree species, it is possible that species assemblages at higher trophic levels are more likely to verify the conditions of Model 3.

The dispersal-limited neutral model of [7] is historically important in the context of ecology and biogeography. However, Hubbell's two-parameter ($\theta, m$) generalization of Model 1 is not framed as an urn model or a random allocation model. An urn representation of Hubbell's model is possible, but it is not straightforward[2]. This is a special form of a hierarchical urn construction, such that the construction of the second urn depends on that of the first urn, but not the other way around. Here, I define Model 2 as a neutral model of $K$ subdivided assemblages, or multi-deme model, a more useful alternative in practical situations. Though only one result is presented and studied here, many other results have been obtained, and it is an area where coalescent based numerical analyses are most useful [12, 51].

[2]Define two urns, the first with a black ball with weight $\theta$, the second empty. First, draw $J$ balls from the first urn exactly as in the Hoppe urn scheme, generating the sequence of ball colors: $n_1^{[1]}, \ldots, n_k^{[1]}$, with $n^{[1]} = J = \sum_{i=1}^{k} n_i^{[1]}$. Here the notation "$^{[1]}$" represents the first urn. Then turn to the second urn. At step $n$, either (1) pick one ball from the *first* urn with probability $I/(I+n)$; this ball is of color $i$ with probability $n_i^{[1]}/J$; add a new ball of the same color $i$ in the *second* urn (so: $n_i^{[2]} \to n_i^{[2]} + 1$), or (2) with probability $n/(I+n)$, pick a ball in the second urn, and add one ball of the same color. In the second urn, the total number of $n^{[2]} = n = \sum_{i=1}^{k} n_i^{[2]}$ (usually, $n \ll J$).

### 4.2. *Implications for tropical forests*

A remarkable feature is that neutral models fit extremely well tree species abundance data [5, 7]. The only departure from this neutral fit appears to be for the most abundant species, which I have here informally called ultradominant species. Ultra-dominant species are the most abundant species that tend to depart from the prediction of the canonical neutral model (Model 1), and a simple empirical method is proposed here to detect these species. When ultra-dominant species are removed from the analysis, the neutral model reproduces empirical observations extremely well in the three regional tropical tree data sets explored here (Figure 2). The nature of ultra-dominant species is unclear, but it would be interesting to explore the commonalities of ultra-dominant species, and possible biological explanations for their occurrence.

A second insight from this analysis stems from the fit of the multi-deme model (Model 2) to the same three regional tropical tree data sets. The finding of Figure 3 is that most local sites significantly depart from a hypothesized random sample of the regional species pool, detected by $m$ values much lower than unity. In a dispersal-limited interpretation [7], this suggests that a relatively constant proportion of the reproduction events are local, the rest being immigration events. The alternative explanation is that biotic or abiotic effects exert a major influence on the floristic composition of each local community, and $m < 1$ values in the multi-deme model reflect these environmental filtering effects [33].

In the Ewens canonical model, $\theta$ is a natural measure of local diversity, which is a convenient property of the model. However, non-parameteric methods of biodiversity estimation have also been developed, and they are often presented as more robust than parametric methods [72, 73]. It is not the goal of the present contribution to discuss this issue at length. One known limitation of Model 1 is that the $\theta$ parameter cannot be estimated with confidence for small sample sizes. A method such as Model 2 may provide a useful alternative to Model 1 and it would be interesting to explore the scale dependence of the $m$ parameter (i.e., how $m$ varies with decreasing sample sizes $n$).

The notion of neutrality has generated a heated debate in ecology [7,20,74–76]. Part of the debate has been motivated by a narrow definition of neutrality. The questions raised relate to the fact that this model ignores so many important features as to become useless or event dangerous [77]. Species differ not only in abundance but also in ecological traits, and individuals within species also differ in size, metabolic capacity, and fitness. A second class of critiques of ecological neutrality is that tests of the theory are not robust because independent estimates of the parameters cannot be accessed [74]. A last class of critiques relate to the consistency of the ecological neutral model with respect to the dominant theory in ecology, niche theory, which interprets patterns of species occurrence in the light of competition between species [78]. Sampling methods that make the assumption of neutrality (in the sense of exchangeability), are however more general than commonly discussed in the ecology literature. One strength of these models presented here is that they are naturally associated with a method of exact inference, which makes it possible to compare model and data quantitatively. By analogy, the analysis of selectively neutral alleles in natural populations is a useful tool set of population genetics, and helps infer past population sizes, and phases of demographic expansions and bottlenecks [8, 79].

