Sample-Based Estimation of Tree Species Richness in a Wet Tropical Forest Compartment

Steen Magnussen[1] **and Raphaël Pélissier**[2]

Abstract.—Petersen's capture-recapture ratio estimator and the well-known bootstrap estimator are compared across a range of simulated low-intensity simple random sampling with fixed-area plots of 100 m^2 in a rich wet tropical forest compartment with 93 tree species in the Western Ghats of India. Petersen's ratio estimator was uniformly superior to the bootstrap estimator in terms of average error (bias) and mean absolute error. The observed richness always had the largest negative bias. A large negative bias of 25 percent persisted even when approximately 10 percent of the area was sampled. Estimated confidence intervals had poor coverage rates. A proposed variance estimator for the observed richness performed well.

Introduction

Obtaining an unbiased and precise estimate of the number of forest tree species (*S*) currently growing in a region, State, or country poses a challenge. The number of species observed in a statistically valid sample is downwardly biased, and historic data and tree distribution maps may not reflect current realities (Guralnick and Van Cleve 2005).

A forest survey would ideally provide an unbiased and precise estimate of *S* for the populations of interest. Research into the species estimation problem was pioneered by Arrhenius (1921), Fisher *et al.* (1943), and Good and Toulmin (1956). We now have a plethora of estimators and estimation procedures (Bunge and Fitzpatrick 1993, Walther and Moore 2005). Rare species, easily missed in typically low-intensity forest survey sampling, exert a disproportionate influence on the results (Link 2003, Mao and Colwell 2005). Samples with a poor representation of rare species cannot be expected to yield reliable estimates of *S*.

Can we expect a typical low-intensity forest survey to provide an acceptable estimate of *S*? Experience with sample-based estimation of *S* for tree species is limited. Schreuder *et al.* (1999) assessed 10 modifications of Chao's and Lee's nonparametric estimators by resampling two large data sets with 4,060 forest inventory plots from Missouri and 12,260, from Minnesota, respectively. Sample sizes in the order of 500 to 700 were deemed necessary to keep bias below 15 percent. Sample sizes of 80 produced a negative bias of about 40 percent. Palmer (1990) performed resampling with very small circular plots of 2 m^2 in the Duke Forest (North Carolina, United States) and found that the nonparametric second-order jackknifed and that the bootstrap estimators performed best in terms of accuracy and precision. Hellmann and Fowler (1999), in a similar resampling study with 25 m^2 plots, found the second-order jackknifed estimator to be the best for low-intensity sampling (< 10 percent of area sampled). Gimaret-Carpentier *et al.* (1998a) found Chao's estimator(s) to be superior to the generalized jackknifed estimator for estimating richness in a wet, species-rich tropical forest.

The objective of this study is to introduce and assess the performance of Petersen's ratio estimator of richness (Thompson 1992) in low-intensity simple random sampling with fixed-area plots in a wet, species-rich tropical forest compartment. Petersen's ratio estimator, which rests on a minimal set of assumptions, is easy to calculate and lends itself to a bootstrap estimation of sampling errors, but has so far not been used for the purpose of tree species-richness estimation. The bootstrap estimator serves as a reference benchmark as it is a widely known and equally simple estimator (Bunge *et al.* 1995, Schreuder *et al.* 1999).

[1] Canadian Forest Service, Natural Resources Canada, 506 West Burnside Road, Victoria, BC V8Z 1M5, Canada. E-mail: steen.magnussen@nrcan.gc.ca.
[2] UMR Botanique et Bioinformatique de l'Architecture des Plantes (AMAP), TA40/PS2, 34398 Montpellier Cedex 05, France.

Material and Methods

Data from a 28-ha forest compartment in the Kadamakal Reserve Forest (Kadagu District, Karnatiaka State, India) near the village of Uppangala in the Western Ghats mountain range (lat. 12º30'N by long. 75º39'W; 500–600 m ALT) are used for this study. The forest type is Dipterocarpus indicus–Kingiodendron pinnatum–Humboldtia brunonis (Pascal 1982). Five strips with a width of 20 m, oriented north-south, 100 m apart, and 180- to 370-m long, were stem mapped (Pascal and Pélissier 1996). Species and spatial location were determined for all trees with a diameter at breast height (d.b.h.) ≥ 30 cm. Pascal and Pélissier (1996) found 1981 such trees (635 trees per ha), representing 93 species ($S = 93$).

