A novel method for estimating the number of species within a region

Elad Shtilerman¹, Colin J. Thompson³, Lewi Stone²,⁵, Michael Bode⁴ and Mark Burgman⁴

¹Porter School of Environmental Studies, and ²Department of Zoology, Tel Aviv University, Tel Aviv, Israel
³Department of Mathematics and Statistics, and ⁴ARC Centre of Excellence for Environmental Decisions, School of Botany, University of Melbourne, Parkville, Victoria, Australia
⁵School of Mathematical and Geospatial Sciences, RMIT University, Melbourne, Victoria, Australia

Ecologists are often required to estimate the number of species in a region or designated area. A number of diversity indices are available for this purpose and are based on sampling the area using quadrats or other means, and estimating the total number of species from these samples. In this paper, a novel theory and method for estimating the number of species is developed. The theory involves the use of the Laplace method for approximating asymptotic integrals. The method is shown to be successful by testing random simulated datasets. In addition, several real survey datasets are tested, including forests that contain a large number (tens to hundreds) of tree species, and an aquatic system with a large number of fish species. The method is shown to give accurate results, and in almost all cases found to be superior to existing tools for estimating diversity.

1. Introduction

Estimating how many species occupy a region from sample data is an important research area in biodiversity science and ecology [1–5]. Such estimates have numerous applications, including the comparison of relative richness of different ecosystems [6], the evaluation of different sampling methodologies [7], the identification of ‘hot-spot’ areas for conservation action [8,9], the estimation of the total number of species on the planet [10,11] and monitoring changes in the number of species in time to allow for inference of climatic effects or other disturbances.

Various methods have been used to study these problems, including log–log [1] and log–linear [4] species area curves, the Chao statistic [12], bootstrap and extrapolation methods [13], jackknife estimates [14], and the abundance coverage-based estimator (ACE) [15,16]. Each method makes different statistical assumptions about the ways in which species form communities. Good reviews of the available methods may be found in earlier studies [6,13,17–19]. There is little ecological theory or evidence to support any one of these methods over the others [20] and the results of the various estimators can be substantially different. Without a well-grounded theory, species richness estimation methods can be compared only by empirically testing on benchmark ecosystem datasets that are mapped out in detail. There has been recent interest [6,15,17,18,21] in resolving the differences in the various estimation methods and it is now understood to be a research direction that needs much more attention.

Here, we present a new approach based on asymptotic analysis of sampling data. To set the stage, we suppose that the true number of species in the entire area under investigation is \( S^* \). Suppose also that we have accurate surveys of \( n \) quadrats that together only cover a small part of the total area. Given that the number of species is known in each quadrat, is it possible to estimate the total number of species, \( S^* \)?

In answering this question, our starting point, as in many other studies (e.g. [13]), is the well-known expression

\[
S(n) = \sum_{j=1}^{n} \left( 1 - (1 - p_j)^n \right).
\]  (1.1)
This expression relates the expected number of species \( S(n) \) found in samples of \( n \) quadrats to the total true number of species \( S^* \) in the region. In this formula, \( p_j \) is the probability that species \( j \) is found in a single quadrat. The term \((1 - p_j)^n\) is thus the probability that species \( j \) is not found in any of the \( n \) quadrats, and therefore \( 1 - (1 - p_j)^n \) is the probability it is found in at least one out of \( n \) quadrats. Hence the sum is equal to the mean number of species found in \( n \) quadrats.

Previous methods typically assume specific functional forms for \( S(n) \), such as exponential or hyperbolic [13], and extrapolate from the data to estimate \( S^* \), which from equation (1.1) is equal to the large \( n \) limit of \( S(n) \). Our approach, on the other hand, assumes a general smooth but unknown scaling functional form for the \( p_j \) probabilities, from which we deduce an asymptotic form for \( S(n) \) to estimate \( S^* \) in the large \( n \) limit. A similar approach has been presented recently by Chao et al. [22] to determine asymptotic relations between low species abundance numbers (see also [17]). Although the asymptotic methods described are similar to ours, the underlying assumptions are quite different. Our asymptotic forms are derived in appendix A where comparisons are made with the work of Chao et al. [22].

