Improved estimation of neutral model parameters for multiple samples with different degrees of dispersal limitation
Etienne, Rampal S.

Published in:
Ecology

DOI:
10.1890/08-0750.1

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2009

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Improved 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

Abstract. The standard neutral model of biodiversity, as introduced by S. P. Hubbell, is currently increasingly used as a null model for the structure of ecological communities. In such applications, estimation of the model parameters is essential. An exact maximum likelihood approach has been developed for data sets consisting of multiple, spatially separated, samples of species abundances. This approach is only computationally tractable when it is assumed that all these samples have the same amount of dispersal limitation. Recently, an approximate approach has been proposed that does not require this assumption. However, this approach cannot estimate the fundamental biodiversity number \( h \) when there are only a few samples or many, very different, samples. In this note, I present a modification of this approximate approach that does not suffer from this shortcoming. I illustrate it with simulated and real data sets.

Key words: Etienne sampling formula; Ewens sampling formula; fundamental biodiversity number; fundamental dispersal number; maximum likelihood.

INTRODUCTION

The now well-known neutral theory of biodiversity (Hubbell 2001) is increasingly being used as a null model for the structure of ecological communities, particularly for the distribution of abundances of the species in a community (Alonso et al. 2006; see Gotelli and McGill 2006 for a discussion of neutral vs. null models). The most widely used model implementation of this theory is a mainland–island model where the local community receives immigrants from the regional community (the metacommunity), which itself is governed by speciation and extinction. Both local and regional abundances vary stochastically, a process termed ecological drift (Hubbell 2001). The model contains two parameters: the fundamental biodiversity parameter \( \theta \) and the fundamental dispersal or immigration parameter \( I \) (Etienne and Alonso 2005). The former parameter summarizes the diversity of the metacommunity community and is a composite of the metacommunity size \( J_M \) and the speciation rate \( v \). The latter parameter measures the amount of dispersal limitation of the local community and may be interpreted as the number of potential immigrants competing with local individuals for vacant sites (resources) (Etienne and Olff 2004a). When dealing with samples of species abundances, the parameter \( m \) has often been used instead of \( I \), where \( m = I/(I + J - 1) \) with \( J \) the total number of individuals in the species abundance data set. This parameter may be interpreted as the probability that any of the \( I \) immigrants will occupy the vacant site in a community of size \( J \).

Applications of the model often involve estimation of these parameters from samples of species abundances because these parameters may provide new and relatively quick insights into the ecology of a system, even if neutrality is violated (Latimer et al. 2005, Jabot et al. 2008). Exact sampling formulae have been developed with which these parameters can be simultaneously estimated using maximum likelihood (Etienne and Olff 2004b, Etienne 2005, 2007). These exact formulae are tractable in practice only when there is a single sample, or when there are multiple, spatially separated samples (i.e., from different local communities connected to the same metacommunity) which are assumed to have the same degree of dispersal limitation. That is, only a single value of \( I \) can be estimated. Exact and simultaneous estimation of a separate value for \( I \) for each local sample is currently only theoretically possible (Etienne 2007). Fortunately, an approximate approach has been proposed that does not require this assumption (Munoz et al. 2007). It is approximate in the sense that it does not
estimate the parameters ($\theta$ and the various $I$) simultaneously and it needs Monte Carlo samples from the data set to estimate $\theta$. Unfortunately, this approach does not work when there are only a few samples available, or when there are many samples that are very different. Here I show why it does not work, and I present a modification of this approach that solves this problem. I illustrate it with simulation data and real data. A program code is provided in the online material.

**IMPROVED TWO-STAGE ESTIMATION**

Denote the species abundance data set from which we want to infer the model parameters by the vector $n = (n_{11}, \ldots, n_{1k}, n_{21}, \ldots, n_{2k}, \ldots, n_{N1}, \ldots, n_{NK})$: Here, $N$ is the number of local samples, $S$ is the total number of different species in the entire data set, and $n_{ik}$ is the abundance of species $k$ in local sample $i$. The two-stage estimation approach of Munoz et al. (2007) consists of the following two steps.