The five 20-m-wide survey lines, totaling 1,560 m in length, were subdivided into 312 100 m^2 rectangular (5 m by 20 m) plots. Simple random sampling (SRS) with sample sizes $n = 10, 15, ..., 30$ plots without replacement was simulated. Accordingly, between 3.2 and 9.6 percent of the area was sampled. The area sampled is denoted by A_s. Sampling, followed by estimation of species richness (S), was repeated 2,000 times for each sample size.

Let S_{OBS} be the number of species encountered in n sample plots. Encountered species are labeled by an index i $(i = 1,...,S_{OBS})$. The sample data consist of a size $S_{OBS} \times n$ binary matrix δ with element $\delta_{ij} = 1$ if the ith species occurred in the jth plot and zero (0) otherwise. A design-unbiased estimator of the sampling variance of S_{OBS} is not available. The distribution of S_{OBS} has been assumed Poisson with a mean and a variance equal to S_{OBS}. Instead $S_{OBS}^2 \times \left(\sum_{i=1}^{S_{OBS}} \delta_{i\cdot}\right)^{-1}$ is proposed as an estimator of the sampling variance on the grounds that $\left(\sum_{i=1}^{S_{OBS}} \delta_{i\cdot}\right) \times S_{OBS}^{-1}$ is the average number of plots per unique species in the sample.

To arrive at Petersen's capture–recapture ratio estimator of richness, we first consider the n sample plots as composed of two independent half-samples. Let $S_{OBS}^{(1)}$ be the number of species found in the first half and $S_{OBS}^{(2)}$ the number of species in the second half. We have $S_{OBS} = S_{OBS}^{(1)} + S_{OBS}^{(2)}$. Some species are seen in both half-samples; let this number be denoted by $S_{OBS}^{(1)\cap(2)}$. Petersen's capture–recapture estimator (Thompson 1992) of S is then

$$\hat{S}_{PET} = \eta \times \frac{S_{OBS}^{(2)}}{S_{OBS}^{(1)\cap(2)}} \times S_{OBS}^{(1)} \quad (1)$$

where η is a multiplier that scales the estimate from the half sample to the complete sample of size n. Here $\eta = \left(S_{OBS}^{(1)} + S_{OBS}^{(2)}\right) / S_{OBS}^{(1)}$. In case $S_{OBS}^{(1)\cap(2)} = 0$, a modification suggested by Chapman (Seber 1982) would be used. To avoid estimating SPET from a single arbitrary data split, we computed $\hat{\bar{S}}_{PET}$ as the average of $\hat{S}_{PET}^{i}, i = 1,2,\ldots,1\,000$ where $\hat{S}_{PET}^{i}$ is an estimate based on the ith random split of the n sample records. The variance of $\hat{S}_{PET}$ was estimated as $Var\left(\hat{S}_{PET}^{i}\right)$.

Smith and van Belle (1984) first suggested a bootstrap estimation of S. A bootstrap sample of size n is drawn with replacement from the n observed sample records. Let S_{BOOT}^{r} be the number of unique species in the rth such bootstrap sample. The difference, $\Delta S_{BOOT} = S_{BOOT}^{r} - S_{OBS}$, is a bootstrap estimate of bias (Efron and Tibshirani 1993); thus

$$E_r\left(S_{BOOT}^{r} - S_{OBS}\right) = \Delta S_{BOOT} = \sum_{i=1}^{S_{OBS}} (1 - p_i)^n \quad (2)$$

with expectation taken across all possible size n bootstrap samples. ΔS_{BOOT} is an estimate of the number of species "missed" in the sample (bias). From equation (1) we obtain the bootstrap estimator of S:

$$\hat{S}_{BOOT} = S_{OBS} + \widehat{\Delta S}_{BOOT} \quad (3)$$

A variance estimator for $\hat{S}_{BOOT}$ conditional on S_{OBS} is

$$\widehat{\text{var}}\left(\hat{S}_{BOOT}\right) = \sum_{i=1}^{S_{OBS}} \left(1 - q_i^{\,n}\right) q_i^{\,n} + \sum_{i=1}^{S_{OBS}} \sum_{j \neq i}^{S_{OBS}} q_{ij}^{n} - q_i^{n} q_j^{n} \quad (4)$$

where q_i is the proportion of sample plots that do not contain the ith species and q_{ij} is the proportion that contains neither the ith nor the jth species.

The two richness estimators either explicitly or implicitly assume an infinite population size. To take the finite population size into consideration (Valliant *et al.* 2000), we corrected the estimates in equations (1) and (3) by

$$\hat{S}'_{M} = S_{OBS} + (1 - f_{pc})(\hat{S}_{M} - S_{OBS}) \quad (5)$$

where $f_{pc} = A_s \times A^{-1}$ with $M = \{PET, BOOT\}$. This correction ensures that $f_{pc} \to 1$ means $\hat{S}'_M \to S_{OBS}$ as required. A corresponding correction was applied to estimators of sampling variance.

The performance of S_{PET} and S_{BOOT} will be assessed by their average error (estimate of bias), precision (actual and average of estimated sampling errors), accuracy as estimated by the mean absolute difference (*Mad)* between an estimate and the true value, the proportion of estimates within 10 percent of the true value ($\delta 10$), and finally the coverage rate of estimated 95 percent confidence intervals $(pCI95)$.

Results

Observed richness had, as expected, the largest negative average error (estimate of bias), as detailed in table 1. Even with 10 percent of the area sampled, the bias was –56 percent. The average relative error of the observed richness declined at a decreasing rate as sample size increased. S_{PET} was clearly a better estimator than S_{BOOT} in terms of its average relative errors (bias), which were roughly half of those associated with S_{BOOT}. The rate of decline in the average relative error was similar for the three estimators.

Relative standard errors of the richness estimates were about 4 to 5 percent for n = 10 and 3 to 4 percent for n = 30 (table 1). Hence, the decline in the standard error for an increase in n was much slower than $-2^{-1}n^{-1.5}$, as expected for conventional forest inventory estimates of population totals, namely averages. Average estimates of precision for *PET* were quite conservative: about three times larger than the empirically estimated errors (table 1). In contrast, those for *BOOT* were somewhat liberal (too small) at n = 10, but at larger sample sizes ($n \geq 20$) they matched the empirical estimates to within 0.5 percent. The proposed variance estimator for OBS appears attractive inasmuch as the observed and the average of the estimated errors were within 0.5 percent of each other.

Mean absolute differences (table 2) were dominated by the bias component; as such, the results largely mirror those detailed above for the average error. The fraction of estimates within 10 percent of the actual value of 93 was low for *PET* (≤ 8 percent) for all sample sizes. It was 0 for both *OBS* and *BOOT*. Estimated 95 percent confidence intervals of *BOOT* and *OBS* estimates of richness failed to include the actual value (table 2). Results were not much better for *PET*, with coverage rates increasing from just 15 percent at n = 10 to 34 percent at n = 30.

Table 1.—*Mean error (estimate of bias) of richness estimates. Actual (s.e.) and average of estimated sampling errors ($\overline{s.e.}$) are in parentheses (s.e./ $\overline{s.e.}$). Errors are in percent of true richness S = 93. Means are across 2,000 replicate samples.*

Estimator	Sample size (A_s/A×100)				
	10 (3.2)	15 (4.8)	20 (6.4)	25 (8.0)	30 (9.6)
S_{obs}	– 75 (4/4)	– 69 (4/4)	– 64 (4/4)	– 60 (4/4)	– 56 (4/4)
$\hat{S}_{PET}$	– 46 (4/11)	– 38 (4/11)	– 34 (4/10)	– 29 (3/0)	– 25 (3/9)
$\hat{S}_{BOOT}$	– 69 (5/3)	– 62 (4/3)	– 57 (4/3)	– 57 (3/3)	– 48 (3/3)

