

A new sampling formula for neutral biodiversity: A new sampling formula

Rampal Etienne


Ecology Letters

Cite this paper

Downloaded from [Academia.edu](#) 

[Get the citation in MLA, APA, or Chicago styles](#)

Related papers

[Download a PDF Pack](#) of the best related papers 



[Confronting different models of community structure to species-abundance data: a Bayesia...](#)
Rampal Etienne

[A dispersal-limited sampling theory for species and alleles: A dispersal-limited sampling theory](#)

David Alonso

[A neutral sampling formula for multiple samples and an `exact' test of neutrality](#)

Rampal Etienne

LETTER

A new sampling formula for neutral biodiversity

Rampal S. Etienne
 Community and Conservation
 Ecology Group, University of
 Groningen, PO Box 14, 9750 AA
 Haren, The Netherlands
 Correspondence: E-mail:
 r.s.etienne@rug.nl

Abstract

The neutral model of biodiversity, proposed by Hubbell (*The Unified Neutral Theory of Biodiversity and Biogeography*, Princeton University Press, Princeton, NJ, 2001) to explain the diversity of functionally equivalent species, has been subject of hot debate in community ecology. Whereas Hubbell studied the model mostly by simulations, recently analytical treatments have yielded expressions of the expected number of species of a particular abundance in a local community with dispersal limitation. Moreover, a formula has been offered for the joint likelihood of observing a given species-abundance dataset in a local community with dispersal limitation, but this formula is too complicated to allow practical applications. Here, I present a much simplified expression that can be regarded as an enhanced version of the famous Ewens sampling formula. It can be used in maximum likelihood methods for quick estimation of the model parameters, using all information in the data, and for model comparison. I also show how to rapidly generate examples of species-abundance distributions for a given set of model parameters and how to calculate Simpson's diversity index.

Keywords

Biodiversity, community, Ewens sampling formula, Hubbell, neutral model, urn scheme.

Ecology Letters (2005) 8: 253–260

INTRODUCTION

Community ecology has recently been endowed with the neutral model of biodiversity (Hubbell 1997, 2001) that aimed at giving a simple explanation of biodiversity patterns such as species-abundance distributions and species–area curves. According to this model these patterns solely result from the stochastic processes of birth, death, speciation and immigration. This model has been heavily debated (e.g. Yu *et al.* 1998; Abrams 2001; Brown 2001; De Mazancourt 2001; Bengtsson 2002; Chave & Leigh 2002; Clark & McLachlan 2003; Fargione *et al.* 2003; Harte 2003; Ricklefs 2003), but it is more or less accepted as a useful null model that merits further study. In this paper I introduce a new sampling formula for the neutral model that has important applications in the confrontation of the neutral model to data. For a complete understanding of this formula, I first briefly review the neutral model and recent advances in neutral theory.

In Hubbell's (2001) model, when individuals in a local community die, they are immediately replaced by offspring of other local individuals or by immigrants from the regional species pool (the metacommunity in Hubbell's terminology), keeping the total number of individuals constant (the

zero-sum assumption). The replacement probability is proportional to each species abundance in the local community (when replaced by a local individual), or in the regional species pool (when replaced by an immigrant); this proportionality is the neutrality assumption. The regional species pool is in a balance between speciation and extinction. The model contains two parameters: the immigration probability m and the fundamental biodiversity number θ that is a measure of species diversity in the regional species pool. Hubbell (2001) defines θ and $\theta := 2J_M v'$, where J_M is the number of individuals in the regional species pool and v' is the speciation probability per unit birth. Vallade & Houchmandzadeh (2003) define $\theta := \frac{v(J_M-1)}{1-v}$, which is, apart from a factor of 2, equivalent to Hubbell's (2001) definition in the limit that Hubbell takes in his derivation ($J_M \gg 1$), because $v' := \frac{v}{1-v}$ where v is the speciation probability per time-step. The factor of 2 results from whether or not one allows multiple speciations to occur within one time-step. Regarding the immigration probability, $m < 1$ means that immigration is limited (dispersal limitation in Hubbell's terminology). The recognition of the potentially crucial role of dispersal limitation in determining biodiversity is one of the main achievements of Hubbell's work (Hubbell 1997, 2001; Hubbell *et al.* 1999),

although it has been noted before that dispersal limitation can promote coexistence of species (Tilman 1994; Hurtt & Pacala 1995; Loreau & Mouquet 1999).