1. **Munoz et al. estimation of $\theta$**—Repeat the following procedure $M$ times and denote each iteration with the subscript $j$. Take a single individual from each local sample, record its species, and estimate $\hat{\theta}_j$ from the resulting species abundance data set of $N$ individuals as follows. Because each local sample supplies only a single individual, the $N$ individuals thus collected can be regarded as a sample from the metacommunity (they have different immigrating ancestors [Etienne and Olff 2004b, Etienne 2007]). Because the metacommunity is governed by the Ewens sampling formula (Ewens 1972, Hubbell 2001), this sampling formula can be used to estimate $\theta_j$. Denoting the abundance vector of the $N$ individuals in the $j$th iteration by $N_j$, the Ewens sampling formula reads for this case as follows:

$$P[N_j | \theta_j, N] = F_j \frac{\theta_j^{N_j}}{\Gamma(\theta_j + N)}$$

(1)

where $F_j$ is a prefactor that does not depend on $\theta$, $S$ is the number of species in the metacommunity sample, which is usually less than the total number of species $S$, and $(\theta_j)_N$ is the Pochhammer notation,

$$(\theta_j)_N = \prod_{s=1}^{N} \left( \theta_j + s - 1 \right) = \frac{\Gamma(\theta_j + N)}{\Gamma(\theta_j)}$$

(2)

where $\Gamma$ refers to the Gamma function. It can be easily shown (Tavaré and Ewens 1997) that the maximum
likelihood estimate is the solution $\hat{\theta}_j$ (to be obtained numerically to arbitrary precision) of
\[
S_j = \hat{\theta}_j [\Psi(\hat{\theta}_j + N) - \Psi(\hat{\theta}_j)]
\]
where $\Psi(x)$ is the psi function or digamma function, that is, the derivative of the logarithm of the gamma function (see, e.g., Etienne and Olff 2004a). The final estimate of $\hat{\theta}$ is then simply calculated as the average of the $M$ values of $\hat{\theta}_j (j = 1 \ldots M)$:
\[
\hat{\theta}_{\text{M2007}} = \frac{1}{M} \sum_{j=1}^{M} \hat{\theta}_j.
\]

Eq. 3 is not only the equation for the maximum likelihood estimator; it is also the equation for the estimator based on the method of moments, that is, it equates the observed number of species $S_j$ to the expected number of species in a (metacommunity) sample of size $N$. Thus, there is no difference between these two estimation methods in this case.

2. Munoz et al. estimation of $I_i$.—The fundamental dispersal number $I_i$ for each local sample $i$ can now be easily estimated with maximum likelihood by applying my single-sample sampling formula (Etienne 2005) to each local sample $i$ keeping $\theta$ fixed to $\hat{\theta}_{\text{M2007}}$, because given $\theta$, the local samples are independent and can be treated as single samples.

V. Improved estimation of $\theta$.—The approach outlined above works as long as $S_j < N$, because as soon as $S_j = N$, the estimate of $\theta$ becomes infinite, and any average no longer makes sense. To circumvent this problem, I propose not to estimate each $\hat{\theta}_j$, but to write down the likelihood for all $M$ iterations,

\[
P[N_1, \ldots, N_M | \theta, N] = \prod_{j=1}^{M} F_{\theta_j} \frac{\theta_j^{N_j}}{\theta^N}
\]

and then use maximum likelihood to estimate $\theta$. By setting the derivative of this likelihood with respect to $\theta$ to zero, one can show that the maximum likelihood estimator is given by the solution of

\[
\frac{1}{M} \sum_{j=1}^{M} S_j = \hat{\theta}_{\text{new}} [\Psi(\hat{\theta}_{\text{new}} + N) - \Psi(\hat{\theta}_{\text{new}})].
\]

Again, this solution can be obtained numerically to arbitrary precision. Thus, instead of averaging the estimates $\hat{\theta}_j$, one must average the number of species $S_j$. If $M$ is large enough, this average is smaller than $N$ and finite values of $\hat{\theta}_{\text{new}}$ will be obtained. Only if the local samples have no species in common at all, we find an estimate of $\hat{\theta}_{\text{new}} = \infty$, but this is indeed the best estimate in this case. Note again that Eq. 6 also gives the estimator based on the method of moments: the observed number of species averaged over $M$ simulated samples of size $N$ is equated to the number of species in a metacommunity sample of size $N$ predicted by the model. This interpretation gives way to an exact (rather than by Monte Carlo sampling) computation of $\hat{\theta}_{\text{new}}$, because the mean number of species in a sample of size $N$ where each individual is taken from a different local community $i$, can in principle be computed exactly:

\[
\langle S(k) \rangle = \sum_{[k]} N \prod_{i=1}^{N} \frac{H_i}{J_i} S[i] S[k]
\]
k = (k₁, ..., kₙ) with kᵢ the species label of the individual from site i; the sum is over all possible vectors k. Thus, \( \hat{\theta}_{\text{new}} \) is the solution of

\[
\langle S(k) \rangle = \hat{\theta}_{\text{new}} [\Psi(\hat{\theta}_{\text{new}} + N) - \Psi(\hat{\theta}_{\text{new}})].
\]  