In ecology, the development of the neutral mathematical tool set has not been as rapid as in genetics. One limitation has been that, unlike in genetics, methods for analyzing huge data sets has been relatively less in need in ecology. The situation is changing now with the rapid rise of DNA-based ecology [15], in which samples are not of individuals but of sequences, and with applications in tropical forest research [80]. Provided that sequence abundance data can be interpreted biologically, the question of the structure of sequence abundance distributions can be explored with the same approach as outlined here.

Why should a model as simple as Model 1 provide such an excellent fit to regional tropical tree abundance data? The neutral hypothesis cannot be valid over ecological and evolutionary time scales. Yet, regional species assemblages result from such a myriad of local processes, both biotic and abiotic, that large enough samples of regional patterns average over these effects. Half a century ago, Robert H. Whittaker wrote [17]: “The enigma of the diversity of the tropical rainforest should be expected to open itself

to no single key, and may be enigmatic to the extent we have yet to comprehend the full implications of biotic differentiation and interaction, the complexity [...] that is feasible and has evolved in these forests". To this day, this analysis still holds, and the successes of the neutral theory to replicate broad patterns remains a mystery. Even if the details of species persistence and coexistence may be explained by fundamentally different processes [2], the laws of large numbers provide important insights into key questions on the regional species abundance patterns.

## Declaration of interests

The author does not work for, advise, own shares in, or receive funds from any organization that could benefit from this article, and has declared no affiliations other than his research institution.

## Funding

I gratefully acknowledge funding from "Investissement d'Avenir" grants managed by the Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-25-01; TULIP, ref. ANR-10-LABX-0041; ANAEE-France: ANR-11-INBS-0001), CNES long-term funding, the European Space Agency CCI-BIOMASS and FRM4BIOMASS projects. This research is product of the SynTreeSys group funded by the synthesis center CESAB of the French Foundation for Research on Biodiversity (FRB).

## Supplementary data

Supporting information for this article is available on the journal's website under https://doi.org/10.5802/crbiol.162 or from the author. The package `neutr` in the R Statistical Language [61] is available at https://github.com/jeromechave/neutr.

## References

[1] D. Rabinowitz, "Seven forms of rarity", in *The Biological Aspects of Rare Plant Conservation* (H. Synge, ed.), John Wiley & Sons, Chichester, 1981, p. 205-217.

[2] A. R. Kruckeberg, D. Rabinowitz, "Biological aspects of endemism in higher plants", *Ann. Rev. Ecol. Syst.* **16** (1985), no. 1, p. 447-479.

[3] W. E. Kunin, K. J. Gaston, "The biology of rarity: patterns, causes and consequences", *Trends Ecol. Evol.* **8** (1993), no. 8, p. 298-301.

[4] Q. C. B. Cronk, "The past and present vegetation of St Helena", *J. Biogeogr.* **16** (1989), no. 1, p. 47-64.

[5] H. ter Steege, N. C. A. Pitman, D. Sabatier, C. Baraloto, R. P. Salomao, J. E. Guevara *et al.*, "Hyperdominance in the Amazonian tree flora", *Science* **342** (2013), no. 6156, article no. 1243092.

[6] D. L. M. Cooper, S. L. Lewis, M. J. P. Sullivan, P. I. Prado, H. ter Steege, N. Barbier *et al.*, "Consistent patterns of common species across tropical tree communities", *Nature* **625** (2024), no. 7996, p. 728-734.

[7] S. P. Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography*, Monographs in Population Biology, vol. 32, Princeton University Press, Princeton, NJ, 2001.

[8] W. J. Ewens, *Mathematical Population Genetics: Theoretical Introduction*, vol. 1, Springer, New York, 2004.

