Ecology Letters, (2004) **7**: 901–910 doi: 10.1111/j.1461-0248.2004.00640.x

IDEAS AND PERSPECTIVES

Sampling Hubbell's neutral theory of biodiversity

David Alonso[1,2]* **and Alan J. McKane**[3]
[1]*Ecology and Evolutionary Biology, University of Michigan 830 North University Avenue, Ann Arbor MI 48109-1048, USA*
[2]*ICREA-Complex Systems Lab, Universitat Pompeu Fabra, Dr Aiguader 80, 08003 Barcelona, Spain*
[3]*Department of Theoretical Physics, University of Manchester, Manchester M13 9PL, UK*
**Correspondence: E-mail: dalonso@umich.edu*

Abstract

In the context of neutral theories of community ecology, a novel genealogy-based framework has recently furnished an analytic extension of Ewens' sampling multivariate abundance distribution, which also applies to a random sample from a local community. Here, instead of taking a multivariate approach, we further develop the sampling theory of Hubbell's neutral spatially implicit theory and derive simple abundance distributions for a random sample both from a local community and a metacommunity. Our result is given in terms of the average number of species with a given abundance in any randomly extracted sample. Contrary to what has been widely assumed, a random sample from a metacommunity is not fully described by the Fisher log-series, but by a new distribution. This new sample distribution matches the log-series expectation at high biodiversity values ($\theta > 1$) but clearly departs from it for species-poor metacommunities ($\theta < 1$). Our theoretical framework should be helpful in the better assessment of diversity and testing of the neutral theory by using abundance data.

Keywords

Abundance distributions, Hubbell's neutral theory, Poissonian zero-sum multinomial, sampling theory.

Ecology Letters (2004) 7: 901–910

INTRODUCTION

In order to address the fundamental question of how species abundances change in space and time, Hubbell's unified theory builds on the foundation of the original theory of island biogeography (MacArthur & Wilson 1967). The theory is developed on the basis of two principles (Hubbell 2001): zero-sum dynamics, and a per capita ecological equivalence among all individuals of every species in a given trophically defined community, which also involves neutral speciation. In particular, the principle of ecological equivalence has generated intensive discussion among ecologists (Yu *et al.* 1998; Bell 2001; Condit *et al.* 2002; Enquist *et al.* 2002; Clark & McLachlan 2003; Magurran & Henderson 2003; McGill 2003; Kneitel & Chase 2004). However, this controversy has been partly a result of the difficulty in devising clear methods to test it (McGill 2003; Volkov *et al.* 2003).

The discussion about the merits of Hubbell's neutral theory as compared with alternative theories based on the differential ecological adaptation of species (Tilman 1982, 1990; Chesson 2000; Mouquet & Loreau 2002, 2003; Kneitel & Chase 2004) is important both from the theoretical and the applied point of view, because the neutral theory is able to describe the complexity of natural communities in a very concise way – only two numbers are needed to characterize a natural community in a given locality. These are θ, the fundamental biodiversity number (the potential species richness of the community), and m, the immigration parameter (its degree of isolation).

Recently, various analytical approaches to the study of the theory have been formulated (Vallade & Houchmandzadeh 2003; Volkov *et al.* 2003; Etienne & Olff 2004b; McKane *et al.* 2004). In the hope of unifying previous approaches and providing sound methods to test the theory, we calculate the exact sample species abundance distributions (SADs) predicted by the neutral theory on the basis of a general theory of community sampling (Dewdney 1998). In particular, we show clearly that the metacommunity abundance distribution is not well described by a logseries distribution (Fig. 1b). As a consequence, we show that the logseries-based solution of Volkov *et al.* (2003) is good for species-rich communities, but gives incorrect predictions in species-poor communities. Although biodiversity issues are unavoidable in relation to species-rich communities, such as rain forests and coral reefs, some of the world's most extensive and ancient ecosystems

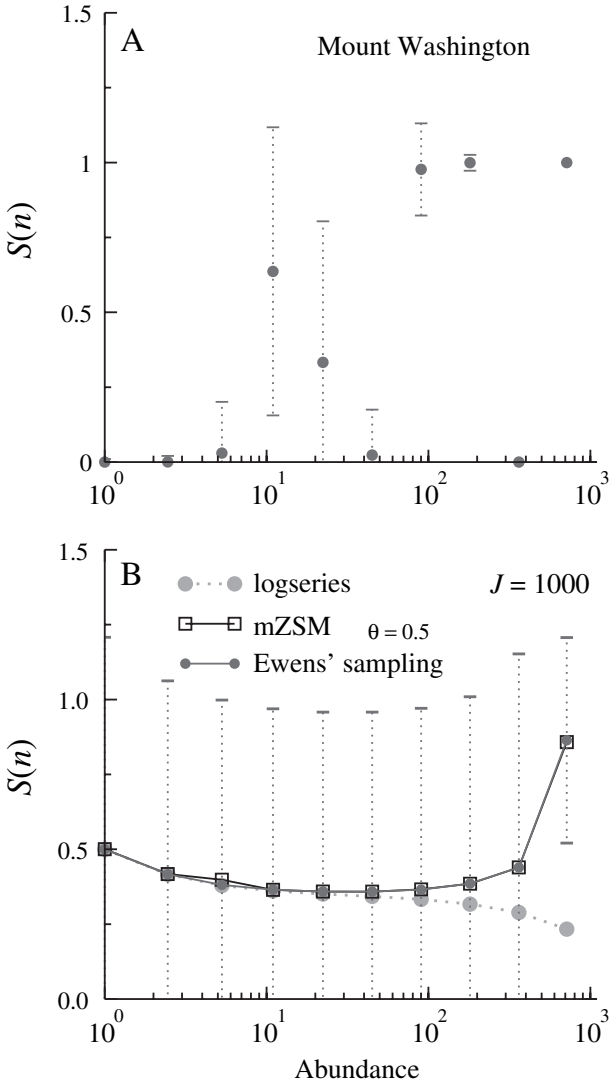

Figure 1 pZSM vs. log-series in species-poor communities. (a) We have plotted the sample SAD after averaging over randomly generated samples from the metacommunity made up of four species at the frequencies observed on Mount Washington. In both plots, the average number of species in each abundance class has been represented along with the expected standard deviation. (b) Abundance distribution for a random sample from a metacommunity such as a boreal forest [$\theta = 0.5$ (e.g. Hubbell 2001, p. 147)] predicted by neutral theory (mZSM) and by the log-series approximation. Samples are generated by using Ewens' algorithm. For instance, in a random sample of 500 individuals from a low diversity metacommunity ($\theta = 0.15$) there is a 0.43 probability of finding one species being represented by more than half the sample. The same probability calculated with the log-series yields an incorrect prediction of 0.09.

contain few species. It is important to know whether the functioning of such systems, for instance boreal forests, bogs, or heathlands, is consistent with the neutral theory. Given that it is far more reliable to sort individuals into species when analysing species-poor communities, tests based on abundance data from these communities will be more reliable than the same tests performed by using data from species-rich communities. Our results improve the ability to test the neutral theory by using abundance data.