While Hubbell (2001) was able to present analytical results for the model without dispersal limitation ($m = 1$) because he could borrow these from the neutral model in population genetics (Ewens 1972), he provided only simulation results for the biologically more interesting model with dispersal limitation ($m < 1$). Fortunately, the first analytical results for the latter case have been found recently. Two lines of analytical treatment can be distinguished. One line uses a mean-field master equation approach with a Markovian description of states and transitions (Vallade & Houchmandzadeh 2003; Volkov *et al.* 2003; Alonso & McKane 2004; McKane *et al.* 2004). This has resulted in exact analytical expressions and various approximations for the *expected number of species with a certain abundance* in a dispersal-limited local community. The other line takes a coalescent-type approach where community members are traced back to the ancestors that once immigrated into the community (Etienne & Olff 2004a,b). This line has resulted in a full and exact analytical expression for the *multivariate probability of observing a specific species-abundance distribution* in a sample of J individuals from the local community. The species-abundance distribution D contains the abundances of each species: if there are S species, then $D = (n_1, n_2, \dots, n_S)$. The multivariate probability is thus the likelihood $P[D|\theta, m, J]$. However, the expression for this likelihood obtained by Etienne & Olff (2004b) is hardly tractable in practice. The new sampling formula that I will present below is a much simplified version of this expression that can readily be used in practical applications (see below). For a correct interpretation of this formula, I will summarize Etienne & Olff's (2004b) approach.

Etienne & Olff's (2004b) first observation is that each individual in the local community can be traced back to a single ancestor that immigrated into the local community. This ancestor is of the same species as its descendant, because there is no speciation in the local community, so all information in a species-abundance distribution is contained in the combination of the ancestral tree of the community and the species labels of the immigrating ancestors. Thus, for the n_i individuals of species i there are a_i ancestors with $a_i \leq n_i$ and for all J individuals there are $A := \sum_{i=1}^S a_i \leq J$ ancestors in total. The second observation is that these ancestors form a sample from the regional species pool where there is only speciation and extinction. For this regional species pool the species-abundance distribution is given by the Ewens sampling formula (ESF) (Ewens 1972; Karlin & McGregor 1972; Tavaré & Ewens 1997; Hubbell 2001). Going forwards in time, each ancestor j of species i can be calculated to have

given rise to $n_{i,j}$ individuals in the local community, and the a_i ancestors of species i gave rise to $\sum_{j=1}^{a_i} n_{i,j} = n_i$ individuals of species i in the local community. Thus, if we had knowledge about each individual's ancestor, we would have a species-ancestry-abundance distribution $D_+ = ((n_{1,1}, \dots, n_{1,a_1}), (n_{2,1}, \dots, n_{2,a_2}), \dots, (n_{S,1}, \dots, n_{S,a_S}))$. We usually do not have this knowledge, so the probability of our species-abundance dataset D being observed is a sum over all possible species-ancestry-abundance datasets D_+ that are compatible with it. The resulting expression is (Etienne & Olff 2004b, combine eqns 5, 8 and 9)

$$P[D|\theta, m, J] = \sum_{\{D_+\}} \frac{J! \prod_{i=1}^A (i-1)!^{\varphi_i} \prod_{j=1}^J (j-1)!^{\phi_j} I^A \theta^S}{\prod_{i=1}^S \left(\prod_{j=1}^{n_i} j!^{\phi_{i,j}} \phi_{i,j}! \right) \prod_{k=1}^S \psi_k! (I)_j (\theta)_A} \quad (1)$$

In this equation I is the number of immigrants that compete with the local individuals for vacant spots. I is related to the immigration probability m by $I = m(J-1)/(1-m)$; this is similar to the way θ is related to v [$\theta := v(J_M-1)/(1-v)$ in Vallade & Houchmandzadeh (2003) as noted above]. Furthermore, $\phi_{i,j}$ is the number of ancestors of species i that have exactly j descendants in the local community, ϕ_j is the total number of ancestors that have exactly j descendants in the local community, φ_i is the number of species with exactly i ancestors, and ψ_k is the number of species that have the same ancestry-abundance distribution $(n_{i,1}, \dots, n_{i,a_i})$; k is an arbitrary counter of these ancestry-abundance distributions and therefore cannot exceed S . The notation $(x)_y$ is the Pochhammer symbol defined as

$$(x)_y := \prod_{i=1}^y (x+i-1) \quad (2)$$

It is equal to

$$(x)_y = \sum_{i=1}^y \bar{\tau}(y, i) x^i \quad (3)$$

where $\bar{\tau}(y, i)$ is the unsigned Stirling number of the first kind, so $\bar{\tau}(y, i)$ is the i th coefficient in the expansion of $(x)_y$. For special cases of i there are simple expressions for $\bar{\tau}(y, i)$. For example, $\bar{\tau}(y, 1) = (y-1)!$, an equality that I will use below.

Because the set $\{D_+\}$ contains an enormous amount of states D_+ , eqn 1 is very complicated. It can however be greatly simplified as I will show now.