[9] J. Pitman, "The two-parameter generalization of Ewens' random partition structure", Tech. Report 345, Department of Statistics, UC Berkeley, 1992.

[10] R. Durrett, "Stochastic spatial models", *SIAM Rev.* **41** (1999), no. 4, p. 677-718.

[11] J. Pitman, in *Combinatorial Stochastic Processes: Ecole d'Eté de Probabilités de Saint-Flour XXXII - 2002* (J. Picard, ed.), Springer Berlin, Heidelberg, 2006.

[12] J. H. Wakeley, *Coalescent Theory: an Introduction*, Roberts & Company Publishers, Greenwood Village, CO, 2009.

[13] H. Crane, "The ubiquitous Ewens sampling formula", *Stat. Sci.* **31** (2016), no. 1, p. 1-19.

[14] S. Tavaré, "The magical Ewens sampling formula", *Bull. Lond. Math. Soc.* **53** (2021), no. 6, p. 1563-1582.

[15] P. Taberlet, A. Bonin, L. Zinger, E. Coissac, *Environmental DNA: For Biodiversity Research and Monitoring*, Oxford University Press, Oxford, 2018.

[16] R. Durrett, S. Levin, "The importance of being discrete (and spatial)", *Theor. Popul. Biol.* **46** (1994), no. 3, p. 363-394.

[17] R. H. Whittaker, "Evolution and measurement of species diversity", *Taxon* **21** (1972), no. 2–3, p. 213-251.

[18] W. C. Allee, O. Park, A. E. Emerson, T. Park, K. P. Schmidt, *Principles of Animal Ecology*, 1st ed., Saunders and Co, Philadelphia and London, 1949.

[19] R. Vane-Wright, "Ecological and behavioural origins of diversity in butterflies", in *Symposia of the Royal Entomological Society of London*, vol. 9, Royal Entomological Society of London, London, p. 56-70.

[20] J. Chave, "Neutral theory and community ecology", *Ecol. Lett.* **7** (2004), no. 3, p. 241-253.

[21] J. F. C. Kingman, "Random discrete distributions", *J. R. Stat. Soc. Ser. B (Methodological)* **1** (1975), p. 1-22.

[22] J. Pitman, "Poisson–Kingman partitions", in *Lecture Notes—Monograph Series*, vol. 40, 2003, p. 1-34.

[23] W. J. Ewens, "The sampling theory of selectively neutral alleles", *Theor. Popul. Biol.* **3** (1972), no. 1, p. 87-112.

[24] N. Johnson, S. Kotz, *Urn Models and their Application; An Approach to Modern Discrete Probability Theory*, John Wiley & Sons, New York, 1977.

[25] M. Mowbray, D. Gollmann, "Electing the doge of Venice: anal-

ysis of a 13th century protocol", in *20th IEEE Computer Security Foundations Symposium (CSF'07)*, IEEE, 2007, p. 295-310.

[26] J. Fernández-Gracia, K. Suchecki, J. J. Ramasco, M. San Miguel, V. M. Eguíluz, "Is the voter model a model for voters?", *Phys. Rev. Lett.* **112** (2014), no. 15, article no. 158701.

[27] E. Ben-Naim, N. W. Hengartner, S. Redner, F. Vazquez, "Randomness in competitions", *J. Stat. Phys.* **151** (2013), no. 3, p. 458-474.

[28] G. Polya, *Patterns of Plausible Inference*, Princeton University Press, Princeton, NJ, 1954.

[29] F. M. Hoppe, "Pólya-like urns and the Ewens' sampling formula", *J. Math. Biol.* **20** (1984), no. 1, p. 91-94.

[30] J. Wakeley, "Nonequilibrium migration in human history", *Genetics* **153** (1999), p. 1863-1871.

[31] J. Wakeley, "The coalescent in an island model of population subdivision with variation among demes", *Theor. Popul. Biol.* **59** (2001), no. 2, p. 133-144.

[32] R. S. Etienne, "A new sampling formula for neutral biodiversity", *Ecol. Lett.* **8** (2005), no. 3, p. 253-260.