In the second part of the article, we re-analyse several data sets: Williams' classical data (Fisher *et al.* 1943) on Lepidoptera in light traps, the Barro Colorado Island (BCI) tree plot (Condit *et al.* 2002), and a species-poor tree community from Mount Washington, NH, USA (Braun 1950). Our intention here is to provide worked examples showing the power and the limitations of this theoretical approach. In this context, we show how to use likelihood tests, which penalize for extra parameters, and tailored Monte Carlo tests, looking for significance in our fitted abundance curves. This framework is reliant on the statistical power we gain when analysing real abundance data. We find that deviations from neutral theory expectations are not large. Therefore, our analysis provides some evidence that neutral zero-sum dynamics and dispersal limitation could be assumed to be the main factors controlling community dynamics, at least as a first approximation and given the limited amount of data we have analysed. However, in general, noise in real data seems to be higher than that predicted by the neutral theory. Replicated, extensive data sets and sound methods to test against alternative theories, are necessary to uncover other mechanisms controlling community dynamics.

A MASTER EQUATION APPROACH

Two complementary analytic approaches have been used to develop Hubbell's neutral zero-sum theory. Hubbell (2001) re-interprets Ewens' sampling distribution (Tabare & Ewens 1997), initially introduced in the context of genetics (Ewens 1972; Karlin & McGregor 1972), and uses it as a species abundance distribution of a metacommunity undergoing neutral zero-sum dynamics. Recently, by extending this approach, Etienne & Olff (2004b) have derived the corresponding multivariate sample distribution, also of the local community. However, an alternative approach is possible. Frequently insight is gained when ecological interactions are formulated as a one-step stochastic process governed by a master equation in continuous time (Renshaw 1991; McKane *et al.* 2000; Solé *et al.* 2000; Stollenwerk & Briggs 2000; Stollenwerk & Jansen 2003; Alonso 2004). This formulation has been recently used to address neutral zero-sum dynamics by several authors (Vallade & Houchmandzadeh 2003; Volkov *et al.* 2003; McKane *et al.* 2004). Incidentally, it is also important to remark that so far all these approaches study the spatially implicit formulation of Hubbell's theory assuming the

point-mutation mode of speciation (see Chapter 4 Hubbell 2001). It is essentially only in this case that analytic expressions for various quantities of interest have been given so far (but see Chave & Leigh 2002; He & Hubbell 2003; Houchmandzadeh & Vallade 2003; Chave 2004). As far as we know, there are no, even approximate, analytic expressions for the corresponding SADs in the spatially-explicit formulation of the theory.

Here our starting point is a careful formulation of metacommunity dynamics. The metacommunity is isolated [a biogeographical region (Rosenzweig 1995)] and, therefore, its dynamics is only controlled by two processes: mutation and reproduction. Vallade & Houchmandzadeh (2003) describe metacommunity dynamics in terms of a non-linear one-step stochastic process where birth–death transition rates can be written as:

$$g_n = T(n+1|n) = \beta \frac{n(J_M - n)}{J_M(J_M - 1)}, \quad (1)$$

$$r_n = T(n-1|n) = \beta \frac{n(J_M - n)}{J_M(J_M - 1)} + v\frac{n}{J_M}. \quad (2)$$

These transition rates can be linearized for large J_M, giving rise to:

$$g_n = T(n+1|n) = \beta \frac{n}{J_M}, \quad (3)$$

$$r_n = T(n-1|n) = (\beta + v)\frac{n}{J_M}. \quad (4)$$

It is convenient in this case to rescale time by introducing $\tau = t/J_M$. This now gives transition probabilities:

$$\tilde{g}_n = \beta n; \quad \tilde{r}_n = (\beta + v)n. \quad (5)$$

This linear representation for r_n and g_n is precisely the starting point chosen by Volkov *et al.* (2003) in order to derive the stationary abundance distribution at the speciation-extinction equilibrium. Kendall (1948) provides the general solution for a one-step stochastic process with linear transition rates in the context of a birth–death process with immigration. Here, the linear approximation provided by eqn 5 leads to the log-series abundance distribution at the metacommunity level (Volkov *et al.* 2003). On the other hand, Vallade & Houchmandzadeh (2003) find the exact species abundance distribution at the speciation-extinction equilibrium in the metacommunity, without considering any further approximation. To stress the connection between this exact solution and empirical data, in this report we build the corresponding sample SADs in different asymptotically meaningful situations.

To keep a uniform notation, it will be useful to re-write the result from Vallade & Houchmandzadeh (2003) as

$$S_M(n) = \frac{\theta}{n}\frac{\Gamma(J_M + 1)}{\Gamma(J_M + 1 - n)}\frac{\Gamma(J_M + \theta - n)}{\Gamma(J_M + \theta)}, \quad (6)$$

where $S_M(n)$ is the expected number of species represented by n individuals within the metacommunity, and θ, the fundamental biodiversity number, is defined as $\theta = (J_M - 1)v/\beta$. Since reproduction and death are coupled by the zero-sum rule, the metacommunity size is fixed and denoted by J_M, v is the probability that an individual undergoes a mutation per unit time and β is the probability that it reproduces per unit time. It must be noted that Hubbell's definition of v is dimensionless (probability of giving rise to a new species per birth). So, the above definition of the biodiversity number is closely related to Hubbell's definition, $\theta = 2J_M v$.

