Journal of Theoretical Biology 257 (2009) 510–514

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Maximum likelihood estimation of neutral model parameters for multiple samples with different degrees of dispersal limitation

Rampal S. Etienne *

Community and Conservation Ecology Group, Centre for Ecological and Evolutionary Studies, University of Groningen, P.O. Box 14, 9750 AA Haren, The Netherlands

ARTICLE INFO

Article history:
Received 17 October 2008
Received in revised form
8 December 2008
Accepted 8 December 2008
Available online 30 December 2008

Keywords:
Ewens sampling formula
Etienne sampling formula
Maximum likelihood
Fundamental biodiversity number
Fundamental dispersal number

ABSTRACT

In a recent paper, I presented a sampling formula for species abundances from multiple samples according to the prevailing neutral model of biodiversity, but practical implementation for parameter estimation was only possible when these samples were from local communities that were assumed to be equally dispersal limited. Here I show how the same sampling formula can also be used to estimate model parameters using maximum likelihood when the samples have different degrees of dispersal limitation. Moreover, it performs better than other, approximate, parameter estimation approaches. I also show how to calculate errors in the parameter estimates, which has so far been largely ignored in the development of and debate on neutral theory.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

In a recent paper (Etienne, 2007), I presented a sampling formula for the joint probability of a data set of species abundances in multiple local samples. This sampling formula assumes the prevailing spatially implicit neutral model of biodiversity (Hubbell, 2001; Volkov et al., 2003; Etienne, 2005) where local communities draw immigrants from a metacommunity that is in a balance between speciation and extinction. In that paper I stated that the formula was only applicable in practice for maximum likelihood estimation of model parameters if all samples are assumed to be equally dispersal limited, that is, that they have the same values of the fundamental dispersal number I_i; I_i is related to the immigration probability m_i by $I_i = m_i/(1-m_i)(J_i-1)$, see Etienne and Alonso (2005). I refer to Tomašových (2008) for an application. The reason for this limited applicability was that only under the assumption of equal dispersal limitation the formula, which involved a very large number of sums, simplified to something computationally tractable. Here I show that the formula also simplifies to a computationally tractable (albeit still demanding) form even if the assumption on the fundamental dispersal number I is dropped. This allows simultaneous maximum likelihood estimation of the fundamental biodiversity number θ and each of the fundamental dispersal numbers I_i for each sample i. The utility of the sampling formula is, thus, substantially extended. Furthermore I demonstrate that the maximum likelihood parameter estimation based on the sampling formula outperforms other, approximate, approaches that have been developed in the meantime. Finally, I note that not only the neutral model parameters themselves can be estimated, but also the errors in the parameters.

Jabot et al. (2008) pointed out that m and I actually do not, just, represent a measure of dispersal, but of recruitment which encompasses both dispersal and establishment. Only if establishment is assumed identical for both immigrant and local individuals, then m and I can be interpreted as measures of dispersal (limitation). From hereon I will assume the broader interpretation in terms of recruitment. Consistency then requires to call I the fundamental recruitment number.

As a final remark in this introduction, I would like to point out that different I-values for samples from different geographic locations is not in contradiction with the neutrality assumption, because individuals of different species are still functionally equivalent. This function is now made dependent on the geographic location, but it is still the same for all species in the same geographic location. In more abstract terms: individuals in the same location are exchangeable, but different locations are not (see also Etienne, 2007).

2. The sampling formula

Suppose that there are N samples from N different local communities, each of which contains J_i individuals ($i = 1,\ldots,N$), summing to a total of J individuals in all samples together and a total of S different species. The N samples sizes can be

* Tel.: +31503632230.
E-mail address: r.s.etienne@rug.nl

0022-5193/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2008.12.016

summarized by the vector $\vec{J} = (J_1, \ldots, J_N)$. The species found in these samples are indicated by an arbitrary order $k = 1, \ldots, S$ and the data set of all species abundances $\vec{D}$ can be written as a vector of vectors $\vec{D} = (\vec{D}_1, \ldots, \vec{D}_N) = ((n_{11}, \ldots, n_{1S}), \ldots, (n_{N1}, \ldots, n_{NS}))$ where n_{ik} represents the number of individuals of species k in sample i. Given θ and $\vec{I} = (I_1, \ldots, I_N)$, the sampling formula for such a data set reads (Etienne, 2007)