A NEW SAMPLING FORMULA

The new sampling formula, derived from (1) in Appendix A, is given by (recall that $m = I/(I+J-1)$):

$$P[D|\theta, m, J] = \frac{J!}{\prod_{i=1}^S n_i \prod_{j=1}^J \Phi_j!} \frac{\theta^S}{(I)_J} \sum_{\mathcal{A}=S}^J K(D, \mathcal{A}) \frac{I^{\mathcal{A}}}{(\theta)_{\mathcal{A}}} \quad (4)$$

Here, Φ_j is the number of species that have abundance j and $K(D, \mathcal{A})$ is defined as

$$K(D, \mathcal{A}) := \sum_{\{a_1, \dots, a_S\} \mid \sum_{i=1}^S a_i = \mathcal{A}} \prod_{i=1}^S \frac{\bar{s}(n_i, a_i) \bar{s}(a_i, 1)}{\bar{s}(n_i, 1)} \quad (5)$$

where the summation is over $a_i = 1, \dots, n_i$ for all i with the restriction that the a_i sum to \mathcal{A} . More details on $K(D, \mathcal{A})$ are given in Appendix A. The $K(D, \mathcal{A})$ for the example datasets mentioned below, and a computer code to generate them, are in the Supplementary Material supplied with the online version of this paper. As the computation of the $K(D, \mathcal{A})$ involves computationally intensive calculation of the Stirling numbers, it may take up to several hours, depending on the processor, the efficiency of the programming language and code, the desired accuracy and the actual data, particularly the largest abundance. This can be considerably shortened if a table of Stirling numbers is used (which is available from the author upon request).

Equation 4 is a great simplification over eqn 1, because the sum of a complicated expression over an enormous number of states D_+ which are also difficult to enumerate, is

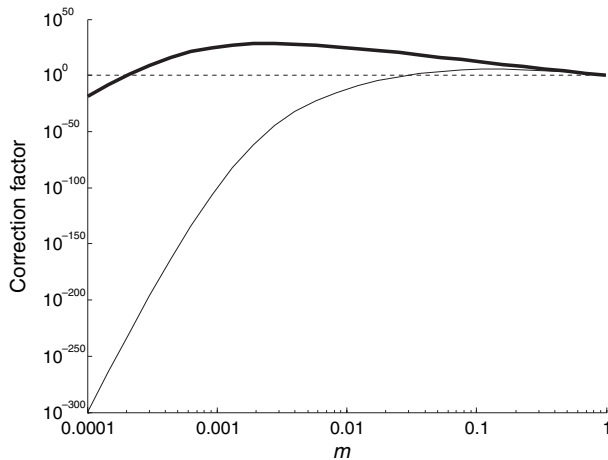


Figure 1 The correction factor of eqn 6 vs. the immigration probability m for the 1982 census of the BCI dataset (thin line) and the CM dataset (thick line). The dotted line is the line for which the correction factor equals 1 and is shown for convenience. For each graph the value of θ is arbitrarily chosen equal to the maximum likelihood estimator of the corresponding dataset (i.e. $\theta_{\text{BCI}} = 48.4$; $\theta_{\text{CM}} = 29.8$, see Table 1). For this choice, the correction factor attains its maximum value at the maximum likelihood estimator of m ($m_{\text{BCI}} = 0.133$; $m_{\text{CM}} = 0.0026$, see Table 1).

replaced by a sum of a relatively simple expression over $J - S + 1$ values of \mathcal{A} .

Rewriting (4) as

$$P[D|\theta, m, J] = \frac{J!}{\prod_{i=1}^S n_i \prod_{j=1}^J \Phi_j!} \frac{\theta^S}{(I)_J} \times \sum_{\mathcal{A}=S}^J \left(K(D, \mathcal{A}) \frac{(\theta)_J}{(\theta)_{\mathcal{A}}} \frac{I^{\mathcal{A}}}{(I)_J} \right) \quad (6)$$

one obtains the ESF multiplied by a correction factor. Without dispersal limitation, that is, for $m = 1$, the model reduces to the ESF, as it should, because I becomes infinite and then there is only one term ($\mathcal{A} = J$) determining the sum in (4) making the correction factor equal to unity. Figure 1 shows the correction factor as a function of m for the two datasets used below.

APPLICATIONS

Parameter estimation

The new sampling formula enables a quick estimation of model parameters for a given dataset, for example, using maximum likelihood. Contrary to previous estimations (Hubbell 2001; McGill 2003a; Volkov *et al.* 2003; Alonso & McKane 2004), estimations with the new sampling formula use all information in the data and take into account that the abundances are correlated due to the zero-sum constraint.