THE SAMPLING DISTRIBUTIONS

The distinction between a distribution within a given community and the distribution observed in a sample from this community is still not widely appreciated (Pielou 1969). Dewdney (1998) showed that by assuming random sampling, the community distribution and the sample distribution will tend to be very similar. For instance, when the community is described by a continuous abundance distribution such as the famous lognormal distribution (Preston 1962), the sample distribution becomes a finite discrete representation of the distribution at the community level (Bulmer 1974). However, they could differ depending on how the sampling process has been carried out. Since Hubbell's theory is formally based on the Ewens' sampling theory of selectively neutral alleles (Ewens 1972; Karlin & McGregor 1972), in essence it is a sampling theory. As a consequence, any analytic expression for the distribution of species abundances in the recent literature (Hubbell 2001; Vallade & Houchmandzadeh 2003; Volkov *et al.* 2003; Etienne & Olff 2004b) applies either to the whole community or a small sample from it. This means that by understanding the community size as our sample size we are turning the community distribution into a sample distribution. However, this is only true if the sampling is random. We think that this point has not been stressed enough. For instance, consider an island connected by migration to a mainland. If we sample the island and try to test to what extent limited dispersal from the continent and zero-sum neutral dynamics explains our data, we should first randomize our local sampling within the island. Otherwise, if we have only taken a sample from a particular island locality, dispersal-limitation effects within the island (especially if the island is large) mean that the expressions we have for Hubbell's spatially implicit theory cease to apply to this case in a straightforward way. Here, by highlighting the sampling nature of Hubbell's theory, we would like to re-interpret the analytic SADs of this theory in the light of a general theory of sampling process (Pielou 1969; Dewdney 1998). This will emphasize how other sampling processes could modify these expressions.

We denote by $S(n)$ the expected number of species represented by n individuals in a sample from a given ecological community. In the most general case, this sample distribution is linked to the real abundance distribution, $Q(k)$, in the community by (Dewdney 1998; Lande *et al.* 2003):

$$S(n) = \sum_{k=1}^{J_M} f(n|k)Q(k), \tag{7}$$

where $f(n|k)$ gives the probability for any species having k individuals in the community to enter the sample with exactly n individuals. This conditional probability, the sample transformation function (Dewdney 1998), is the core of any sampling theory. It depends on the assumptions of how individuals enter the sample, for instance: is there species aggregation or can we assume unbiased random sampling?

In the context of the neutral theory, the exact solution for the expected number of species represented by n individuals in a local sample of J individuals can be recast as a sample distribution (eqn 7). Since we know the SAD in the metacommunity, $S_M(n)$ (Vallade & Houchmandzadeh 2003), all that is required is to determine the set of probabilities $f(n|k)$. These probabilities, rather than being random Poissonian, should be affected by dispersal limitation, because species in the sample do not appear randomly from the metacommunity. An analytical expression for the probability $P_s(n; N, m, x)$ of finding a certain species with relative abundance in the metacommunity x, being represented by n individuals in a local community of size N and connected by migration to a much larger metacommunity, has recently been found independently by several authors (Vallade & Houchmandzadeh 2003; Volkov *et al.* 2003; McKane *et al.* 2004, and see Appendix S1 in the Supplementary Material, where its mathematical form is also given). As stated before, by assuming unbiased random sampling at the local level, any sample of J individuals can be considered as an equivalent local community of that size. Therefore, we notice that the probability $P_s(n; J, m, x)$ is nothing else but $f(n|k)$, where J is the sample size, and $k = xJ_M$ is the abundance in the metacommunity. These probabilities transform the distribution at the metacommunity level, which is inaccessible, into the sample distribution that will be actually encountered when we get a sample from a dispersal-limited locality.

We may now introduce the expressions for $P_s(n; J, m, x)$ and $S_M(n)$, given by eqn 6, into the general eqn 7, to obtain the general solution for a local community of Hubbell's unified theory in the form of a sample distribution:

$$S(n) = \sum_{k=1}^{J_M} P_s(n; J, m, k/J_M)S_M(k), \tag{8}$$

where $n = 0,\ldots,J$, and J is the sample size.

We remark again that assuming random local sampling, the abundance distribution given in eqn 8 is also true for the whole local community, as was found by Vallade & Houchmandzadeh (2003). By using the general theory of the sampling process of Dewdney (1998), this result can be formalized, since eqn 8 can be shown to be invariant under hypergeometric sampling.

The infinite metacommunity assumption

Since metacommunties are large, they may be considered to be effectively infinite. If this assumption is made, then in the asymptotic limit as J_M tends to infinity, the stationary state at the speciation-extinction equilibrium can be expressed by a continuous abundance distribution (Vallade & Houchmandzadeh 2003):

$$f_M(x)\mathrm{d}x = \frac{\theta}{x}(1-x)^{\theta-1}\,\mathrm{d}x, \tag{9}$$

which represents the number of species with a relative abundance x within the abundance interval $(x, x + \mathrm{d}x)$ in the metacommunity. This expression may be deduced directly from eqn 6 by using the asymptotic expression for gamma functions (Abramowitz & Stegun 1965), and is central to the derivation of the sample SADs we present here.

By using eqn 9, a corresponding asymptotic form (for an infinite metacommunity) of eqn 8 can be also written:

$$S(n) = \theta \int_0^1 P_s(n; J, m, x)\frac{(1-x)^{\theta-1}}{x}\,\mathrm{d}x, \tag{10}$$

where now $n = 1,\ldots,J$. In Appendix S1 (see Supplementary Material) we relate this solution to that in Volkov *et al.* (2003). Notice that none of these asymptotic forms allows an estimation of the total species richness in the metacommunity nor of the total metacommunity size, J_M: eqn 8 is required to estimate these very relevant quantities. By summing over $n = 0,\ldots, J$, eqn 8 gives the expected value for the number of species in the metacommunity. On the other hand, if we only wish to estimate the expected number of species entering a sample of size J, we can use either eqn 8 or eqn 10 by summing now over $n = 1,\ldots, J$. Carrying out this procedure, our estimates for the average number of species in a sample of size J match those obtained using the formulas given in Etienne & Olff (2004a). Our asymptotic approach relies on assuming the metacommunity to be infinite. Numerically, we have checked that eqn 8 is only slightly sensitive to J_M when J_M reaches 10^5–10^6. In a forthcoming paper, we will develop a large J_M approximation to address this point in an analytic way.