This article is organized as follows. In §2, we outline the derivation of the estimation method and describe the main assumptions on which it is formed. In §3, we first show that the method gives good results for randomly generated datasets in various scenarios. Following this, we test our method in comparison with other estimation techniques on various real datasets. These results typically show that in most cases our method outperforms others. In §3b, we present an analysis of the asymptotic \( S(n) \) curves for a large number of samples. This analysis shows that the expected behaviour in the large \( n \) regime is as our theory predicts, therefore validating our underlying assumptions. Finally, in §4, we summarize our main findings.

2. Material and methods

The values of the \( p_j \) in equation (1.1) are unknown. We can, however, assume without loss of generality that the species are ordered or labelled so that

\[
0 < p_1 \leq p_2 \leq p_3 \leq \cdots \leq p_n \leq 1. \tag{2.1}
\]

We now assume that \( p_j \) are independent of \( n \) and can be expressed in terms of a (monotonic) scaling function \( f(x) \) for \( 0 \leq x \leq 1 \) as

\[
p_j = f\left(\frac{j}{S^*}\right), \quad j = 1, 2, 3, \ldots, S^*. \tag{2.2}
\]

where \( x = j/S^* \). Such a function will always exist though its precise functional form, by definition, is unknown. We assume only that \( f(x) \) is continuous and convex, and that \( f(0) = 0 \). These assumptions are justified from observing ordered \( p_j \) of real datasets, which reveals that there are typically many rare species and that the \( p_j \) curve rises very gradually close to the origin that leads to a convex form.

To illustrate this, figure 1 shows \( p_j \) for quadrats of \( 30 \times 30 \) m (0.09 ha) within the Barro Colorado Island (BCI) forest dataset [23–25]. Figure 1a shows \( p_j \) for all species—it is evident that it has the convex form that we assume. Figure 1b is a zoom in on small \( j/S^* \) (the rare species). The solid line is a quadratic fit \((y = 0.42x^2 - 0.008x + 0.0013)\). Note that the coefficient of the quadratic term is several orders of magnitude larger than the others, indicating that the dominant term is the quadratic and not the linear or constant terms. The general shape of the curve and the quadratic fit confirm our assumptions about \( f(x) \). We discuss this further in §3b.

To continue the outline of the derivation, we express equation (1.1) as

\[
S(n) = S^* (1 - I(n)), \tag{2.3}
\]

with \( I(n) = 1/S^* \sum_{j=1}^{S^*} (1 - p_j)^n \). For large \( S^* \), given our assumptions, we can approximate \( I(n) \), using equation (2.2), as an integral

\[
I(n) = \frac{1}{S^*} \sum_{j=1}^{S^*} \left(1 - f\left(\frac{j}{S^*}\right)^n\right) \sim \int_0^1 \left(1 - f(x)^n\right) dx. \tag{2.4}
\]

For large \( n \), the integrand in equation (2.4) is sharply peaked about its maximum, which by our assumptions occurs at \( x = 0 \). The asymptotic form of \( I(n) \) in equation (2.4) for large \( n \) is then obtained by expanding the integrand around \( x = 0 \). Because of this sharp maximum, the terms for \( f(x) \) close to \( x = 0 \) contribute mostly to the final value of the integral. In mathematical terms, this is known as Laplace’s method [26]. In short, the Laplace approximation method allows one to isolate the terms which contribute the most to the sum in equation (1.1), terms with \( p_j \ll 1 \), and thus obtain an accurate estimate for the total number of species. Therefore, instead of calculating a complicated integral, we just sum the main contributions using the method of Laplace. The rigorous derivation is briefly outlined for the present application in appendix A.

In ecological terms, the dominance of small \( x \)-values in the asymptotic evaluation of equation (2.4) means, from equation (2.2), that rare species are the prime determinants of total species richness \( S^* \), irrespective of the precise form of the relative
densities of species. The same statement in fact applies to the Chao statistic [12] and jackknife estimates [14,27] that are determined by species with low abundance numbers (specifically singletons and doubletons). Intuitively, individuals of the rarer species are far less likely to be observed in a random collection of quadrats that samples only a small part of the total region. Thus, in the derivation rarer species are given a higher weight when calculating the integral.