Table 2.—*Mean absolute error (*Mad*) of richness estimates.* Mad *is in percent of true richness (93). Percent of estimates within 10 percent of true value (δ_{10})and coverage rates of estimated 95 percent confidence intervals (pCI_{95})are in parentheses (δ_{10} / pCI_{95}).*

Estimator	Sample size (A_s/A×100)				
	10 (3.2)	15 (4.8)	20 (6.4)	25 (8.0)	30 (9.6)
S_{obs}	75 (0/0)	69 (0/0)	64 (0/0)	60 (0/0)	56 (0/0)
$\hat{S}_{PET}$	46 (2/15)	38 (3/21)	34 (3/21)	29 (5/28)	25 (8/34)
$\hat{S}_{BOOT}$	69 (0/0)	62 (0/0)	57 (0/0)	57 (0/0)	48 (0/0)

Discussion

Low-intensity forest inventories do not provide estimates of tree species richness on a routine basis. Given the importance that is attached to notions of species richness and biodiversity, however, it would seem reasonable to expect that forest inventories would provide such an estimate. While it is generally recognized that the observed number of species will be downwardly biased, it is probably less appreciated that almost any estimator of species richness will be an improvement over the observed richness. It is generally accepted that there is no universally best estimator of *S*. The choice must be based on documented performance (Chao and Bunge 2002). Because an overestimation of richness can have a negative impact on credibility, an estimator unlikely to produce an inflated estimate is warranted. At low-intensity sampling both Petersen's and the bootstrap estimator are unlikely to produce an inflated estimate. Palmer (1990, 1991) and Hellmann and Fowler (1999) have already confirmed this property of S_{BOOT} . The uniform superiority of Petersen's estimator vis-á-vis the bootstrap estimator holds promise, but it needs to be corroborated by additional studies before one can draw any general conclusion.

Because the study site had many rare and just a few common species we cannot *a priori* expect to obtain very good estimates of richness from low-intensity forest inventory sampling. Condit *et al.* (1996) suggest that a sample of at least 1,000 individually sampled trees, or about 10 percent of a population, is needed in wet, tropical species-rich forests before a sample-based estimate of species richness is within 15 percent of the actual value.

Our study reiterated the importance of choosing a suitable estimator of richness. It is well known that the performance of an estimator depends not only on the statistical sampling designs but also on the population structure and spatial distribution of species (Brose *et al.* 2003, Colwell *et al.* 2004, Keating *et al.* 1998). Only an extensive assessment of a larger suite of estimators in diverse environments and across a series of conventional low-intensity forest inventory designs will allow a resolution to the question of whether we can hope to obtain estimates of tree species richness that are both reasonably accurate and reasonably precise from low-intensity forest inventories. The test designs would have to include sampling with plots of different size, as the effect of plot size is expected to depend strongly on both the estimator and the spatial distribution of species in the population of interest (Condit *et al.* 1996, Gimaret-Carpentier *et al.* 1998b).

Acknowledgments

The French Institute of Pondichéry kindly provided the data, and the Karnataka Forest Department supported the data collection.

Literature Cited

Arrhenius, O. 1921. Species and man. Journal of Ecology. 9: 95-99.

Brose, U.; Martinez, N.D.; Williams, R.J. 2003. Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. Ecology. 84(9): 2364-2377.

Bunge, J.; Fitzpatrick, M. 1993. Estimating the number of species: a review. Journal American Statistical Association. 88(421): 364-373.

Bunge, J.; Fitzpatrick, M.; Handley, J. 1995. Comparison of three estimators of the number of species. Journal Applied Statistics. 22(1): 45-59.

Chao, A.; Bunge, J. 2002. Estimating the number of species in a stochastic abundance model. Biometrics. 58: 531-539.

Colwell, R.K.; Mao, C.X.; Chang, J. 2004. Interpolating, extrapolating, and comparing incidence-based species accumulation curves. Ecology. 85(10): 2717-2727.