As the immigration parameter tends to 1, the general expression for the zero-sum multinomial abundance distribution, eqns 8–10, which could be called the local or

migration-limited zero-sum multinomial (labelled throughout this report as either localZM, or simply ZSM), has a limiting form corresponding to a random sampling from the metacommunity (metaZSM, or mZSM, for short). If the metacommunity is assumed to be finite, the exact sample distribution is again eqn 6, since it can also be shown that this expression is invariant under the hypergeometric transformation. In an infinite metacommunity, as $m \to 1$, eqn 10 naturally gives rise to a random sampling which is binomially distributed, the expected number of individuals in the sample belonging to a species with relative abundance x in the metacommunity being Jx:

$$S(n) = \theta \binom{J}{n} \int_0^1 x^n (1-x)^{J-n} \frac{(1-x)^{\theta-1}}{x}\, dx, \qquad (11)$$

For large samples, the exact sample distribution given by eqn 11 can be very well approximated by using the Poisson distribution (Pielou 1969; Dewdney 1998):

$$S(n) = \theta \int_0^1 \exp(-xJ) \frac{(xJ)^n}{n!} \frac{(1-x)^{\theta-1}}{x}\, dx. \qquad (12)$$

In this case, the integral can be approximately evaluated and given as an analytic formula which is easy to use (see Appendix S2 in Supplementary Material):

$$S(n) = \frac{\theta}{n}\left(1 - \frac{n}{J}\right)^{\theta-1} + \mathcal{O}\left(\frac{1}{J^2}\right). \qquad (13)$$

The asymptotic distribution in eqn 11 is the asymptotic form of the ZSM, when there is no dispersal limitation (metaZSM, mZSM), i.e. the distribution corresponding to a sample from a globally mixed or panmictic infinite metacommunity. Its Poissonian approximation in eqns 12 and 13 can be seen as a compound Poisson distribution (Bulmer 1974). There are two possible concerns about the use of this new approximate distribution. First, the sample size must be large ($J > 100$, see Appendix S2 in the Supplementary Material) and second, when $n = J$ and $\theta < 1$, the exact binomial formula (eqn 11) should be used instead. This framework should provide a better test of the neutral theory at the metacommunity level, as well as a better estimate of low values of the fundamental biodiversity number. In fact, the log-series and the Poisson metaZSM (eqn 13) are almost coincident only if the diversity values are not small. In this context, it is conceptually important to make the point that the Poisson metaZSM (eqn 13) is a sample distribution with the same meaning as that given by Fisher *et al.* (1943) to his log-series distribution: the expected number of species with n individuals in a random sample from a community. Both distributions are conceptually different from the metacommunity steady-state SAD (eqn 6). This SAD at the metacommunity level can only be approximated by a log-series by taking linear transition rates, which is equivalent to first assuming that zero-sum dynamics is irrelevant in an infinite system and then solving the resulting linear problem (Volkov *et al.* 2003). If we proceed the other way around, i.e. we first solve the exact problem and obtain eqn 6, and then look for the asymptotic behaviour of this solution as J_M tends to infinity, we obtain eqn 9, which gives rise to the sample SADs we are discussing in this paper. In contrast to the solution reported by Volkov *et al.* (2003), the sample SADs reported here keep the fingerprint of the zero-sum dynamics at the metacommunity level, even in the case when we take the metacommunity to be infinite.

How is the metaZSM to be tested empirically? As Hubbell (2001, p. 318) suggests, 'the best estimates of θ are likely to be obtained from pooling samples collected all across the metacommunity'. However, as a consequence of our findings, instead of the log-series distribution the new metaZSM should be used as the expected sample SAD when dealing with species-poor communities. In order to test the theory, we need to sample a species-poor metacommunity at random, and see whether the metaZSM applies. Alternatively, we can pool together a large number of samples from different localities of a metacommunity and perform a re-sampling of R small pseudo-replicas of size J from the whole pool of all individuals. This mimics a random sampling of the metacommunity. Since we have R replicas, by averaging we can empirically calculate the probability of having a species represented by $n = 0,1,2,\ldots, J$ individuals in a sample of size J. If the metacommunity is very poor ($\theta < 1$, as, for instance, in a boreal forest), and this probability is well described by a metaZSM, the shape of the abundance distribution should show an upward bend at the right end of the curve, as seen in Fig. 1b. If this were the case, we should conclude that the empirical measured sample SAD is consistent with Hubbell's zero-sum neutral metacommunity dynamics (see the worked example on the boreal forest of Mount Washington, NH, USA and Fig. 1a).

METHODS

Data

In order to show the applicability of this theoretical framework, we have mainly used two data sets. The first one is Williams' classical data on Lepidoptera (Fisher *et al.* 1943). These data were collected by means of a light trap in England during the years 1933-1936. Only specimens completely characterized by their full species name were included in the analysis, mainly belonging to certain families (Sphingidae, Noctuidae, Arctiidae, Geometriidae, and a few other related families). The second is the 50 ha rainforest plot of Barro Colorado Island (BCI), where trees larger than

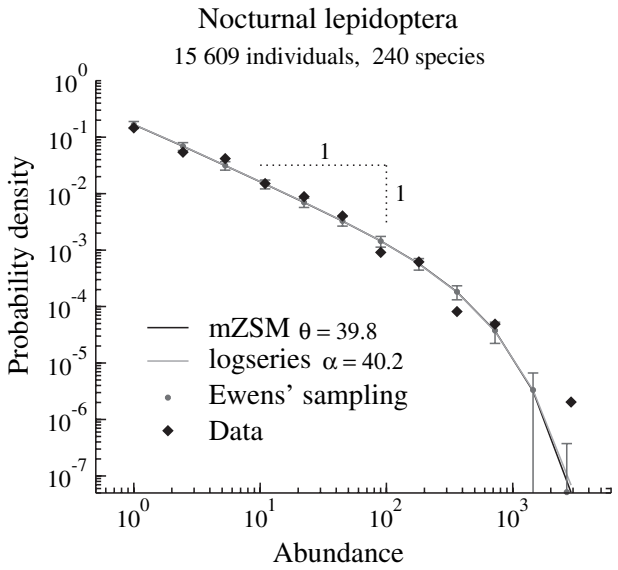

Figure 2 Abundance data on Lepidoptera in light traps. Data on Lepidoptera collected during 4 years in England (Fisher *et al.* 1943). Here 12 abundance intervals are defined: $2^{i-1} \leq n < 2^i, i = 1,\ldots,12$. A Monte Carlo generated curve using Ewens' algorithm (Hubbell 2001) is also shown. Standard errors (over 1000 random generated samples of 15 609 individuals) are plotted at each abundance interval.