Although the functional form of \( f(x) \) is not known, based on our assumptions it is reasonable to assume that for small \( x \)
\[
f(x) \sim ax^n \text{ where } a > 0 \text{ and } x \ll 1.
\] (2.5)

If \( a = 1 \), this would correspond to a linear form so that the probability that species \( j \) is found in a quadrat increases linearly with the species index \( j \). If \( a = 2 \), this would correspond to a quadratic increase. For real datasets, we have found \( a = 2 \) is often the more realistic choice (see §3b). If one makes the additional assumption that \( f(x) \) is smooth around \( x = 0 \), and recalling that \( f(0) = 0 \), one can make the power series approximation
\[
f(x) = x^{\alpha}(a_0 + a_1 x + a_2 x^2 + \cdots).
\] (2.6)

Inserting this expression in the integral of equation (2.4), Laplace’s method allows us to find the asymptotic form of \( I(n) \) and thus \( S(n) \) in equation (2.3) (see appendix A for details). Because of the sharp maximum in the integrand, around \( x = 0 \), the terms for \( f(x) \) that contribute mostly to the final value of the integral are readily identified. As shown in appendix A, based on the simplest approximation for \( f(x) \), \( S(n) \) in equation (2.3) has the asymptotic form
\[
S(n) \sim S^* \left( 1 - \frac{A}{n^{\alpha/\alpha}} \right)
\] (2.7)
for large \( n \) (and \( S^* \)), where \( A \) is a constant. For \( a = 1 \), equation (2.7) is simply an asymptotic version of the well-known and popular two-parameter hyperbolic (or Michaelis–Menten) model for \( S(n) \) [13]. In general, equation (2.7) implies that a plot of \( S(n) \) versus \( n^{-1/\alpha} \) should approach a straight line with intercept \( S^* \) and slope \(-AS^*/n\) for large \( n \). In §3b, we explicitly show this asymptotic behaviour for several datasets.

We refer to \( S(n) \) as calculated by equation (2.7) as our first-order predictor for the total number of species \( S^* \) (from here on ‘Laplace 1’). In practice, even better estimates of \( S^* \) may be achieved by refining the model to include higher-order terms obtained directly from Laplace’s method. The second-order approximation (from here on ‘Laplace 2’) which is derived in appendix A is given by
\[
S(n) \sim S^* \left( 1 - \frac{A}{n^{\alpha/\alpha}} + \frac{B}{n^{2/\alpha}} \right).
\] (2.8)

Assuming that the function \( f(x) \) is convex and analytic, the parameter \( \alpha \) appearing in the equations of Laplace 1 and 2 is chosen to take the value \( \alpha = 2 \), as discussed in §3b. As we have already noted, and as shown in figure 1, this choice works extremely well for the BCI data and other datasets we test. Thus, working with the species accumulation curve \( S(n) \) obtained from quadrat data, one can numerically fit model equation (2.8) to estimate the three parameters \( A, B \) and (most importantly) total species numbers \( S^* \).

As discussed shortly, Laplace 2 is our model of choice because of its accurate estimates and relative ease of implementation. In the following section, we present some case studies and compare our model predictions with those of others mentioned previously.

3. Results

(a) Analysis of real and simulated datasets

We first show that the method yields accurate and consistent estimates based on computer-generated random datasets. This requires the construction of a random set of \( n_q \) quadrats which can be considered typical samples of the entire area. The species accumulation curve \( S(n) \) may then be obtained from the sampled quadrats. Although there are numerous ways to achieve this, the method suggested below appears to have the most advantages.

(i) First, it is necessary to choose the \( p_j \), the probabilities that species \( j \) is found on a given quadrat. To accomplish this, we select an arbitrary polynomial \( f(x) \) (polynomial of \( x = j/S^* \) satisfying our basic assumptions, so that \( f(x) \) is of the form:
\[
f(x) = x^n(a_0 + a_1 x + a_2 x^2 + \cdots).
\] (3.1)

For test purposes, we chose rather arbitrarily the convex monotonically increasing function \( f(x) = 0.23 (x^2 + x^3 + x^4) \), but note that the results we present are quite robust to changes in the polynomial coefficients.