(7b)

In practice, the number of different vectors k increases very fast with N, so this exact method is only feasible for small N. This is exactly the situation in which the Monte Carlo method requires many iterations to yield an accurate estimate of \( \hat{\theta}_{\text{new}} \).

2'. Estimation of \( I_i \):—The estimation of \( I_i \) for each sample i proceeds as before where one fixes \( \theta \) at \( \hat{\theta}_{\text{new}} \). In the Supplement, I have provided a code to perform the estimation of \( \theta \) and \( I_i \).

Results

I illustrate the new approach with simulated data and real data and compare them to the one-stage estimation approach of Etienne (2007) and to the two-stage approach of Munoz et al. (2007). I will refer to each of these three approaches as the new two-stage approach, the one-stage approach, and the two-stage approach, respectively.
approach, and the old two-stage approach, respectively. I mainly focus on the estimation of $\theta$, because this is where the new two-stage approach developed here differs from the old two-stage approach, and because the values of $m_i$ (or $I_i$) depend on the estimate for $\theta$. For the simulated data, I selected two types of data that differ in the number of local samples. In order to show that the new two-stage approach presented here works well when only a few local samples are available, in which case the old two-stage approach does not work at all. I first generated 1000 simulated data sets of three local samples each using the urn method (based on coalescence) explained in Etienne (2007). I did this for various scenarios of sample sizes and parameter values (see Table 1). For each data set, I calculated $\hat{\theta}_{\text{new}}$ with the exact method of Eq. 7a, which can be done within reasonable time because $N = 3$ is small. For comparison, I also calculated the exact estimate $\hat{\theta}_{\text{E2007}}$ using the full sampling formula of the one-stage approach.

One observes that the averaged estimates for the simulated data sets produce a biased value of $\theta$, but some bias can be expected because it also occurs for the Ewens sampling formula (Tavare and Ewens 1997). This bias decreases with sample size as scenarios 1–5 show; a sample size of 500 each already gives averaged values of $\hat{\theta}_{\text{new}}$ and $\hat{\theta}_{\text{E2007}}$ with little bias. However, the coefficient of variation in $\hat{\theta}_{\text{new}}$ is substantially larger than that in $\hat{\theta}_{\text{E2007}}$. This means that $\hat{\theta}_{\text{new}}$ is less reliable than $\hat{\theta}_{\text{E2007}}$.

Fig. 1A shows the relationship between $\hat{\theta}_{\text{new}}$ and $\hat{\theta}_{\text{E2007}}$ for the 1000 simulated data sets in scenario 1. The two parameter estimates are clearly correlated, but $\hat{\theta}_{\text{new}}$ tends to be larger (smaller) than $\hat{\theta}_{\text{E2007}}$ when both are larger (smaller) than the true value $\theta$. The estimates for $\hat{\theta}_{\text{new}}$ become worse when diversity is lower (lower $\theta$ and lower $m_i$). In contrast, $\hat{\theta}_{\text{E2007}}$ seems rather insensitive to the diversity of the community, as far as the bias concerned; however, the coefficient of variation increases when diversity decreases. The estimates for the dispersal limitation parameters, $m_i$, contain little bias in general; only when $\theta$ is low, the bias is substantial. Again, for low diversity the coefficient of variation of $m_i$ is high.

In sum, when the assumption of equal dispersal limitation is met (i.e., same $I_i$, even if $m_i$ values are different, as in scenarios 18–20 in Table 1), the one-stage approach gives more reliable parameter estimates than the approach presented here, but the latter approach nevertheless performs fairly well as long as diversity is not too low and sample sizes are sufficiently large. When this assumption is not met (the exact situation for which the approach presented here was developed), the new two-stage approach gives better estimates than the one-stage approach.