10 cm DBH were counted and sorted out up to the level of species (Condit *et al.* 2002). Again, only specimens completely characterized by their full species name were considered. Finally, as an example of a poor-species community, we used data from a boreal forest (mid-elevation) on Mount Washington (NH, USA) described in Braun (1950) and also analysed in Hubbell (2001, page 147). These data are given as the average frequencies of this community which contains only four species (Spruce: 75%; Balsam Fir: 16.5%; Paper Birch: 8.0%; Yellow Birch: 1.4%).

The way abundance data were plotted (Fig. 2) is intended to reduce the noise usually present in this type of data (Fig. S1 and Fig. 3). In Fig. 3, the *y* axis is the number of species at each abundance interval. In Fig. 2, the probability density is defined as the number of species having abundances *n* within an interval divided by the abundance interval length and the total number of species in the sample (Pueyo 2003) (see Supplementary Material).

General fitting method: maximum likelihood

As a first step, in order to obtain maximum likelihood estimates, we need to calculate the probability of obtaining a data set supposing that the model and its parameter values are known. From any sample abundance distribution, for example eqns 7–13, we can define the probability of a species being represented by *n* individuals in the sample:

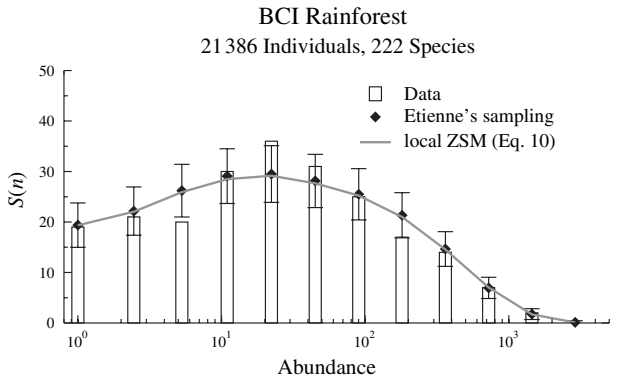

Figure 3 Data on BCI tree species community. A Monte Carlo generated curve using Etienne's algorithm (R.S. Etienne, unpublished data) is also shown. Standard errors (over 1000 randomly generated samples of 21 386 individuals) are plotted at each abundance interval.

$$p(n) = \frac{S(n)}{\sum_{i=1}^{J} S(i)}, \tag{14}$$

where J is the sample size. Therefore, if we sort species in terms of their abundances (S_1 is the number of singleton species in the sample, and so on), the probability of obtaining a collection, $S_1,\ldots,S_a$, of a given size J can be written as:

$$Pr\{S_1,\ldots,S_a|\text{Model}\} = \frac{S!}{S_1!\ldots S_m!} p(1)^{S_1}\ldots p(a)^{S_a}, \tag{15}$$

where S_i is the number of species with i individuals in the sample, S is the number of species in the sample and a is the maximum abundance observed in the sample. Since we are interested in how the likelihood of our data set changes when changing the parameters defining the model, we take as a likelihood function:

$$\mathcal{L}\{S_1,\ldots,S_a|\theta,m\} = p(1)^{S_1}\ldots p(a)^{S_a}, \tag{16}$$

where we have labelled the model in terms of the parameters our abundance distributions depend on. In practice, our maximum likelihood estimates are those maximizing eqn 16, i.e. the likelihood of the data set. Alternatively, when looking for the maximum, it is usually suggested that the negative of the logarithm of the likelihood should be used (Hilborn & Mangel 1997):

$$\mathbf{L}\{S_1,\ldots,S_m|\theta,m\} = -\sum_{i=1}^{a} S_i \log(p(i)). \tag{17}$$

The minimum of this log-likelihood function gives the maximum likelihood estimate of the parameter and its width information about the confidence we have in our estimates. Confidence intervals were calculated using the method described in Hilborn & Mangel (1997) (see Supplementary Material).

The derivation of this likelihood function is consistent with the asymptotic formulas we are using, since eqn 14 is exact assuming an infinite system. An exact formulation for this likelihood, valid also for finite metacommunities, is given by Etienne & Olff (2004b), as the multivariate probability of randomly extracting a particular collection of individuals from a local community. However, our alternative likelihood for large metacommunities is simpler and has a solid foundation. In particular, it is related to the general concept of 'composite likelihood'. In high dimensional problems, when the multivariate likelihood is unknown, or very difficult to calculate, composite likelihood methods apply and have been used successfully for maximum likelihood estimation in ecology and other areas (Heagerty & Lele 1998; Lele & Taper 2002).

The likelihood ratio test

The ZSM (eqn 10) is a two-parameter abundance distribution, $S(n) = F(n; \theta, m)$, while the mZSM (eqn 13) is a one-parameter abundance distribution, $S(n) = G(n; \theta)$. Within the framework developed here, we have shown that $F(n; \theta, m)$ collapses into $G(n; \theta)$ when the degree of isolation of the local community being sampled is zero ($m = 1$). This enables us to use the likelihood ratio test (see Hilborn & Mangel 1997, and references therein) to assess for the significance of the immigration parameter ($m < 1$) and see to what extent dispersal limitation plays a significant role in shaping our abundance data (see Supplementary Material).

Monte Carlo tests

The fitting of the abundance curve for the Lepidoptera community was assessed by performing two Monte Carlo tests. These tests mimic a random sampling either from the metacommunity or from the local community, where a given large number of independent pseudo-samples are generated. These pseudo-data can be used to build a Monte Carlo χ^2 test as an alternative to parametric tests based on the χ^2 probability distribution (Hilborn & Mangel 1997). In cases where the assumptions for the application of classical χ^2 tests are not completely fulfilled or the provided results are not sufficiently clear, Monte Carlo χ^2 tests are specially indicated (see Supplementary Material).

