Ecology, 89(7), 2008, pp. 2052–2060

AN INCIDENCE-BASED RICHNESS ESTIMATOR FOR QUADRATS SAMPLED WITHOUT REPLACEMENT

Tsung-Jen Shen[1,3] and Fangliang He[2]

[1]*Department of Applied Mathematics and Institute of Statistics, National Chung Hsing University, 250 Kuo Kuang Road, Tai-Chung, Taiwan*

[2]*Department of Renewable Resources, University of Alberta, Edmonton, Alberta T6G 2H1 Canada*

Abstract. Most richness estimators currently in use are derived from models that consider sampling with replacement or from the assumption of infinite populations. Neither of the assumptions is suitable for sampling sessile organisms such as plants where quadrats are often sampled without replacement and the area of study is always limited. In this paper, we propose an incidence-based parametric richness estimator that considers quadrat sampling without replacement in a fixed area. The estimator is derived from a zero-truncated binomial distribution for the number of quadrats containing a given species (e.g., species *i*) and a modified beta distribution for the probability of presence–absence of a species in a quadrat. The maximum likelihood estimate of richness is explicitly given and can be easily solved. The variance of the estimate is also obtained. The performance of the estimator is tested against nine other existing incidence-based estimators using two tree data sets where the true numbers of species are known. Results show that the new estimator is insensitive to sample size and outperforms the other methods as judged by the root mean squared errors. The superiority of the new method is particularly noticeable when large quadrat size is used, suggesting that a few large quadrats are preferred over many small ones when sampling diversity.

Key words: maximum likelihood estimate; modified beta distribution; presence–absence; quadrat sampling; richness estimator; sampling without replacement; zero-truncated binomial distribution.

Introduction

The number of species (or richness) in an area is the most basic diversity measurement. Although complete enumeration of species might be possible in a relatively small area of a few hectares, richness has to be estimated from sampling in situations where census is not feasible. Many methods have been developed to address this problem (Palmer 1990, Bunge and Fitzpatrick 1993, Colwell and Codington 1994, Hellmann and Fowler 1999, Chiarucci et al. 2003, Chao 2005, Magnussen et al. 2006). In general, there are three types of estimators: species–area curves, abundance-based methods, and incidence-based methods. Abundance-based methods require information about abundance, while incidence-based methods are based on species presence–absence data. These estimators are not only designed for handling different types of data but also differ in their mathematical assumptions. Some of them are derived purely from mathematical inspiration with little practical significance (e.g., the multinomial estimator), while others are developed from practical consideration. As a result, their performances vary considerably. Comparison analysis for many of the estimators has been conducted using various empirical data (Palmer 1990, Colwell and Codington 1994, Chazdon et al. 1998, Magnussen et al. 2006), but there is still lack of general consensus. Among the many criteria used to judge estimators, unbiasedness and insensitivity to sample size are most basic. The latter is particularly useful for the practical purpose. Given the difficulty and cost of sampling, it is highly desired to develop estimators that can give a reasonable estimate even when the sample size is small.

Sampling with and without replacement are two basic sample devices. It is intuitive that sampling without replacement is more efficient as long as sedentary organisms (e.g., plants) are concerned, although the majority of the methods currently in use are actually designed for sampling with replacement or for infinite populations. This reflects the historical fact that most richness estimators were initially derived from the mark–recapture method for animals. Schreuder et al. (1999) attempted to adjust this type of estimator by adding a finite population correction term; see also Magnussen et al. (2006). But this has not proven to be successful (see *Empirical test*). As we will show later, if these methods are unconditionally applied to data from sampling without replacement, considerable overestimation can result if sampling proportion, denoted by $q = t/T$, becomes large (see *The proposed model*).

In this study, we developed an incidence-based richness estimator for quadrats sampled without replacement (see Plate 1). Our new method only requires

Manuscript received 17 September 2007; revised 19 November 2007; accepted 21 November 2007. Corresponding Editor: A. M. Ellison.

[3] E-mail: tjshen@amath.nchu.edu.tw

data on species presence–absence in each quadrat. The method was derived from two basic assumptions: (1) the number of occupied quadrats of a species follows a zero-truncated binomial distribution, and (2) the probability of presence/absence of a species in a quadrat follows a modified beta distribution. It is an unbiased estimator when the parameters of the modified beta distribution are given, and its variance is also given. A comparison analysis against the other nine incidence-based estimators using two large-scale empirical data sets has shown that our method is relatively insensitive to sample size and outperforms the other estimators.