As an illustration, I estimated the model parameters for two datasets: a neotropical tree community on a 50-ha plot on Barro Colorado Island (BCI), Panama (Condit *et al.* 1996, 2002) and a neotropical fish community in Caño Maraca (CM), a creek-floodplain ecosystem in Venezuela (Winemiller 1990). The BCI dataset has been the prime example in treatments of the neutral model (Hubbell 2001; McGill 2003a; Volkov *et al.* 2003; Alonso & McKane 2004; Etienne & Olff 2004b). I used the species-abundance data of each of five censuses (1982, 1985, 1990, 1995 and 2000) as they appeared on <http://ctfs.si.edu/datasets/bci/bcidata/bciN100.html> in September 2004. These data are somewhat different from those reported by Condit *et al.* (1996, 2002), because occasionally errors are found that require correction in old datasets, and because species identifications are sometimes changed (R. Condit, personal communication). I analysed all censuses instead of just one to test whether they yield similar results, as one should expect, and to get an impression of the uncertainty of the estimates. The CM dataset has not served as an example very often in neutral model analyses; to my knowledge only Hubbell (2001) mentions it.

The model parameters were estimated using likelihood maximization or, equivalently, $-\log(\text{likelihood})$ minimization, that is, I used standard numerical optimization algorithms (e.g. Matlab's `FMINSEARCH` command, which uses the Nelder–Mead method) to determine the values of θ and m (or I) for which the likelihood (4) attains its maximum value. It is also possible to derive analytical expressions for these values by setting the partial derivatives of (4) with respect to θ and I to 0 and solving for θ and I . As this solution must be determined numerically, there is no real advantage of this approach over numerical optimization routines. Once the $K(D, A)$ are calculated, likelihood maximization is a matter of seconds to minutes.

The results of the parameter estimation for the two datasets are listed in Table 1. For comparison, the results for the ESF (corresponding to the model without dispersal limitation, $m = 1$) are also listed. From the P -values of the likelihood ratio test between the new sampling formula (with variable m) and the ESF (with fixed m equal to 1) it is clear that the model with dispersal limitation fits much better than the model without dispersal limitation.

As a test of the neutral model, the estimate of the immigration m probability can be compared with an independent estimate using the average dispersal distance of the community members. For BCI, the average dispersal distance is known (39 m, Condit *et al.* 2002). If for simplicity, one assumes the 50 ha BCI plot to be perfectly circular, the dispersal kernel to be exponential and isotropic, and the tree density in the metacommunity pool to be constant and equal to the local tree density, then one can show (see Supplementary Material) that a rough estimate of m is 0.12. Given the numerous simplifications, this is very close to the average of the estimates obtained from the species-abundance distribution ($m \approx 0.11$).

Sequential construction

Distributions (sampling formulas) in the Ewens family correspond to sequential construction schemes, also called urn schemes, that generate samples from the distribution. Hubbell (2001), borrowing from Ewens (1972), presented the scheme that corresponds to the ESF, to generate a species-abundance distribution when there is no dispersal limitation. These urns are known as Hoppe urns in population genetics, after Hoppe (1984, 1987). Likewise, there is a sequential construction scheme for dispersal limited communities, which corresponds to the new sampling formula. It is given in the Supplementary Material in the form of a pseudo code. Actually, it generates the species-ancestry-abundance distribution, given the model parameters θ and m and sample size J . Each individual in the sample is given an ancestry label and a species label according to some specific rules. When all individuals have been labelled, the species-ancestry-abundance distribution can easily be computed by counting all individuals with the same ancestry label and species label. If one ignores the ancestry labels (or effectively sums over the ancestries) and only looks at the species labels, one obtains the species-abundance distribution. The routine was already provided to Alonso & McKane (2004) to compare their expressions with those of Etienne & Olff (2004b), but it is presented here for the first time.

Simpson's diversity index

Simpson's (1949) diversity index is defined as the probability that two individuals randomly sampled from the community belong to different species [its analogue in population genetics is known as Nei's (1987) measure of heterozygosity]. I denote it here by $D_2 := 1 - F_2$ where F_2 is the probability that the two individuals are of the same species.

Table 1 Summary statistics of the Barro Colorado (BCI) tree dataset and the Caño Maraca (CM) fish dataset: number of individuals J , number of species S , number of singleton species T , maximum likelihood parameters $\hat{\theta}$ and \hat{m} , and the corresponding log-likelihood $\ln P$ using the new sampling formula and the ESF (which is equal to the new sampling formula for $m = 1$). The deviance (twice the difference in log-likelihood) between the two models and the corresponding P -value of the χ^2 -distribution with 1 degree of freedom is also given. The new sampling formula gives a significantly better fit than the ESF in all cases

Dataset		Some statistics			New sampling formula			Ewens sampling formula		Model comparison	
Location	Census	J	S	T	$\hat{\theta}$	\hat{m}	$\ln P(D \hat{\theta}, \hat{m}, J)$	$\hat{\theta}$	$\ln P(D \hat{\theta}, J)$	Deviance	P -value
BCI	1982	20 881	238	25	48.4	0.133	-307.5	37.6	-313.0	11.0	9.1×10^{-4}
	1985	20 719	237	29	48.8	0.126	-301.7	37.4	-306.8	10.2	1.4×10^{-3}
	1990	21 233	229	18	48.3	0.100	-302.2	35.8	-312.3	20.2	7.0×10^{-6}
	1995	21 455	227	19	47.8	0.098	-310.1	35.3	-319.8	19.4	1.1×10^{-5}
	2000	21 205	226	14	47.3	0.102	-303.6	35.2	-315.3	23.4	1.3×10^{-6}
CM	1984	28 975	83	7	29.8	0.0026	-256.7	10.4	-266.8	20.1	7.3×10^{-6}