To be able to compare the estimates based on the new two-stage approach presented here with the old two-stage approach, I generated 1000 simulated data sets of 50 local samples, each local sample containing 400 individuals, using $\theta = 50$ and $m = 0.1$. I estimated both $\hat{\theta}_{\text{M2007}}$ and $\hat{\theta}_{\text{new}}$ for each of these 1000 simulated data sets, using $M = 10000$ iterations. Fig. 1B shows the correlation between the estimates for these 1000 simulated data sets. The mean value of $\hat{\theta}_{\text{new}}$ (51.55) is closer to the true value (50) than is $\hat{\theta}_{\text{M2007}}$ (54.81), which is therefore more biased. Also the variance of $\hat{\theta}_{\text{new}}$ (55.21) is lower than that of $\hat{\theta}_{\text{M2007}}$ (68.21). The two estimates are highly correlated as may be expected.

For the real data, I used two examples. The first consists of the three tropical tree communities in the Panama Canal Zone (Condit et al. 2002), which I also used in Etienne (2007). The three plots, Sherman (5.96 ha of which 5 ha is in the data file), Barro Colorado Island (50 ha), and Cocoli (4 ha) lie along a precipitation gradient (3030 mm/yr, 2616 mm/yr, and 1950 mm/yr, respectively [Condit et al. 2004]). Table 2 shows the results of the parameter estimation, for these three plots and for various local samples from the large BCI plot (see Etienne 2007). The estimates for the one-stage approach and the new two-stage approach proposed in this note are very different. This is partly because the

<table>
<thead>
<tr>
<th>Data set</th>
<th>Sample sizes and species richness</th>
<th>$\hat{\theta}_{\text{E2007}}$</th>
<th>$I_i$</th>
<th>$\hat{\theta}_{\text{new}}$</th>
<th>$I_i$</th>
<th>$\hat{\theta}_{\text{Sherman}}$</th>
<th>$I_i$</th>
<th>$\hat{\theta}_{\text{BCI}}$</th>
<th>$I_i$</th>
<th>$\hat{\theta}_{\text{Cocoli}}$</th>
<th>$I_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sherman + BCI + Cocoli</td>
<td>(2860, 21457, 1079)</td>
<td>(125, 225, 99)</td>
<td>259.3</td>
<td>44.24</td>
<td>341.7</td>
<td>33.87</td>
<td>53.71</td>
<td>30.79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sherman + BCI + Cocoli</td>
<td>(2860, 2359, 1079)</td>
<td>(125, 152, 99)</td>
<td>270.5</td>
<td>39.18</td>
<td>247.7</td>
<td>37.09</td>
<td>53.66</td>
<td>32.61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sherman + BCI1 + Cocoli</td>
<td>(2860, 2151, 1079)</td>
<td>(125, 150, 99)</td>
<td>273.9</td>
<td>39.21</td>
<td>232.6</td>
<td>37.09</td>
<td>55.43</td>
<td>33.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sherman + BCI3 + Cocoli</td>
<td>(2860, 2076, 1079)</td>
<td>(125, 162, 99)</td>
<td>280.0</td>
<td>41.18</td>
<td>322.6</td>
<td>34.35</td>
<td>57.49</td>
<td>31.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sherman + BCI4 + Cocoli</td>
<td>(2860, 2027, 1079)</td>
<td>(125, 162, 99)</td>
<td>282.2</td>
<td>42.63</td>
<td>359.8</td>
<td>33.46</td>
<td>61.38</td>
<td>30.55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sherman + BCI5 + Cocoli</td>
<td>(2860, 2000, 1079)</td>
<td>(125, 166, 99)</td>
<td>290.8</td>
<td>41.71</td>
<td>348.5</td>
<td>33.71</td>
<td>59.40</td>
<td>30.70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sherman + BCI6 + Cocoli</td>
<td>(2860, 2050, 1079)</td>
<td>(125, 153, 99)</td>
<td>297.3</td>
<td>39.13</td>
<td>399.5</td>
<td>32.71</td>
<td>49.31</td>
<td>30.12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sherman + BCI7 + Cocoli</td>
<td>(2860, 2364, 1079)</td>
<td>(125, 147, 99)</td>
<td>298.6</td>
<td>37.27</td>
<td>448.4</td>
<td>31.98</td>
<td>42.26</td>
<td>29.69</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sherman + BCI8 + Cocoli</td>
<td>(2800, 2225, 1079)</td>
<td>(125, 138, 99)</td>
<td>300.4</td>
<td>37.65</td>
<td>427.1</td>
<td>32.28</td>
<td>39.83</td>
<td>29.86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sherman + BCI8 + Cocoli</td>
<td>(2800, 2076, 1079)</td>
<td>(125, 145, 99)</td>
<td>301.5</td>
<td>40.47</td>
<td>399.5</td>
<td>32.28</td>
<td>44.01</td>
<td>29.86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sherman + BCI9 + Cocoli</td>
<td>(2860, 2129, 1079)</td>
<td>(125, 157, 99)</td>
<td>271.5</td>
<td>40.47</td>
<td>369.3</td>
<td>33.27</td>
<td>51.94</td>
<td>30.44</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The different BCI denote different subsamples from the full BCI (Barro Colorado Island) 50-ha plot.
bias and variance are higher for the new two-stage approach (see Table 1), but mostly because the former approach assumes that $I$ is the same for all samples, whereas the latter approach shows that this assumption is not justified: BCI clearly has a higher fundamental immigration number than Sherman and Coccoli. Because $I$ in fact measures recruitment limitation which involves dispersal and establishment (Jabot et al. 2008), this may be explained by the central location of BCI: because of this central location, it receives dispersers from all directions, and it has intermediate precipitation, allowing establishment of species adapted to both wetter and drier circumstances. One needs to realize, however, that it is also possible that the model assumption of a single metacommunity for the three local communities is incorrect. Although studying other models of community structure is beyond the scope of this note, this is a interesting direction for future research.