[33] F. Jabot, R. S. Etienne, J. Chave, "Reconciling neutral community models and environmental filtering: theory and an empirical test", *Oikos* **117** (2008), no. 9, p. 1308-1320.

[34] D. Blackwell, J. B. MacQueen, "Ferguson distributions via Pólya urn schemes", *Ann. Stat.* **1** (1973), no. 2, p. 353-355.

[35] R. Lande, "Extinction thresholds in demographic models of territorial populations", *Am. Nat.* **130** (1987), no. 4, p. 624-635.

[36] J. M. Rhymer, D. Simberloff, "Extinction by hybridization and introgression", *Ann. Rev. Ecol. Syst.* **27** (1996), no. 1, p. 83-109.

[37] F. Courchamp, T. Clutton-Brock, B. Grenfell, "Inverse density dependence and the Allee effect", *Trends Ecol. Evol.* **14** (1999), no. 10, p. 405-410.

[38] H. A. Simon, "On a class of skew distribution functions", *Biometrika* **42** (1955), no. 3/4, p. 425-440.

[39] S. Engen, "On species frequency models", *Biometrika* **61** (1974), no. 2, p. 263-270.

[40] R. S. Etienne, H. Olff, "A novel genealogical approach to neutral biodiversity theory", *Ecol. Lett.* **7** (2004), no. 3, p. 170-175.

[41] G. Malécot, *Les Mathématiques de l'Hérédité*, Masson, Paris, 1948.

[42] J. Chave, E. G. Leigh Jr, "A spatially explicit neutral model of $\beta$-diversity in tropical forests", *Theor. Popul. Biol.* **62** (2002), no. 2, p. 153-168.

[43] S. Karlin, J. McGregor, "Addendum to a paper of W. Ewens", *Theor. Popul. Biol.* **3** (1972), p. 113-116.

[44] 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. Animal Ecol.* **12** (1943), no. 1, p. 42-58.

[45] E. C. Pielou, *Ecological Diversity*, Wiley & Sons, New York, 1975.

[46] S. Engen, *Stochastic Abundance Models with Emphasis on Biological Communities and Species Diversity*, Monographs on Statistics and Applied Probability, vol. 1, Springer, Berlin and Heidelberg, 1978.

[47] G. Watterson, "The homozygosity test of neutrality", *Genetics* **88** (1978), no. 2, p. 405-417.

[48] J. W. McCloskey, *A model for the distribution of individuals by species in an environment*, PhD thesis, Michigan State University, 1965 (unpublished).

[49] R. H. MacArthur, "On the relative abundance of bird species", *Proc. Natl. Acad. Sci. USA* **43** (1957), no. 3, p. 293-295.

[50] J. E. Cohen, "Alternate derivations of a species-abundance relation", *Am. Nat.* **102** (1968), no. 924, p. 165-172.

[51] F. Rousset, *Genetic Structure and Selection in Subdivided Populations*, vol. 40, Princeton University Press, Princeton, NJ, 2004.

[52] I. Hanski, "Metapopulation dynamics", *Nature* **396** (1998), no. 6706, p. 41-49.

[53] T. Maruyama, "Effective number of alleles in a subdivided population", *Theor. Popul. Biol.* **1** (1970), no. 3, p. 273-306.

[54] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry*, Elsevier, Amsterdam, 1992.

[55] T. Janzen, B. Haegeman, R. S. Etienne, "A sampling formula for ecological communities with multiple dispersal syndromes", *J. Theor. Biol.* **374** (2015), p. 94-106.

[56] R. H. MacArthur, E. O. Wilson, *The Theory of Island Biogeography*, Princeton University Press, Princeton, NJ, 1969.

[57] F. Munoz, P. Couteron, B. Ramesh, R. S. Etienne, "Estimating parameters of neutral communities: from one single large to several small samples", *Ecology* **88** (2007), no. 10, p. 2482-2488.

[58] M. Perman, J. Pitman, M. Yor, "Size-biased sampling of Poisson point processes and excursions", *Probab. Theory Relat. Fields* **92** (1992), no. 1, p. 21-39.