For the neutral model without dispersal limitation, it is simply given by (Hubbell 2001)

$$D_2 := \frac{\theta}{\theta + 1} \quad (7)$$

With dispersal limitation, the formula is only slightly more complicated:

$$D_2 = \frac{\theta}{\theta + 1} \frac{I}{I + 1} \quad (8)$$

The derivation is as follows. For an individual to be of a different species than another, it must have a different immigrating ancestor. This has probability $I/(I + 1)$. Furthermore, the two ancestors must be of different species which has probability $\theta/(\theta + 1)$. Hence, (8) follows. Note also that (8) is closely related to the second step (i.e. $j = 2$) of the sequential construction scheme (see Supplementary Material).

DISCUSSION

I have presented a new sampling formula for the neutral model with dispersal limitation ($m < 1$). Like the ESF, it is called a sampling formula because it applies to a sample from a community and because there is a sequential construction scheme associated with it that can be used to rapidly generate samples for a given set of model parameters. I have developed the sampling formula in the context of community ecology, but it is also applicable in population genetics, or other fields where sampling formulas of the Ewens family are relevant (e.g. macro-economy, Aoki 2002).

Because the new sampling formula is in fact the likelihood, it is ideally suited for parameter estimation procedures and model comparisons that are based upon the likelihood. I illustrated this with a simple example of classical likelihood maximization, but a full Bayesian analysis is possible as well. Etienne & Olff (2004b) already showed results of a Bayesian analysis with the complicated expression (1). This Bayesian analysis can be greatly simplified with the new sampling formula. The Bayesian framework is also a natural setting to compare the neutral model with other proposed models of community structure (e.g. lognormal, McGill 2003a; Volkov *et al.* 2003). This is reported in a companion paper (Etienne & Olff 2005). It addresses the issue whether species-abundance data contain sufficient information to discriminate between several models, which has recently been doubted (Harte 2003; McGill 2003b).

Alonso & McKane (2004) also present a likelihood function based on their mean-field master equation approach to the neutral model, but their expression does not fully obey the zero-sum constraint, and is therefore only

an approximation of which the accuracy is still unknown. Moreover, their expression conditions on the number of species rather than the number of individuals, as it should in the neutral zero-sum framework (Etienne & Olff 2005).

The maximum likelihood estimates obtained for the 1982 census of the BCI dataset are somewhat different from the estimates obtained by Etienne & Olff (2004b). There are three reasons why these results are not comparable. First, the values reported by Etienne & Olff (2004b) represent the posterior mode, which is different from the ML estimate. Secondly, Etienne & Olff (2004b) used an earlier version of the 1982 census, as reported in Condit *et al.* (1996). Thirdly, the estimate of Etienne & Olff (2004b) is based on the complicated expression (1) requiring many latent variables. It is therefore likely that the Markov Chain Monte Carlo simulation on which their estimation is based has not completely converged yet.

In (8) one clearly sees that diversity in a neutral local community is determined by regional richness and immigration. Low diversity can either be due to low regional richness or high dispersal limitation (or both). Simpson's diversity index alone cannot tell us which one of these possibilities is the case. More information is needed. The full species-abundance dataset does provide sufficient information. With the new sampling formula it is now possible to completely disclose this information.

The new sampling formula provides a correct expression for the distribution of abundances in Hubbell's neutral model with dispersal limitation. This model is only a spatially implicit model. A theory, like Hubbell's (2001), that stresses the importance of dispersal in structuring communities, strongly calls for a spatially explicit model as well. Some progress in that direction has already been made (Chave & Leigh 2002). Further development of this spatially explicit model and comparison with the spatially implicit model will teach us the strengths and limitations of the spatially implicit model and hence of the new sampling formula.

ACKNOWLEDGEMENTS

I thank Han Olff, Annette Ostling, Nicholas Gotelli and two anonymous referees for helpful comments on an earlier version of the manuscript, and Emile Apol for useful discussions. Data of the neotropical fish assemblage in Caño Maraca, Venezuela, were kindly provided by K.O. Winemiller. See Winemiller (1990) for full acknowledgement. The Forest Dynamics Plot of Barro Colorado Island has been made possible through the generous support of the US National Science Foundation, The John D. and Catherine T. MacArthur Foundation, and the Smithsonian Tropical Research Institute and through the hard work of over 100 people from 10 countries over the past two

decades. The BCI Forest Dynamics Plot is part the Center for Tropical Forest Science, a global network of large-scale demographic tree plots. Funding for the author was partly provided by the Priority Program 'Biodiversity in Disturbed Ecosystems' of the Netherlands Organization for Scientific Research. Part of the work for this article was carried out at the Tropical Nature Conservation and Vertebrate Ecology group, Wageningen University and Research Centre, Wageningen, The Netherlands, and at Environmental Science, Faculty of Geosciences, Utrecht University, Utrecht, The Netherlands.