(ii) We then ‘build’ the \( n_q \) quadrats by randomly selecting the species present on each quadrat. Essentially, this requires scanning over all \( S^* \) species and allowing species \( j \) to reside on a given quadrat with probability \( p_j \).

(iii) Once the \( n_q \) quadrats are generated, it is possible to construct the \( S(n) \) curve. As mentioned, \( S(n) \) is the number of species observed in \( n \) of the total \( n_q \) quadrats. In our case, the order of the accumulation of species among the \( n \) quadrats is insignificant, and therefore, in order to get a smooth \( S(n) \) curve, we averaged over 100 permutations out of the \( n_q \)!

This randomization process is described and elaborated by Colwell & Coddington [13, §4e].

(iv) A standard MATLAB routine (nlinfit) is then used to fit model equation (2.8) to the \( S(n) \) curve to obtain the estimate for \( S^* \), the number of species. Carrying out the fit actually means we are extrapolating the \( S(n) \) curve for \( n \to \infty \).

One should note that in the case of random data, which are generated according to the above procedure, the quadrat size is irrelevant as \( p_j \) determines directly the probability that a species will be in a quadrat. However, when analysing real data, given that the number of quadrats is fixed, then the quadrat size will obviously have a large impact on the estimation capability.

Figure 2a–c demonstrates how the method works. The diamond markers are the values for \( S(n) \) as a function of the number of quadrats \( n \). The curves are fitted with the first- and second-order models to obtain the estimated number of species. Figure 2a,b is for random datasets generated with \( f(x) = 0.23 (x^2 + x^3 + x^4) \) with 320 species, while figure 2c is for the BCI dataset (also 320 species with a quadrat size of 25 × 25 m).

Looking more closely at figure 2a,b, based on \( n = 8 \) quadrats, one observes the curve begins with \( S(1) = 53 \) and ends with \( S(8) = 170 \). The goal of the predictor is to extrapolate the graph for large \( n \). Clearly, by eye alone it is impossible to make an accurate guess for \( S^* \) where the curve saturates.
Table 1. The table shows the mean and standard deviation of the number of species estimated for 500 repetitions.

<table>
<thead>
<tr>
<th>n_q</th>
<th>order</th>
<th>random, ( f_1(x) S^*=200 )</th>
<th>random, ( f_1(x) S^*=320 )</th>
<th>random, ( f_2(x) S^*=320 )</th>
<th>BCI, ( S^*=320 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>first</td>
<td>144 ± 6</td>
<td>229 ± 8</td>
<td>197 ± 10</td>
<td>235 ± 10</td>
</tr>
<tr>
<td>20</td>
<td>first</td>
<td>160 ± 5</td>
<td>256 ± 6</td>
<td>247 ± 7</td>
<td>259 ± 7</td>
</tr>
<tr>
<td>8</td>
<td>second</td>
<td>184 ± 11</td>
<td>293 ± 15</td>
<td>303 ± 18</td>
<td>284 ± 16</td>
</tr>
<tr>
<td>20</td>
<td>second</td>
<td>189 ± 11</td>
<td>302 ± 9</td>
<td>328 ± 11</td>
<td>287 ± 10</td>
</tr>
</tbody>
</table>

Yet the second-order predictor remarkably estimates \( S^* \) within 5–10% of the true value (\( S^*=320 \)).

Table 2 summarizes results for several scenarios including results for random datasets generated with two functions \( f_1(x) = 0.23(x^3 + x^2 + x^2) \) and \( f_2(x) = 0.3x^2 \). The general trends are that the method is more accurate (i) when using second order instead of first order, (ii) when the number of quadrats is increased and (iii) when the number of species is larger. Moreover, the method is robust to different choices of the function \( f(x) \). Overall, table 1 shows that our method gives accurate estimates with a relatively small number of quadrats.

In order to examine our method in detail, we used five forest datasets (BCI, Sherman, Cocoli, Luquillo and Oosting [23–25]), a fish survey dataset consisting of data of 243 hauls from the Mediterranean Sea [26,27], and a random dataset created as previously described. The parameters we used and general information regarding the datasets is summarized in table 2.