$$P[\vec{D}|\vec{I},\theta,\vec{J}] = \frac{1}{\prod_j \Phi_j!}\left(\prod_{i=1}^{N}\frac{J_i!}{(I_i)_{J_i}\prod_{k=1}^{S} n_{ik}!}\right) \times \sum_{\{a_{11},\ldots,a_{NS}\}}\left(\prod_{k=1}^{S}\left((a_k-1)!\prod_{i=1}^{N}\bar{s}(n_{ik},a_{ik})\right)\prod_{i=1}^{N} I_i^{A_i}\right)\frac{\theta^S}{(\theta)_A} \quad (1)$$

where Φ_j is the number of species that have abundance vector $\vec{j}$ across the samples, $\bar{s}(x,y)$ denotes the unsigned Stirling number of the first kind and $(x)_y$ denotes the Pochhammer notation (Etienne, 2005, 2007). Furthermore, I have defined $A_i = \sum_k a_{ik}$ and $a_k = \sum_i a_{ik}$ and $A = \sum_i A_i = \sum_{i,k} a_{ik} = \sum_k a_k$. Eq. (1) assumes that species are not labeled, but samples are, the most common use of abundance distributions. Different assumptions on the labeling only affect the prefactor; see Etienne (2007) for more details. The sampling formula can serve as a likelihood in maximum likelihood parameter estimation (Etienne, 2007).

3. Simplification of the sampling formula

The definition of A_i is crucial in the simplification, because with it we can write

$$\begin{aligned} P[\vec{D}|\vec{I},\theta,\vec{J}] &= \frac{1}{\prod_j \Phi_j!}\left(\prod_{i=1}^{N}\frac{J_i!}{(I_i)_{J_i}\prod_{k=1}^{S} n_{ik}!}\right) \times \sum_{\{a_{11},\ldots,a_{NS}\}}\left(\prod_{k=1}^{S}\left((a_k-1)!\prod_{i=1}^{N}\bar{s}(n_{ik},a_{ik})\right)\prod_{i=1}^{N} I_i^{\sum_{k=1}^{S} a_{ik}}\right)\frac{\theta^S}{(\theta)_A} \\ &= \frac{1}{\prod_j \Phi_j!}\left(\prod_{i=1}^{N}\frac{J_i!}{(I_i)_{J_i}\prod_{k=1}^{S} n_{ik}!}\right) \times \sum_{\{a_{11},\ldots,a_{NS}\}}\left(\prod_{k=1}^{S}\left((a_k-1)!\prod_{i=1}^{N}\bar{s}(n_{ik},a_{ik})\right) \times \prod_{i=1}^{N}\left(\prod_{k=1}^{S} I_i^{a_{ik}}\right)\right)\frac{\theta^S}{(\theta)_A} \end{aligned} \quad (2)$$

where in the first line I have simply substituted this definition and in the second line I have factored out the a_{ik} in the exponent. Note now that we have two products over i and k after the summation. This can be simplified to a single product

$$P[\vec{D}|\vec{I},\theta,\vec{J}] = \frac{1}{\prod_j \Phi_j!}\left(\prod_{i=1}^{N}\frac{J_i!}{(I_i)_{J_i}\prod_{k=1}^{S} n_{ik}!}\right) \times \sum_{\{a_{11},\ldots,a_{NS}\}}\left(\prod_{k=1}^{S}\left((a_k-1)!\prod_{i=1}^{N}\bar{s}(n_{ik},a_{ik})I_i^{a_{ik}}\right)\right)\frac{\theta^S}{(\theta)_A} \quad (3)$$

One can write this more compactly as

$$P[\vec{D}|\vec{I},\theta,\vec{J}] = \frac{1}{\prod_j \Phi_j!}\left(\prod_{i=1}^{N}\frac{J_i!}{(I_i)_{J_i}\prod_{k=1}^{S} n_{ik}!}\right)\sum_A M(\vec{D},A,\vec{I})\frac{\theta^S}{(\theta)_A} \quad (4a)$$

where

$$M(\vec{D},A,\vec{I}) := \sum_{\{a_{11},\ldots,a_{NS}|\sum_{i,k} a_{ik}=A\}}\prod_{k=1}^{S} \times\left((a_k-1)!\prod_{i=1}^{N}\bar{s}(n_{ik},a_{ik})I_i^{a_{ik}}\right) \quad (4b)$$