SUPPLEMENTARY MATERIAL

The following material is available from <http://www.blackwellpublishing.com/products/journals/suppmat/ELE/ELE717/ELE717sm.htm>

Appendix S1 A pdf-file on the independent calculation of m using dispersal kernels.

Appendix S2 A pdf-file with an explanation and pseudo-code of the sequential construction scheme.

Appendix S3 A zipfile containing a file with source code to compute the $K(D, A)$ for a given dataset and the resulting output files for the datasets used in this article.

Appendix S4 A zipfile containing a file with source code of the sequential construction scheme and some files for illustration of the scheme.

Each zipfile also contains a readme.txt file explaining the contents in more detail.

REFERENCES

- Abrams, P.A. (2001). A world without competition. *Nature*, 412, 858–859.
- Alonso, D. & McKane, A.J. (2004). Sampling Hubbell's neutral theory of biodiversity. *Ecol. Lett.*, 7, 901–910.
- Aoki, M. (2002). Open models of share markets with two dominant types of participants. *J. Econ. Behav. Organ.*, 49, 199–216.
- Bengtsson, J. (2002). The unified neutral theory of biodiversity and biogeography by Hubbell S.P. *Ecol. Econ.*, 42, 497–498.
- Brown, J.H. (2001). Toward a general theory of biodiversity. *Evolution*, 55, 2137–2138.
- Chave, J. & Leigh, E.H. (2002). A spatially explicit neutral model of β -diversity in tropical forests. *Theor. Popul. Biol.*, 62, 153–168.
- Clark, J.S. & McLachlan, J.S. (2003). Stability of forest biodiversity. *Nature*, 423, 635–638.
- Condit, R., Hubbell, S.P. & Foster, R.B. (1996). Changes in tree species abundance in a neotropical forest: impact of climate change. *J. Trop. Ecol.*, 12, 231–256. Data can be downloaded from <http://www.ctfs.si.edu/data/data/data.htm>.
- Condit, R., Pitman, N., Leigh, E.G., Chave, J., Terborgh, J., Foster, R.B. *et al.* (2002). Beta-diversity in tropical forest trees. *Science*, 295, 666–669.
- De Mazancourt, C. (2001). Consequences of community drift. *Science*, 293, 1772.
- Etienne, R.S. & Olff, H. (2004a). How dispersal limitation shapes species – body size distributions in local communities. *Am. Nat.*, 163, 69–83.
- Etienne, R.S. & Olff, H. (2004b). A novel genealogical approach to neutral biodiversity theory. *Ecol. Lett.*, 7, 170–175.
- Etienne, R.S. & Olff, H. (2005). Confronting different models of community structure to species–abundance data: a Bayesian model comparison. *Ecology Letters*, in press.
- Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, 3, 87–112.
- Fargione, J., Brown, C.S. & Tilman, D. (2003). Community assembly and invasion: an experimental test of neutral versus niche processes. *Proc. Natl Acad. Sci. USA*, 100, 8916–8920.
- Harte, J. (2003). Tail of death and resurrection. *Nature*, 424, 1006–1007.
- Hoppe, F. (1984). Pólya-like urns and the Ewens' sampling formula. *J. Math. Biol.*, 20, 91–94.
- Hoppe, F. (1987). The sampling theory of neutral alleles and an urn model in population-genetics. *J. Math. Biol.*, 25, 123–159.
- Hubbell, S.P. (1997). A unified theory of biogeography and relative species abundance and its application to tropical rain forests and coral reefs. *Coral Reefs*, 16, S9–S21.
- Hubbell, S.P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton, NJ.
- Hubbell, S.P., Foster, R.B., O'Brien, S.T., Harms, K.E., Condit, R., Wechsler, B. *et al.* (1999). Light-gap disturbances, recruitment limitation, and tree diversity in a neotropical forest. *Science*, 283, 554–557.
- Hurtt, G.C. & Pacala, S.W. (1995). The consequences of recruitment limitation: reconciling chance, history and competitive differences between plants. *J. Theor. Biol.*, 176, 1–12.
- Karlin, S. & McGregor, J. (1972). Addendum to a paper of W. Ewens. *Theor. Popul. Biol.*, 3, 113–116.
- Loreau, M. & Mouquet, N. (1999). Immigration and the maintenance of local species diversity. *Am. Nat.*, 154, 427–440.
- McGill, B.J. (2003a). A test of the unified neutral theory of biodiversity. *Nature*, 422, 881–885.
- McGill, B.J. (2003b). Strong and weak tests of macroecological theory. *Oikos*, 102, 679–685.
- McKane, A.J., Alonso, D. & Solé, R.V. (2004). Analytic solution of Hubbell's model of local community dynamics. *Theor. Popul. Biol.*, 65, 67–73.
- Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York, NY.
- Ricklefs, R.E. (2003). A comment on Hubbell's zero-sum ecological drift model. *Oikos*, 100, 185–192.
- Simpson, E.H. (1949). Measurement of diversity. *Nature*, 163, 688.
- Tavaré, S. & Ewens, W.J. (1997). Multivariate Ewens distribution. In: *Discrete Multivariate Distributions* (eds Johnson, N.L., Kotz, S. & Balakrishnan, N.). Wiley, New York, NY, pp. 232–246.
- Tilman, D. (1994). Competition and biodiversity in spatially structured habitats. *Ecology*, 75, 2–16.
- Vallade, M. & Houchmandzadeh, B. (2003). Analytical 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.
- Winemiller, K.O. (1990). Spatial and temporal variation in tropical fish tropic networks. *Ecol. Monogr.*, 60, 331–367.