RESULTS

Species-poor communities

Since data from the boreal forest we analysed are given as the real relative frequencies encountered in the field through different samples taken over several years, we will assume that they roughly describe the metacommunity composition of this type of mid-elevated boreal forest. In order to investigate the shape of a SAD from such a metacommunity, we have generated samples of 1000 trees according to the given relative frequencies of the four metacommunity species. This mimics a random sampling from such a metacommunity. We have calculated the sample SAD after averaging a number of randomly generated samples from the metacommunity made up by only the four species at the frequencies observed on Mount Washington (see Fig. 1a). The shape of the abundance distribution actually shows an upward bend at the right end of the abundance curve. The logseries completely fails to capture this feature. Although the way we have generated the samples – by assuming real frequencies at the metacommunity level – is artificial, this simple example is meant to highlight the fact that poor communities will naturally show this upward bend behaviour at the right end of the abundance curve, and that the right sample distribution predicted by the neutral theory, the metaZSM (but not the logseries), is able to account for this feature.

Species-rich communities

In Table 1 we summarize our results corresponding to the two data sets analysed from two different species-rich communities: Williams' Lepidopetera community (Fisher *et al.* 1943) and BCI (Condit *et al.* 2002).

Table 1 Williams' Lepidoptera community (Fisher *et al.* 1943) and BCI (Condit *et al.,* 2002)

	Lepidoptera	BCI
Logseries	$\alpha = 40.2(28.1,51.1)$	
$S(n) = \alpha\frac{x^n}{n}$	$\chi^2_o = 15.7, df = 10$	
metaZSM	$\theta = 39.8(29.7,51.9)$	$\theta = 34.4(25.6,45.3)$
(eqn 13)	$\chi^2_o = 16, df = 10$	$\chi^2_o = 22, df = 9$
$Pr(\chi^2 < \chi^2_o \mid df)$	0.9	
$Pr(\chi^2_{MC} < \chi^2_o)$	0.95	
localZSM	$\theta = 41(31.1,52.7)$	$\theta = 44(32.,56.)$
(eqn 10)	$m = 0.77(0.3,0.95)$	$m = 0.15(0.05,0.4)$
	$\chi^2_o = 15.5, df = 9$	$\chi^2_o = 4.43, df = 8$
$Pr(\chi^2 < \chi^2_o \mid df)$	0.92	0.18
$Pr(\chi^2_{MC} < \chi^2_o)$	0.95	
Ratio likelihood test	R = 3.96	R = 14.0
$[Pr(\chi^2 < 3.84 \mid 1) = 0.95]$		

Probabilities $Pr(\chi^2 < \chi^2_o \mid df)$ have been calculated by using the classical χ^2 distribution with df degrees of freedom. Probabilities $Pr(\chi^2_{MC} < \chi^2_o)$ have been computed by performing Monte Carlo tests. R stands for the likelihood ratio which is distributed following a χ^2 distribution with one degree of freedom (see Supplementary Material).

BCI data set

The abundance distribution of the tree community in Barro Colorado island has been used as an example where the ZSM works well (Hubbell 2001). This has been challenged (McGill 2003; Etienne & Olff 2004b), but the final conclusion of these criticisms highlights the low discriminative power of species abundance data (Etienne & Olff 2004b). Moreover, the parameters defining the classical lognormal (Preston 1962) are not derived from any dynamical theory of community organization. In the future, it will become necessary to challenge the ZSM fit against well-defined non-neutral theories of community dynamics. In Fig. 3, we plot the ZSM (eqn 10) and the solution calculated by using Etienne's algorithm, which also gives the expected variance of each abundance interval. Both theoretical solutions match each other and describe the empirical data successfully. The ratio likelihood test reveals the high significance of the parameter m.

Lepidoptera community

As can be seen in Fig. 2, the mZSM and log-series give results for the SAD which are very close to each other. By using large samples, and for large values of θ, the logseries closely resembles the actual sample distribution, the mZSM. In particular, the estimates of the fundamental biodiversity numbers are reliable by using either the mZSM or the log-series (Fig. 2, Table 1). In general, the logseries tends to overestimate the fundamental biodiversity number, but this is not significant in Williams' data.

Since the Lepidoptera data were used as a first example of the Fisher logseries (Fisher *et al.* 1943), in the light of the neutral theory the natural underlying null hypothesis is that the sample is a random sample from the metacommunity. Furthermore, these data come from a survey carried out over 4 years, so we can explore whether the temporal dimension of the sampling might have turned these data into an approximate random sampling from the metacommunity. However, a careful Monte Carlo test by using Ewens' sampling allows us to reject this hypothesis. In 95% of cases, the Monte Carlo χ^2 statistic was lower than the observed value in the data (Fig. S1). This result shows that we can reject this null hypothesis at the confidence level of 0.05 (see Supplementary Material).

We can therefore conclude that the data do not seem to be a random sample from the metacommunity undergoing zero-sum neutral dynamics. In fact, the original analysis of Williams (Fisher *et al.* 1943) pointed to the fact that the data deviated from the log-series to some extent. In particular, small samples, which capture the commonest species in the system, were more homogeneous and contained fewer species (Fisher *et al.* 1943) than predicted. In the context of neutral theory, this could be a fingerprint of some dispersal limitation affecting the local community. To consider this possibility, we used eqn 10 to fit the data (Fig. S1). Since we have shown that for the general ZSM, the sample distribution for the local community ($m < 1$, eqn 10), collapses into the mZSM as $m \to 1$ (eqn 13), we performed a likelihood ratio test (see Methods) to assess the significance of the added parameter m. Twice the difference in negative log-likelihoods turned out to be only 3.98 (see Supplementary Material). Therefore, it must be concluded that the ZSM, which gives a better goodness-of-fit by taking into account dispersal limitation, is only slightly, but significantly, better than the one-parametric mZSM. On the other hand, the deviation of the real data from the theoretical localZSM reveals that the data is quite noisy. By using the maximum likelihood estimates for the parameters m and θ, the Monte Carlo test would allow a rejection of the general two-parametric ZSM model (localZSM) to be made at a confidence level of 0.05 but not at the confidence level of 0.01. In fact, by assuming the localZSM with this parameter set, there is a 5% probability of obtaining random pseudo-samples deviating from the expected theoretical values by even more than the deviation observed in the real data.