For each dataset, we tested eight methods, of which three were parametric: Laplace 2, Negative exponential and Michaelis–Menten (the last two involve fitting the \( S(n) \) curve to functions \( A(1 - \exp(-Bn)) \) and \( An/B + n \), respectively). The other five were non-parametric: Chao 2, jackknife first and second order, bootstrap [13], and ACE [15,16], which are all based on different statistics of the quadrat data.

Figure 3 shows the results of the comparison of the methods for all datasets except Oosting, which is found in the electronic supplementary material.

We note from the results of the comparison that when the number of species is large (figure 3a–c,f) the Laplace 2 predictor outperforms the other methods tested. As the theory is based on an asymptotic expansion for a large number of species and for a large number of quadrats, this is expected. But even for a smaller number of species Laplace 2 appears at least as good as any other method, and with smaller confidence intervals. What also stands out is that the same pattern of relative success of the different predictors repeats itself for different datasets (e.g. negative exponential always underestimates by a large percentage, Chao 2 has relatively large confidence intervals, etc.).

In addition to comparing the various methods on different datasets, we performed an analysis comparing the different methods for different sampling areas (number of quadrats). The results of this analysis are in figure 4. It is
apparent that increasing the area covered by the quadrats leads to better and more accurate estimates of the number of species. Again, the pattern of relative success of the different methods is preserved for different fractions sampled of the total area.

(b) Asymptotic analysis of empirical data
In order to find the best-fitting \( \alpha \) parameter, we analysed empirical datasets for their asymptotic behaviour. As we are interested in the limit of large \( n \) and large \( S^* \), we chose the datasets of BCI, Sherman and Cocoli, as they have a large number of species and contain a relatively large area, enabling us to use a large number of quadrats. For each dataset, we generated many repetitions of the \( S(n) \) curve for large \( n \). In figure 5, we show the resulting \( S(n) \) curves plotted versus \( n^{-1/2} \) and below them versus \( n^{-1} \). It is evident that the linear fits in figure 5a–c are much more accurate than the linear fits in figure 5d–f. For this reason, throughout the analysis we chose \( \alpha = 2 \). It is worthwhile to note also that the confidence intervals become smaller for a larger number of quadrats \( n \) (smaller \( n^{-1/2} \)), as is expected by theory based on a large \( n \) approximation.

4. Discussion
In this work, we presented a novel method to tackle the problem of species number estimation based on quadrat data.
We have tested the method in several scenarios, showing that it gives accurate results and outperforms other existing methods in most cases, and performs as well as the jackknife second-order method when the number of species is small. In addition, we have empirically demonstrated that our model’s underlying assumptions are fulfilled, therefore reinforcing its reliability. We found that our model is insensitive to details of quadrant choices and different \( p_j \) probabilities, and therefore is a good candidate to be used in new areas that are being sampled.

Acknowledgements. We thank Dori Edelist for providing the comprehensive dataset of fish hauls in the Mediterranean.

Funding statement. M. Bode was funded by the ARC (DE130100572). The BCI forest dynamics research project was made possible by National Science Foundation grants to Stephen P. Hubbell (DEB-064086, DEB-0425651, DEB-0346488, DEB-0129874, DEB-00753102, DEB-9909347, DEB-9615226, DEB-965226, DEB-9405932, DEB-9049393, DEB-921033, DEB-9100008, DEB-890869, DEB-8805042, DEB-8206992 and DEB-7922197) support from the Center for Tropical Forest Science, the Smithsonian Tropical Research Institute, the John D. and Catherine T. MacArthur Foundation, the Mellon Foundation, the Small World Institute Fund, and numerous private individuals, and through the hard work of over 100 people from 10 countries over the past two decades. The plot project is part the Center for Tropical Forest Science, a global network of large-scale demographic tree plots.