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 27 September 2004

First decision made 30 October 2004

Manuscript accepted 23 November 2004

APPENDIX A. DERIVATION OF THE NEW SAMPLING FORMULA (4) FROM (1)

The simplification of (1) is a purely algebraic simplification, that is, no simplifying assumptions are made. The key element of the simplification is the reduction of the sum over an enormous number of species-ancestry-abundance states to a sum over all possible numbers of ancestors. I distinguish six steps to derive (4) from (1).

First, recall that ϕ_j is the number of ancestors that have j descendants in the local community, and that $\phi_{i,j}$ is the number of ancestors of species i that have j descendants in the local community, so $\phi_j = \sum_{i=1}^S \phi_{i,j}$. Hence,

$$\begin{aligned} \frac{\prod_{j=1}^J (j-1)!^{\phi_j}}{\prod_{i=1}^S \prod_{j=1}^J j!^{\phi_{i,j}}} &= \frac{\prod_{j=1}^J (j-1)!^{\sum_{i=1}^S \phi_{i,j}}}{\prod_{i=1}^S \prod_{j=1}^J j!^{\phi_{i,j}}} \\ &= \frac{\prod_{j=1}^J (j-1)!^{\phi_j}}{\prod_{i=1}^S \prod_{j=1}^J j!^{\phi_{i,j}}} \\ &= \frac{1}{\prod_{i=1}^S \prod_{j=1}^J j!^{\phi_{i,j}}} \end{aligned} \quad (A1)$$

and thus (1) becomes

$$P[D|\theta, m, J] = \sum_{\{D_+\}} \frac{J! \prod_{i=1}^A (i-1)!^{\varphi_i}}{\prod_{i=1}^S \left(\prod_{j=1}^{n_i} j!^{\phi_{i,j}} \right) \prod_{k=1}^S \psi_k! (I)_J(\theta)_A} \frac{I^A \theta^S}{(I)_J(\theta)_A} \quad (A2)$$

Secondly, multiply (A2) by

$$\frac{\prod_{i=1}^S n_i!}{\prod_{i=1}^S n_i(n_i-1)!} \quad (A3)$$

(which evidently equals 1). This results in

$$\begin{aligned} P[D|\theta, m, J] &= \frac{J!}{\prod_{i=1}^S n_i(n_i-1)!} \\ &\times \sum_{\{D_+\}} \frac{\prod_{i=1}^S n_i! \prod_{i=1}^A (i-1)!^{\varphi_i}}{\left(\prod_{j=1}^{n_i} j!^{\phi_{i,j}} \right) \prod_{k=1}^S \psi_k! (I)_J(\theta)_A} \frac{I^A \theta^S}{(I)_J(\theta)_A} \end{aligned} \quad (A4)$$

Thirdly, note that

$$\prod_{i=1}^A (i-1)!^{\varphi_i} = \prod_{i=1}^S (a_i-1)! = \prod_{i=1}^S \bar{\tau}(a_i, 1) \quad (A5)$$

recalling that φ_i is the number of species with exactly i ancestors and a_i is the number of ancestors of species i . Equation A4 becomes, with some minor rearrangements,

$$P[D|\theta, m, J] = \frac{J!}{\prod_{i=1}^S n_i(n_i-1)!} \sum_{\{D_+\}} \frac{\prod_{i=1}^S \frac{n_i! \bar{\tau}(a_i, 1)}{\prod_{j=1}^{n_i} j!^{\phi_{i,j}} \phi_{i,j}!}}{\prod_{k=1}^S \psi_k!} \frac{I^A \theta^S}{(I)_J(\theta)_A} \quad (A6)$$