DISCUSSION

In this report, we have analysed two contrasting examples. Our results rely on clear analytical predictions for the expected sample SADs of the spatially implicit formulation of the theory. The estimated parameters for the BCI rainforest are in agreement with previous values, in particular, with those obtained by the Bayesian approach of Etienne & Olff (2004b). The local community of BCI seems to be clearly dispersal-limited. The inclusion of the immigration parameter considerably improved the performance of the model. The analysis of the same data as Condit *et al.* (2002) using a spatial formulation of the neutral theory is consistent with this result, at least at some spatial scales, suggesting that other mechanisms different from zero-sum dynamics and neutrality control species abundances at other spatial scales. The significance of the parameter m for the Lepidoptera community is not so clear. Although, the localZSM fails to improve the fitting in a clear way, our analysis points to the fact that species entered this local sample with a slight degree of dispersal limitation. However, abundance classes show an underlying variability which is not well captured by the theoretical distribution. Notice that the theoretical variances for each abundance interval are also given (Fig. 3, Fig. S1). This might suggest that there is little evidence that neutral dynamics and dispersal limitation are the main factors controlling local dynamics of Lepidoptera community at this spatio-temporal scale. Similar results have been found in other insect communities (Stork 1997; Hubbell 2001; Lande *et al.* 2003). However, a spatial sampling along with a spatially

explicit neutral theory (Hubbell 2001; Chave & Leigh 2002) might account for the observed pattern in a better way.

Nevertheless, the main limitation of our conclusions is that they are based on only one trait of community organization. In general, whenever possible, different community patterns should be taken together and compared with expectations from the theory to arrive at a coherent answer (Harte 2003). This is particularly true when we lack replication and when abundance data are rather noisy, as seems to be the case in this example (Fig. S1). We stress also the importance of having good statistical methods to test between alternative theories of community dynamics making contrasting predictions about species abundance distributions.

The Lepidoptera abundance distribution analysed here shows a slope −1 at lower abundance classes (Fig. 2). This feature is well captured by the theory. Extensive random samples collected over large areas and long periods of time (Margalef 1994) show the same pattern (Pueyo 2003). This is a signature of neutral zero-sum dynamics at the metacommunity scale. It strongly suggests that neutral zero-sum dynamics pervades metacommunities controlling rarity at large spatio-temporal scales.

Neutral coexistence relies on the ecological equivalence of the different species in a trophic-defined community. Hubbell (2001) proposed that coevolution tends to act to make all species' fitness approximately the same in a given environment through the appearance of well balanced trade-offs. This equalizing effect along with neutral speciation allows a high level of diversity to be maintained. Since a great number of stabilizing mechanisms (Chesson 2000) have been theoretically predicted and empirically identified as responsible for coexistence, mainly at short spatio-temporal scales, it is important to elucidate whether Hubbell's neutral theory can be taken as a zeroth-order approximation at large spatio-temporal scales. In order to have a reliable global picture of biodiversity from the widest perspective, differences among species belonging to the same trophic community might not be so important. However, at smaller scales, a few common species seem to be superabundant and persistent (Clark & McLachlan 2003; Magurran & Henderson 2003). Are these dominant species at the local level a simple consequence of ecological drift or are they also responsible for the lack of rare species (in comparison with the metacommunity) through competitive exclusion or other general ecological mechanisms (Chesson 2000)? To really achieve an understanding of the relative importance of neutrality and dispersal limitation, compared to non-neutral factors in the assembly of local communities, a synthetic theory reconciling neutrality and niche assembly is needed. Some preliminary attempts along these lines have already been made (Etienne & Olff 2004a). We believe this is a great challenge for community ecology in the near future. The unified theory of biodiversity is only the first step in this direction.

ACKNOWLEDGEMENTS

We wish to thank Mercedes Pascual, Simon Levin, Subhash Lele, and Wei Zeng for reading the manuscript and giving suggestions for improvements. Referee reports on earlier versions were extremely helpful in improving and re-structuring the current version. We also thank Rampal Etienne for kindly providing a fast algorithm to sample the local community. Interesting discussions with Salvador Pueyo and Frederic Bartumeus during early stages of this work are also acknowledged. DA would like to thank the MACSIN research group at the UFMG, Belo Horizonte, Brazil, the theoretical physics group at the University of Manchester, U.K., and the CSL in Barcelona, Spain, for constant support and a nice working environment. Funding from the ESF (InterAct grant) and from the NMF is also gratefully acknowledged. DA would like to dedicate this work to the memory of Dr R. Margalef.

SUPPLEMENTARY MATERIAL

The following material is available from http://www.blackwellpublishing.com/products/journals/suppmat/ELE/ELE640/ELE640sm.htm:

Appendix S1 Relationship between different approaches.

Appendix S2 A simple formula for the Poisson mZSM.

Figure S1 Abundance data on Lepidoptera in light traps.

Figure S2 Lepidoptera Community. Monte Carlo test.

REFERENCES

Abramowitz, M. & Stegun, I.A. (eds) (1965). *Handbook of Mathematical Functions*. Dover, New York.

Alonso, D. (2004). *The Stochastic Nature of Ecological Interactions: Communities, Metapopulations, and Epidemics*. PhD Thesis, Polytechnic University of Catalonia, Barcelona.

Bell, G. (2001). Neutral macroecology. *Science*, 293, 2413–2417.

Braun, E.L. (1950). *Deciduous Forests of Eastern North America*. The Free Press, New York.

Bulmer, M.G. (1974). On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics*, 30, 101–110.

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

Chave, J. & Leigh, E.G. (2002). A spatially explicit neutral model of β-diversity in tropical forests. *Theor. Popul. Biol.*, 62, 153–168.

Chesson, P.L. (2000). Mechanisms of maintenance of species diversity. *Ann. Rev. Ecol. Syst.*, 31, 343–366.

Clark, J.S. & McLachlan, J.S. (2003). Stability of forest biodiversity. *Nature*, 423, 635–638.