[59] J. Pitman, "Exchangeable and partially exchangeable random partitions", *Probab. Theory Relat. Fields* **102** (1995), no. 2, p. 145-158.

[60] O. Arrhenius, "Species and area", *J. Ecol.* **9** (1921), no. 1, p. 95-99.

[61] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2023, https://www.R-project.org/.

[62] H. Akaike, "A new look at the statistical model identification", *IEEE Trans. Automat. Control* **19** (1974), no. 6, p. 716-723.

[63] R. K. Hankin, "Introducing untb, an R package for simulating ecological drift under the unified neutral theory of biodiversity", *J. Stat. Softw.* **22** (2007), p. 1-15.

[64] F. Munoz, M. Grenié, P. Denelle, A. Taudière, F. Laroche, C. Tucker, C. Violle, "ecolottery: Simulating and assessing community assembly with environmental filtering and neutral dynamics in R", *Methods Ecol. Evol.* **9** (2018), no. 3, p. 693-703.

[65] C. Vancutsem, F. Achard, J.-F. Pekel, G. Vieilledent, S. Carboni, D. Simonetti, J. Gallego, L. E. O. C. Aragão, R. Nasi, "Long-term (1990–2019) monitoring of forest cover changes in the humid tropics", *Sci. Adv.* **7** (2021), no. 10, article no. eabe1603.

[66] C. Blundo, J. Carilla, R. Grau, A. Malizia, L. Malizia, O. Osinaga-Acosta *et al.*, "Taking the pulse of Earth's tropical forests using networks of highly distributed plots", *Biol. Conserv.* **260** (2021), article no. 108849.

[67] F. Jabot, J. Chave, "Inferring the parameters of the neutral theory of biodiversity using phylogenetic information and implications for tropical forests", *Ecol. Lett.* **12** (2009), no. 3, p. 239-248.

[68] R. Frankham, C. J. Bradshaw, B. W. Brook, "Genetics in conservation management: revised recommendations for the 50/500 rules, Red List criteria and population viability analyses", *Biol. Conserv.* **170** (2014), p. 56-63.

[69] H. ter Steege, S. Mota de Oliveira, N. C. Pitman *et al.*, "Towards a dynamic list of Amazonian tree species", *Sci. Rep.* **9** (2019), no. 1, article no. 3501.

[70] J. F. C. Kingman, "The coalescent", *Stoch. Process. Their Appl.* **13** (1982), no. 3, p. 235-248.

[71] J. F. C. Kingman, "Uses of exchangeability", *Ann. Probab.* **6** (1978), no. 2, p. 183-197.

[72] J. Bunge, M. Fitzpatrick, "Estimating the number of species: a review", *J. Am. Stat. Assoc.* **88** (1993), no. 421, p. 364-373.

[73] A. Chao, "Nonparametric estimation of the number of classes in a population", *Scand. J. Stat.* **11** (1984), p. 265-270.

[74] R. E. Ricklefs, "Unified neutral theory of biodiversity: do the numbers add up?", *Ecology* **87** (2006), p. 1424-1431.

[75] D. Alonso, R. Etienne, A. Mckane, "The merits of neutral theory", *Trends Ecol. Evol.* **21** (2006), no. 8, p. 451-457.

[76] R. E. Ricklefs, S. S. Renner, "Global correlations in tropical tree species richness and abundance reject neutrality", *Science* **335** (2012), no. 6067, p. 464-467.

[77] J. S. Clark, "Beyond neutral science", *Trends Ecol. Evol.* **24** (2009), no. 1, p. 8-15.

[78] P. Chesson, "Mechanisms of maintenance of species diversity", *Ann. Rev. Ecol. Syst.* **31** (2000), no. 1, p. 343-366.

[79] H. Li, R. Durbin, "Inference of human population history from individual whole-genome sequences", *Nature* **475** (2011), no. 7357, p. 493-496.

[80] L. Zinger, J. Donald, S. Brosse, M. A. Gonzalez, A. Iribar, C. Leroy *et al.*, "Advances and prospects of environmental DNA in neotropical rainforests", *Adv. Ecol. Res.* **62** (2020), p. 331-373.