Compare this to

$$P[\vec{D}|\vec{I},\theta,\vec{J}] = \frac{1}{\prod_j \Phi_j!}\left(\prod_{i=1}^{N}\frac{J_i!}{\prod_{k=1}^{S} n_{ik}!}\right)\sum_A M(\vec{D},A)\frac{I^A\theta^S}{(I)_J(\theta)_A} \quad (5)$$

where

$$M(\vec{D},A) := \sum_{\{a_{11},\ldots,a_{NS}|\sum_{i,k} a_{ik}=A\}}\prod_{k=1}^{S}\left((a_k-1)!\prod_{i=1}^{N}\bar{s}(n_{ik},a_{ik})\right) \quad (6)$$

for the case where all I are equal (Etienne, 2007). The main difference is that $I_i^{a_{ik}}$ appears in $M(\vec{D},A,\vec{I})$ but this has only a minor additional computational cost: instead of $\bar{s}(n_{ik},a_{ik})$ for each value a_{ik} in the sum one needs to compute $\bar{s}(n_{ik},a_{ik})I_i^{a_{ik}}$. In maximum likelihood parameter estimation this minor additional computational cost is not negligible because in finding the optimal parameter values $M(\vec{D},A,\vec{I})$ needs to be evaluated every time the parameter values change whereas $M(\vec{D},A)$ only needed to be evaluated once, at the beginning of the optimization procedure. Also, the fact that a_{ik} appears in the exponent of I_i, which is potentially a large number, may incur numerical problems. With the software used (PARI/GP) numerical problems did not occur unless m_i was very close to 1 (I_i very large). The code for maximum likelihood parameter estimation can be found in the online appendix of this paper. It uses the simplex method to find the likelihood optimum. This method is relatively good at finding the global likelihood optimum, but with a high-dimensional parameter space, it is crucial to rerun the optimization algorithm with different initial values to search for the global optimum.

4. Estimation of the errors in the parameters

The maximum likelihood method also allows for computation of the standard error in the estimates by means of the variance–covariance matrix at the likelihood optimum, The variance–covariance matrix M at the likelihood optimum (where $\partial \ln P/\partial\theta = \partial \ln P/\partial I_i = 0$) is the inverse of the observed information matrix I_o which in turn is a matrix of second order derivatives of the loglikelihood evaluated at the optimum: for example, for two samples we have three parameters (θ, I_1 and I_2) and the following variance–covariance matrix:

$$M = I_o^{-1} = \begin{pmatrix} -\frac{\partial^2 \ln P}{\partial\theta^2} & -\frac{\partial^2 \ln P}{\partial\theta\,\partial I_1} & -\frac{\partial^2 \ln P}{\partial\theta\,\partial I_2} \\ -\frac{\partial^2 \ln P}{\partial I_1\,\partial\theta} & -\frac{\partial^2 \ln P}{\partial I_1^2} & -\frac{\partial^2 \ln P}{\partial I_1\,\partial I_2} \\ -\frac{\partial^2 \ln P}{\partial I_2\,\partial\theta} & -\frac{\partial^2 \ln P}{\partial I_2\,\partial I_1} & -\frac{\partial^2 \ln P}{\partial I_2^2} \end{pmatrix}^{-1} \quad (7)$$

Square roots of the diagonal elements (the variances) are the standard errors for the three parameters. The off-diagonal elements (the covariances) provide information on the correlation structure of the estimated parameters. The online material also includes code for estimation of the errors. One can obtain the correlation matrix by dividing each element by the square root of the product of the variances of the two parameters corresponding to that element.

To compute the errors for m_i rather than I_i one needs to perform the following transformation:

$$\frac{\partial^2 \ln P}{\partial\theta\,\partial m_i} = \frac{\partial^2 \ln P}{\partial\theta\,\partial I_i}\frac{\partial I_i}{\partial m_i} \quad (8a)$$

$$\frac{\partial^2 \ln P}{\partial m_i\,\partial m_j} = \frac{\partial^2 \ln P}{\partial I_i\,\partial I_j}\frac{\partial I_i}{\partial m_i}\frac{\partial I_j}{\partial m_j} \quad (8b)$$

In (8b) there are no first order derivatives because I_i only depends on m_i (not on m_j for $j \neq i$) and $\partial \ln P / \partial m_i = 0$ at the likelihood optimum.