Condit, R., Pitman, N., Leigh, J.E.G., Chave, J., Terborgh, J., Foster, R.B. *et al.* (2002). Beta-diversity in tropical forest trees. *Science*, 295, 666–669.

Dewdney, A.K. (1998). A general theory of the sampling process with applications to the veil line. *Theor. Popul. Biol.*, 54, 294–302.

Enquist, B.J., Sanderson, J. & Weiser, M.D. (2002). Modeling macroscopic patterns in ecology. *Science*, 295, 1835–1837.

Etienne, R.S. & Olff, H. (2004a). How dispersal limitation shapes species-body size distributions in local communities. *Am. Nat.*, 163, 68–83.

Etienne, R.S. & Olff, H. (2004b). A novel genealogical approach to neutral biodiversity theory. *Ecol. Lett.*, 7, 170–175.

Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, 3, 87–112.

Fisher, R., Corbet, A. & Williams, C. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.*, 12, 42–58.

Harte, J. (2003). Tail of death and resurrection. *Nature*, 424, 1006–1007.

He, F. & Hubbell, S.P. (2003). Percolation theory for the distribution and abundance of species. *Phys. Rev. Lett*, 91, 198103.

Heagerty, P. & Lele, S. (1998). A composite likelihood approach to binary data in space. *Journal of American Statistical Association*, 93, 1099–1111.

Hilborn, R. & Mangel, M. (1997). *The Ecological Detective. Confronting Models with Data*. Princeton University Press, Princeton, NJ.

Houchmandzadeh, B. & Vallade, M. (2003). Clustering in neutral ecology. *Phys. Rev. E*, 68, 061912.

Hubbell, S.P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton, NJ.

Karlin, S. & McGregor, J. (1972). Addendum to the paper of W. Ewens. *Theor. Popul. Biol.*, 3, 113–116.

Kendall, D.G. (1948). On some modes of population growth leading to R. A. Fisher's logarthmic series distribution. *Biometrika*, 35, 6–15.

Kneitel, J.M. & Chase, M. (2004). Trade-offs in community ecology: linking spatial scales and species coexistence. *Ecol. Lett.*, 7, 69–80.

Lande, R., Engen, S. & Saether, B.-E. (2003). *Stochastic Population Dynamics in Ecology and Conservation. Oxford Series in Ecology and Evolution*. Oxford University Press, Oxford.

Lele, S. & Taper, M.L. (2002). A composite likelihood approach to estimation of co-variance components. *Journal of Statistical Planning and Inference*, 103, 117–135.

MacArthur, R.H. & Wilson, E.O. (1967). *The Theory of Island Biogeography*. Princeton University Press, Princeton, NJ.

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

Margalef, R. (1994). Through the looking glass: how marine phytoplankton appears through the microscope when graded by size and taxonimically sorted. *Science Marine*, 58, 87–101.

McGill, B.J. (2003). A test of the unified theory of biodiversity. *Nature*, 422, 881–885.

McKane, A., Alonso, D. & Solé, R.V. (2000). A mean field stochastic theory for species rich assembled communities. *Phys. Rev. E*, 62, 8466–8484.

McKane, A., Alonso, D. & Solé, R.V. (2004). Analytical solution to Hubbell's neutral model of local community dynamics. *Theor. Popul. Biol.*, 65, 67–73.

Mouquet, N. & Loreau, M. (2002). Coexistence in metacommunities: the regional similarity hypothesis. *Am. Nat.*, 159, 420–426.

Mouquet, N. & Loreau, M. (2003). Community patters in source-sink metacommunities. *Am. Nat.*, 162, 544–557.

Pielou, E.C. (1969). *An Introduction to Mathematical Ecology*. Wiley, New York.

Preston, F.W. (1962). The canonical distribution of commonness and rarity. *Ecology*, 43, 185–215,

Pueyo, S. (2003). *Irreversibility and Criticality in the Biosphere*. PhD Thesis, University of Barcelona, Barcelona.

Renshaw, E. (1991). *Modelling Biological Populations in Space and Time, Vol. 11 of Cambridge Studies in Mathematical Biology*. Cambridge University Press, Cambridge.

Rosenzweig, M.L. (1995). *Species Diversity in Space and Time*. Cambridge University Press, Cambridge.

Solé, R.V., Alonso, D. & McKane, A. (2000). Scaling in a network model of multispecies communities. *Physica A*, 286, 337–344.

Stollenwerk, N. & Briggs, K.M. (2000). Master equation solution of a plant disease model. *Phys. Lett. A*, 274, 84–91.

Stollenwerk, N. & Jansen, V.A.A. (2003). Meningitis, pathogenicity near criticality: the epidemiology of a meningococcal disease as a model for accidental pathogens. *J. Theor. Biol.*, 222, 347–359.

Stork, N.E. (1997). Measuring global biodiversity and its decline. In: *Biodiversity II: Understanding and Protecting Our Biological Resources* (eds Reaka, M.L., Wilson, D.E. & Wilson, E.O.). Joseph Henry Press, Washington, DC, pp. 41–68.

Tabare, S. & Ewens, W.J. (1997). Multivariate Ewens distribution. In: *Discrete Multivariate Distributions* (eds Johnson, N.L., Kotz, S. & Balakrishman, N.). Wiley, New York, pp. 232–246.

Tilman, D. (1982). *Resource Competition and Community Structure*. Princeton University Press, Princeton, NJ.

Tilman, D. (1990). Constraints and trade-offs: toward a predictive theory of competition and succession. *OIKOS*, 53, 3–15.

Vallade, M. & Houchmandzadeh, B. (2003). Analytic solution of a neutral model of biodiversity. *Phys. Rev. E*, 68, 061902.

Volkov, I., Banavar, J.R., Hubbell, S.P. & Maritan, A. (2003). Neutral theory and relative species abundance in ecology. *Nature*, 424, 1035–1037.

Yu, D.W., Terborgh, J.W. & Potts, M.D. (1998). Can high tree species richness be explained by Hubbell's null model. *Ecol. Lett.*, 1, 193–199.

Editor, Nicholas Gotelli
Manuscript received 6 March 2004
First decision made 29 March 2004
Second decision made 26 May 2004
Manuscript accepted 14 June 2004