Fourthly, recalling that ψ_k is the number of species that have the same ancestry-abundance distribution $(n_{i,1}, \dots, n_{i,a})$, observe that, given a particular set of a_1, \dots, a_S ,

$$\begin{aligned} \sum_{\{D_+|a_1, \dots, a_S\}} \frac{1}{\prod_{k=1}^S \psi_k!} \prod_{i=1}^S \left(\frac{n_i!}{\prod_{j=1}^{n_i} j!^{\phi_{i,j}} \phi_{i,j}!} \right) \\ = \frac{1}{\prod_{j=1}^J \Phi_j!} \prod_{i=1}^S \sum_{\{D_+, i|a_i\}} \left(\frac{n_i!}{\prod_{j=1}^{n_i} j!^{\phi_{i,j}} \phi_{i,j}!} \right) \end{aligned} \quad (A7)$$

where Φ_j is the number of species with abundance j . Each summation in the product on the right-hand side is over all ancestry-abundance distributions of species i that have a_i ancestors. Now (A6) reduces to

$$\begin{aligned} P[D|\theta, m, J] &= \frac{J!}{\prod_{i=1}^S n_i(n_i-1)! \prod_{j=1}^J \Phi_j!} \\ &\times \sum_{\{a_1, \dots, a_S\}} \prod_{i=1}^S \sum_{\{D_+, i|a_i\}} \left(\frac{n_i!}{\prod_{j=1}^{n_i} j!^{\phi_{i,j}} \phi_{i,j}!} \right) \bar{\tau}(a_i, 1) \\ &\times \frac{I^A \theta^S}{(I)_J(\theta)_A} \end{aligned} \quad (A8)$$

where the first summation is over all possible combinations of the a_i .

Fifthly, for a moment concentrate on the ancestry distribution of one species i . When the number of ancestors a_i is given, the situation mathematically resembles that of the neutral model without dispersal limitation ($m = 1$) where S is given. For this situation it is known (Tavaré & Ewens 1997) that

$$\sum_{\{D_+, i|a_i\}} \left(\frac{n_i!}{\prod_{j=1}^{n_i} j!^{\phi_{i,j}} \phi_{i,j}!} \right) = \bar{\tau}(n_i, a_i) \quad (A9)$$

and substituting this and the equality $(n_i-1)! = \bar{\tau}(n_i, 1)$ in (A8) yields

$$P[D|\theta, m, J] = \frac{J!}{\prod_{i=1}^S n_i \prod_{j=1}^J \Phi_j!} \times \sum_{\{a_1, \dots, a_S\}} \prod_{i=1}^S \frac{\bar{\tau}(n_i, a_i) \bar{\tau}(a_i, 1)}{\bar{\tau}(n_i, 1)} \frac{I^A \theta^S}{(I)_J (\theta)_A} \quad (\text{A10})$$

Sixthly, define

$$K(D, A) := \sum_{\{a_1, \dots, a_S\} | \sum_{i=1}^S a_i = A} \prod_{i=1}^S \frac{\bar{\tau}(n_i, a_i) \bar{\tau}(a_i, 1)}{\bar{\tau}(n_i, 1)} \quad (\text{A11})$$

where the summation is over $a_i = 1, \dots, n_i$ for all i with the restriction that the a_i sum to A . This means that the $K(D, A)$ are the coefficients in the polynomial

$$\prod_{i=1}^S \sum_{a_i=1}^{n_i} \frac{\bar{\tau}(n_i, a_i) \bar{\tau}(a_i, 1)}{\bar{\tau}(n_i, 1)} x^{a_i} \quad (\text{A12})$$

which enables a relatively simple numerical computation of $K(D, A)$; see the program code in the Supplementary

Material supplied with the online version of this paper. The polynomial (A12) is of order J but the first S orders (including the zeroth order term) have coefficients that are equal to 0; the coefficient of the S th order term, i.e. $K(D, S)$ and the coefficient of the J th order term, i.e. $K(D, J)$ both equal unity, because $\bar{\tau}(y, y) = 1$ for any positive integer y . The definition of $K(D, A)$ leads to

$$\sum_{\{a_1, \dots, a_S\}} \prod_{i=1}^S \frac{\bar{\tau}(n_i, a_i) \bar{\tau}(a_i, 1)}{\bar{\tau}(n_i, 1)} f(A) = \sum_{A=S}^J K(D, A) f(A) \quad (\text{A13})$$

for any function f of A , so (A10) becomes

$$P[D|\theta, m, J] = \frac{J!}{\prod_{i=1}^S n_i \prod_{j=1}^J \Phi_j!} \frac{\theta^S}{(I)_J} \sum_{A=S}^J K(D, A) \frac{I^A}{(\theta)_A} \quad (\text{A14})$$

which is (4).