Appendix A

Laplace’s method [26] is concerned with the asymptotic evaluation of integrals of the form

\[
L(n) = \int_0^b e^{nh(x)} g(x) \, dx,
\]

in the limit of large \( n \). In this limit, the exponential term in equation (A 1) becomes sharply peaked about its maximum. The method then involves expanding \( h(x) \) around its maximum and retaining only the leading order terms in this expansion. In our particular case, equation (2.4), \( a = 0, b = 1, g(x) = 1 \) and

\[
h(x) = \log(1 - f(x)),
\]

with the maximum of \( h(x) \) occurring at the endpoint \( x = 0 \). For the assumed functional form, equation (2.5) for \( f(x) \) near \( x = 0 \), i.e.

\[
f(x) = x^\alpha (a_0 + a_1 x + a_2 x^2 + \cdots) \quad a \geq 1, \quad a_0 > 0.
\]

Expansion of the logarithm in equation (A 2) gives to leading order (when \( a \geq 1 \))

\[
h(x) = -f(x) - \frac{1}{2} f'(x) - \cdots
\]

\[
-\alpha_1 x^\alpha - a_1 x^{\alpha+1} - \cdots - \frac{1}{2} \alpha^2 x^{2\alpha}
\]

\[
-\alpha_1 x^\alpha - a_1 x^{\alpha+1},
\]

where

\[
a = \begin{cases} a_1, & \text{when } a > 1 \\ a_1 + \frac{1}{2} \alpha^2, & \text{when } a = 1. \end{cases}
\]

To leading order for \( a \geq 1 \) and large \( n \) we then have

\[
I(n) = \int_0^1 e^{nh(x)} \, dx
\]

\[
= \int_0^1 \exp(-\alpha_0 x^\alpha - a_1 x^{\alpha+1} - \cdots) \, dx.
\]

Changing variables to \( x = n^{-1/\alpha} y \), we then have

\[
I(n) \sim n^{-1/\alpha} \int_0^{n^{1/\alpha}} \exp\left[-\alpha_0 y^\alpha\right] \exp\left[-a_1 \int_0^{y^{1+\alpha}} \right] dy
\]

\[
\sim n^{-1/\alpha} \int_0^{n^{1/\alpha}} e^{-\alpha_0 y^\alpha} \left[1 - \frac{a}{n^{1/\alpha}} y^{1+\alpha}\right] dy
\]

\[
= An^{-1/\alpha} - Bn^{-2/\alpha} \quad \text{for } n \gg 1,
\]
where

\[ A = \int_0^\infty e^{-a x^2} \, dx \quad \text{and} \quad B = a \int_0^\infty y^{\frac{3}{2} - a} e^{-a y} \, dy \]  \quad (A8)

are constants expressible in terms of gamma functions (depending on \( a \)) [26]. Substituting the leading order terms in equation (A7) into equation (2.3) gives the required results of equations (2.7) and (2.8).

For comparison, the work of Chao et al. [22] is primarily concerned with asymptotic estimates for frequency counts \( f_k \) of species represented by \( k \) individuals from a sample of \( n \). In our notation, from equation (1.1), the expectation value of \( f_0 \), for example, is given by

\[ E(f_0) = \sum_{j=1}^{S*} E([1 - p_j]^n). \]  \quad (A9)

Chao et al. [22] assume that \( p_j \geq a \) for \( j = 1, 2, \ldots, S^* \) and that the \( p_j \) have ‘a common marginal density \( g(x) \)’ so that one has the following equation (before eqn 6 in [22]):

\[ E(f_0) = S^* \int_0^1 (1 - x)^n g(x) \, dx, \]  \quad (A10)

which can be expressed as a Laplace integral (A1) with

\[ h(x) = \log(1 - x). \]  \quad (A11)

While our integral \( l(n) \) in equation (2.4) (see equations (A2) and (A6)) can be expressed in the form of equation (A10) by changing variables to \( y = f(x) \), the resulting \( g(y) \) is typically singular at the maximizing point of \( h(x) \) (e.g. for \( f(x) = x^2 \), \( g(y) = \frac{y^2}{2} \)) and the Laplace method becomes invalid. Comparison of equations (A2), (A3) and (A11) shows in fact that equation (A10) is essentially a special case of \( \alpha = 1 \) in equation (A7). In other words, the methods presented by Chao et al. [22] are confined to asymptotic expansions in powers of \( n^{-1} \), which we have shown in §3B to provide poorer fits to real data in comparison with equation (2.7).

References

24. Condit R. 1998 Tropical forest census plots: methods and results from Barro Colorado Island, Panama and a comparison with other plots. Berlin, Germany: Springer.