5. Results

To examine how well the maximum likelihood estimation based on (4a) performs, I first simulated data sets of three samples with 1000 individuals each using known parameters (see Etienne, 2007 for the algorithm) for various parameter combinations. Then I estimated the parameters using the one-stage (i.e. estimating all parameters at once) approach based on (4a) and using the (approximate) two-stage approach (i.e. first θ is estimated and then the I_i conditional on θ) of Etienne (2009). The latter is an improved version of the two-stage approach of Munoz et al. (2007). Table 1 has the means and coefficients of variation of the maximum likelihood estimates across the 1000 data sets for each parameter combination. While the mean tells us something about the bias of the estimation method (the larger the difference between this mean and the true parameter value, i.e. the value with which the data were generated, the larger the bias), the coefficient of variation informs us about how far away a parameter estimate for an *individual* data set can be from the true value (the larger the c_v, the larger the average individual deviation from the true value); an estimation method can, thus, be unbiased but still be inaccurate for an individual data set, or biased yet accurate when corrected for bias. Clearly, the one-stage approach outperforms the two-stage approach, not only because it produces less biased results, but also because the coefficients of variation are substantially smaller.

As an example of the estimation of the errors in the parameters I reanalyzed the tropical forest data set also used as an example in Etienne (2007). This data set consists of three Panamanian forest plots (Condit et al., 2002): Sherman (5.96 ha of which 5 ha is in the data file), Barro Colorado Island (50 ha) and Cocoli (4 ha). These plots lie along a precipitation gradient (3030 mm/yr, 2616 mm/yr and 1950 mm/yr, respectively, Condit et al., 2004) which may cause them to have very different degrees of recruitment limitation (Jabot et al., 2008). The new methods presented in this paper can help to identify whether they indeed have different degrees of recruitment limitation. I find that BCI has less recruitment limitation than Sherman and Cocoli which are equally recruitment-limited, from which one may conclude that two *I*-values (together with θ) sufficiently describe this data set (Table 2). This result is qualitatively consistent with the estimates based on the two-stage approach (Etienne, 2009). BCI's central location may explain its higher value of *I*. The correlation matrix shows that the estimates for the I_i are not correlated with one another, but they are (strongly) correlated with θ as expected (Etienne, 2005).

Table 2 also contains estimates for the three tree communities where instead of the full BCI plot only a 5 ha subplot is taken (see Etienne, 2007). This has no substantial effect on the parameter estimates which demonstrates the sample size independence of *I* in contrast to *m*.

The time to compute the ML parameters with the above mentioned software depends on three types factors. 1. Environment-related factors: CPU, platform (Windows, Linux), PARI/GP version, compiled or uncompiled (i.e. interpreted) code. 2. Likelihood-optimization-related: initial values used in the optimization, tolerance allowed for the function to be optimized and the parameters. 3. Data-related: number of samples, number of species, number of individuals. As an illustration, the time needed to compute one likelihood value for the tree communities in Panama with the first subsample of BCI was 39 seconds on a 3 GHz

Table 1
Estimates of θ and m_i in various scenarios of simulated data sets for the two-stage approach of Etienne, 2009 and the approach presented here.