The second set of real data is the Western Ghat forest data set of Munoz et al. (2007) that consists of 50 local samples with sizes varying from 104 to 876 (average is 391.1) individuals, totaling 19 555 individuals. The number of species per local sample ranges from 19 to 71 and averages 45.24; the total number of species across all samples is 304. With $M = 100 000$ iterations, I found $\hat{\theta}_{\text{new}} = 51.5$. Furthermore, I recalculated $\hat{\theta}_{\text{M}2007}$ and found $\hat{\theta}_{\text{M}2007} = 54.7$. Compare this with the value $\hat{\theta}_{\text{M}2007} = 62.3$ reported in Munoz et al. (2007), which was based on $M = 1000$ iterations. For this data set, the one-stage approach yields $\hat{\theta}_{\text{E}2007} = 85.2$, which is much higher than the other estimates (due to the fact that it assumes equal dispersal limitation for all local communities).

**Conclusion**

In this note I have improved the approximate two-stage approach of Munoz et al. (2007) to estimate the parameters of the neutral model ($\theta$ and $m$ or $f$) from species abundances in multiple spatially separated local samples. The improvement is that it can also be applied to a small number of local samples, or to many local samples that are very different. In these cases, it is likely that a single draw from each local sample contains only singletons resulting in an infinite $\hat{\theta}_{\text{M}2007}$, but a finite $\hat{\theta}_{\text{new}}$, the estimate produced by the new two-stage approach. For an accurate estimation of $\hat{\theta}_{\text{new}}$, the number of iterations $M$ needed to obtain the expected number of species when drawing a single individual from each local sample must be very large. Fortunately, when the number of local samples $N$ is small, an exact calculation (i.e., without iterations) of $\hat{\theta}_{\text{new}}$ is possible. The estimate for $\hat{\theta}_{\text{new}}$ has little bias when the diversity is not too low ($\theta \sim 50$, $m \sim 0.1$) and as long as sample sizes are sufficiently large ($J > 1000$). The larger the sample size, the smaller the bias. However, the coefficient of variation is always quite large. I therefore suggest caution in using the two-stage approach when the sizes of the local samples are very small or when diversity is low. When the $I_i$ values are likely to be very similar across samples, it is better to use the one-stage approach of Etienne (2007).

**Acknowledgments**

I thank Franck Jabot for stimulating discussions and Emile Apol, Kyle Dexter, and one anonymous reviewer for comments on the manuscript. 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.

**Literature Cited**


**Supplement**

Program code for maximum likelihood estimation of neutral model parameters for multiple samples of species abundances using the two-stage approach described in the paper (Ecological Archives E090-058-S1).