Scenario	Model parameters						Maximum likelihood parameter estimation															
	J	θ	m_1	m_2	m_3		Etienne (2008)						This paper									
							$\hat{\theta}$		$\hat{m}_1$		$\hat{m}_2$		$\hat{m}_3$		$\hat{\theta}$		$\hat{m}_1$		$\hat{m}_2$		$\hat{m}_3$	
							Mean	c_v	Mean	c_v	Mean	c_v	Mean	c_v	Mean	c_v	Mean	c_v	Mean	c_v	Mean	c_v
1	1000	5	0.1	0.2	0.4		5.8096	0.39	0.2003	1.45	0.2765	1.16	0.3975	0.91	4.9689	0.21	0.1119	0.44	0.2353	0.49	0.4727	0.50
2	1000	50	0.1	0.2	0.4		51.8122	0.19	0.1135	0.69	0.2224	0.56	0.4248	0.50	49.9838	0.097	0.1022	0.16	0.2041	0.16	0.4105	0.18
3	1000	500	0.1	0.2	0.4		507.5497	0.12	0.1005	0.088	0.2009	0.089	0.4026	0.11	501.5142	0.067	0.1005	0.08	0.2009	0.077	0.4007	0.076
4	1000	5	0.01	0.05	0.25		5.8089	0.46	0.0438	3.61	0.1319	1.84	0.3460	1.05	4.8982	0.25	0.0108	0.43	0.0572	0.46	0.3658	0.70
5	1000	50	0.01	0.05	0.25		53.6618	0.29	0.0103	0.23	0.0572	0.76	0.2950	0.70	49.9892	0.12	0.0103	0.21	0.0513	0.16	0.2643	0.25
6	1000	500	0.01	0.05	0.25		577.0717	0.36	0.0100	0.17	0.0505	0.14	0.2615	0.34	504.0792	0.11	0.0101	0.17	0.0504	0.11	0.2521	0.091
7	1000	5	0.009	0.09	0.9		5.9486	0.46	0.0374	3.51	0.1854	1.50	0.5619	0.72	5.1082	0.23	0.0098	0.42	0.1019	0.40	0.7720	0.34
8	1000	50	0.009	0.09	0.9		53.8231	0.28	0.0091	0.22	0.1032	0.75	0.7302	0.38	50.5992	0.10	0.0091	0.20	0.0906	0.15	0.8647	0.16
9	1000	500	0.009	0.09	0.9		555.6823	0.31	0.0091	0.18	0.0907	0.14	0.8181	0.20	503.8535	0.075	0.0090	0.18	0.0901	0.090	0.8975	0.081
10	1000	5	0.001	0.002	0.004		10.5145	1.92	0.0217	5.96	0.0427	4.32	0.0617	3.48	5.0388	0.45	0.0012	0.67	0.0027	1.27	0.0066	4.85
11	1000	50	0.001	0.002	0.004		1533.1730	11.70	0.0010	0.51	0.0039	8.93	0.0105	6.43	56.0378	0.55	0.0010	0.42	0.0020	0.35	0.0042	0.30

The values reported are the means and coefficients of variations (c_v) of the parameter estimates over 1000 simulated data sets, each having three samples of size 1000. There are no results listed for $\theta = 500$ and $m_1 = 0.001$, $m_2 = 0.002$, $m_3 = 0.004$ because this configuration frequently results in an abundance data set in which there is no species overlap between samples and, thus, has an infinite ML estimate for θ.

Table 2
Estimates of θ and I_i for the three tropical tree communities in Panama.

Sherman + BCI + Cocoli	Sample sizes and species richness		Maximum likelihood parameter estimation			
	$\vec{J}$	$\vec{S}$	$\widehat{\theta}$	$\widehat{I}_{\text{Sherman}}$	$\widehat{I}_{\text{BCI}}$	$\widehat{I}_{\text{Cocoli}}$
	(2860, 21457, 1079)	(125, 225, 99)	235 ± 23	35.7 ± 3.9	65.3 ± 5.9	31.5 ± 3.9
Sherman + BCI_1 + Cocoli	(2860, 2359, 1079)	(125, 152, 99)	260 ± 29	35.6 ± 3.9	54.2 ± 5.8	30.7 ± 3.7
Sherman + BCI_2 + Cocoli	(2860, 2151, 1079)	(125, 150, 99)	264 ± 30	35.5 ± 3.9	54.4 ± 5.9	30.9 ± 3.8
Sherman + BCI_3 + Cocoli	(2860, 2076, 1079)	(125, 162, 99)	265 ± 29	35.0 ± 3.8	63.5 ± 6.7	31.1 ± 3.8
Sherman + BCI_4 + Cocoli	(2860, 2027, 1079)	(125, 171, 99)	264 ± 29	34.8 ± 3.8	70.5 ± 7.4	31.1 ± 3.8
Sherman + BCI_5 + Cocoli	(2860, 2000, 1079)	(125, 166, 99)	274 ± 30	34.4 ± 3.7	66.9 ± 7.1	31.3 ± 3.8
Sherman + BCI_6 + Cocoli	(2860, 2050, 1079)	(125, 153, 99)	286 ± 32	33.9 ± 3.6	56.1 ± 6.0	31.1 ± 3.8
Sherman + BCI_7 + Cocoli	(2860, 2364, 1079)	(125, 147, 99)	291 ± 33	33.9 ± 3.6	48.2 ± 5.0	30.8 ± 3.8
Sherman + BCI_8 + Cocoli	(2860, 2225, 1079)	(125, 138, 99)	291 ± 34	34.2 ± 3.7	44.8 ± 4.8	30.8 ± 3.8
Sherman + BCI_9 + Cocoli	(2860, 2076, 1079)	(125, 145, 99)	292 ± 34	34.3 ± 3.7	50.1 ± 5.3	30.7 ± 3.7
Sherman + BCI_{10} + Cocoli	(2860, 2129, 1079)	(125, 157, 99)	260 ± 29	35.0 ± 3.8	59.3 ± 6.3	31.3 ± 3.8

Sherman + BCI + Cocoli		Correlation matrix			
		$\widehat{\theta}$	$\widehat{I}_{\text{Sherman}}$	$\widehat{I}_{\text{BCI}}$	$\widehat{I}_{\text{Cocoli}}$
	$\widehat{\theta}$	1	−0.12	−0.37	−0.052
	$\widehat{I}_{\text{Sherman}}$	−0.12	1	0.070	0.0021
	$\widehat{I}_{\text{BCI}}$	−0.37	0.070	1	0.031
	$\widehat{I}_{\text{Cocoli}}$	−0.052	0.0021	−0.031	1

The first row reports the values for the full three data sets; the 10 following rows report the values for each of the 10 subplots of BCI (see also Etienne, 2007). The last part is the correlation matrix for the full data set.

Pentium 4 running uncompiled code in PARI/GPversion 2.3.3 under Windows XP, whereas it took 14 seconds on a single 2 GHz AMD64 node of a cluster running compiled code in PARI/GP version 2.3.4 under Linux.

6. Discussion

I have derived a computationally tractable sampling formula for multiple samples of species abundances, assuming the most widely used spatially implicit neutral model of biodiversity. It does not need the assumption of Etienne (2007) that all samples are equally recruitment-limited (that is, have the same I-value). Maximum likelihood parameter estimation based on this sampling formula can be done for all parameters simultaneously (i.e. it is a one-stage approach in the terminology of Munoz et al., 2007) and outperforms the two-stage approach developed by Munoz et al. (2007) and Etienne (2009) by having less bias and being more accurate (i.e. individual estimates are unlikely to deviate much from the true values).

Because the one-stage approach searches simultaneously for all the parameters that optimize the likelihood, it has another advantage: it potentially recognizes multiple likelihood optima (Etienne et al., 2006). As the number of samples increases, it is unlikely that these optima are similar (and, thus, a clear global optimum exists), because then there is more information in the data on θ (as θ reflects beta diversity). One may find the global likelihood optimum by choosing different sets of starting values of the optimization routine. In contrast, the two-stage approach can only find a single set of parameter estimates which do not necessarily correspond to the global likelihood optimum, although, as stated, the chances that it is far away from the global optimum probably get smaller when the number of samples increase. In any case, the two-stage approach is still useful: it can provide good starting values for the one-stage approach (which otherwise takes long to converge onto the optimum) and, in contrast to the one-stage approach, it remains computationally efficient even when the number of samples becomes large.

Recently, two other approaches to estimate neutral model parameters from species abundances have been put forward. The first approach is by Forster and Warton (2007). They derive a integral likelihood for multiple samples, but this likelihood is less informative because, by being a product over the probabilities for each species' abundance, it conditions on the total number of species as well as on the sample sizes. The sampling formula present in this paper only conditions on the sample sizes and the total number of species is a prediction rather than an assumption. Also, the estimation procedure of Forster and Warton (2007) is, as they state, fraught with numerical problems in evaluating the integral, notwithstanding the fact that they have found clever ways to minimize them. The second approach is by Jabot et al. (2008) who dispense with the metacommunity model altogether and only estimate the I_i assuming the aggregated abundances across all samples as a proxy for the metacommunity abundance distribution. When the number of samples is small or when there are many singletons, this assumption is hard to justify. There is a third approach to estimating neutral model parameters (Munoz et al., 2008) based on the same spatially implicit model, but this approach uses similarity measures similar to Simpson diversity (see also He, 2005) rather than the full abundance vector.

Not only can the sampling formula be used for parameter estimation without the restricting assumption of equal recruitment limitation across all local communities, it is also applicable in the "exact" test of neutrality proposed in Etienne (2007). Furthermore, by being a proper likelihood it enables direct likelihood-based comparisons of the performance of different models of community structure in fitting species abundance data at multiple sites, ranging from model weighting using AIC (Chave et al., 2006; Etienne et al., 2007) to Bayesian comparisons (Etienne and Olff, 2005).

Specifying error estimates will help in interpretation of parameter estimates as in the tropical tree community example. Surprisingly, this has not received much attention in

the development of tools in evaluating the neutral theory of biodiversity. In Etienne (2007), I showed that an estimate of the uncertainty in the parameters can also be obtained by parametric bootstrap (which can also be used to test for neutrality): one simulates many data sets with the ML estimates obtained from the real data and then estimates the ML parameters for each of these simulated data sets (Efron and Tibshirani, 1993); the distribution of these ML estimates informs one about bias and variance in the ML estimates for the real data (see also Burnham and Anderson, 2002). Because this is computationally demanding, the variance–covariance matrix at the likelihood optimum provides a convenient alternative, although it does not give an estimate of the bias. With these two procedures now being available specifying error estimates should become common practice in confrontations of neutral models to diversity data.

Acknowledgments

I thank Franck Jabot and one anonymous reviewer for their helpful comments. Financial support was provided by The Netherlands Organisation for Scientific Research (NWO). Part of the work for this paper was done while I was a Courtesy Research Associate at the University of Oregon.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jtbi.2008.12.016.

References

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference. A Practical Information Theoretic Approach. Springer, NY.

Chave, J., Alonso, D., Etienne, R.S., 2006. Comparing models of species abundance. Nature 441, E1–E2.

Condit, R., Aguilar, S., Hernandes, A., Perez, R., Lao, S., Angehr, G., Hubbell, S., Foster, R., 2004. Tropical forest dynamics across a rainfall gradient and the impact of an El Niño dry season. Journal of Tropical Ecology 20, 51–72.

Condit, R., Pitman, N., Leigh, E.G., Chave, J., Terborgh, J., Foster, R.B., Nunez, P., Aguilar, S., Valencia, R., Villa, G., Muller-Landau, H.C., Losos, E., Hubbell, S.P., 2002. Beta-diversity in tropical forest trees. Science 295, 666–669.

Efron, B., Tibshirani, R.J., 1993. An introduction to the bootstrap. Monographs on Statistics and Applied Probability 57, Chapman & Hall, NY.

Etienne, R.S., 2005. A new sampling formula for neutral biodiversity. Ecology Letters 8, 253–260.

Etienne, R.S., 2007. A neutral sampling formula for multiple samples and an "exact" test of neutrality. Ecology Letters 10, 608–618.

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

Etienne, R.S., Alonso, D., 2005. A dispersal-limited sampling theory for species and alleles. Ecology Letters 8, 1147–1156.

Etienne, R.S., Apol, M.E.F., Olff, H., Weissing, F.J., 2007. Modes of speciation and the neutral theory of biodiversity. Oikos 116, 241–258.

Etienne, R.S., Latimer, A.M., Silander, J.A., Cowling, R.M., 2006. Comment on "neutral ecological theory reveals isolation and rapid speciation in a biodiversity hot spot". Science 311, 610b.

Etienne, R.S., Olff, H., 2005. Confronting different models of community structure to species-abundance data: a Bayesian model comparison. Ecology Letters 8, 493–504.

Forster, M., Warton, D., 2007. A metacommunity-scale comparison of species-abundance distribution models for plant communities of eastern Australia. Ecography 30, 449–458.

He, F.L., 2005. Deriving a neutral model of species abundance from fundamental mechanisms of population dynamics. Functional Ecology 19, 187–193.

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

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

Munoz, F., Couteron, P., Ramesh, B.R., 2008. Beta diversity in spatially implicit neutral models: a new way to assess species migration. American Naturalist 172, 116–127.

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

Tomašových, A., 2008. Evaluating neutrality and the escalation hypothesis in brachiopod communities from shallow high-productivity habitats. Evolutionary Ecology Research 10, 667–698.

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