The Proposed Model

Assume that there are S species labeled by 1, 2, …, S in a fixed region that could be divided into T disjoint quadrats with roughly equal areas. If each species within the sampled quadrats is registered as present or absent, let Φ_i stand for the total number of quadrats with species i in the study region. Note that a realization of Φ_i could be viewed as a measure of the degree of species i scattering on the region. To ensure all species being present when all quadrats in a study region are surveyed, the *zero* point of any qualified probabilistic models must be eliminated. A simple assumption on Φ_i could be given by following a zero-truncated binomial distribution with parameters T and p_i, $i = 1, 2, \ldots, S$. The conditional probability mass function (pmf) of Φ_i given p_i could be formulated as

$$P(\Phi_i = \varphi | p_i) = \binom{T}{\varphi} \frac{p_i^{\varphi}(1-p_i)^{T-\varphi}}{1-(1-p_i)^T} \qquad (1)$$

$$\varphi = 1, 2, \ldots, T.$$

Suppose that a simple random sample of t quadrats is chosen without replacement from the region composed of T quadrats. It is self-evident that once the other $T - t$ quadrats are exhaustively surveyed, the total number of species present in the region will become known. Let X_i denote the number of quadrats of species i in a sample of t quadrats; the conditional pmf of X_i given both Φ_i and p_i is a hypergeometric distribution:

$$P(X_i = x | \Phi_i, p_i) = \frac{\binom{\Phi_i}{x}\binom{T-\Phi_i}{t-x}}{\binom{T}{t}} \qquad (2)$$

where $\max\{0, t - T + \Phi_i\} \leq x \leq \min\{\Phi_i, t\}$.

When Eq. 2 is averaged over all possible realizations of Φ_i, the conditional pmf of X_i only given p_i can be derived from $E_{\Phi_i}[P(X_i = x | \Phi_i, p_i)]$ and explicitly expressed as

$$P(X_i = x | p_i) = \binom{t}{x} \frac{p_i^x (1-p_i)^{t-x}}{1-(1-p_i)^T} - \frac{(1-p_i)^T I(x=0)}{1-(1-p_i)^T}$$

$$x = 0, 1, 2, \ldots, t. \qquad (3)$$

where $I(\cdot)$ is an indicator function. Once all quadrats are sampled from the region, i.e., $t = T$, this condition distribution is exactly the same as the one in Eq. 1. To reduce the number of parameters in Eq. 3 and to simplify the derivation of the unconditional distribution $P(X_i)$ from Eq. 3, we consider p_1, p_2, …, and p_S as a random sample from a distribution with the probability density function (pdf) being a modified beta distribution:

$$\pi(p) = K(\alpha, \beta)[1 - (1-p)^T]p^{\alpha-1}(1-p)^{\beta-1}$$

$$0 < p < 1$$

where $\alpha > 0$, $\beta > 0$, and

$$K(\alpha, \beta) = \left[\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} - \frac{\Gamma(\alpha)\Gamma(\beta+T)}{\Gamma(\alpha+\beta+T)}\right]^{-1}$$

is a normalizing factor. Note this modified beta distribution inherits almost all the properties of the standard beta distribution which is widely used to model data of range [0, 1]. Furthermore, these two distributions are equivalent when T is large. Consequently, the unconditional distribution of X_i is given by

$$P(X_i = x) = \begin{cases} K(\alpha, \beta)\binom{t}{x}\dfrac{\Gamma(x+\alpha)\Gamma(t+\beta-x)}{\Gamma(t+\alpha+\beta)} & x > 0 \\[2ex] K(\alpha, \beta)\left[\dfrac{\Gamma(\alpha)\Gamma(t+\beta)}{\Gamma(t+\alpha+\beta)} - \dfrac{\Gamma(\alpha)\Gamma(T+\beta)}{\Gamma(T+\alpha+\beta)}\right] & x = 0. \end{cases} \qquad (4)$$

Richness Estimators

Let $f_k = \Sigma_{i=1}^{S} I(X_i = k)$ represent the number of species observed in exactly k quadrats out of the sample, $k = 1, 2, \ldots, t$. Note that f_0 is unobservable if $t < T$. Under the proposed model in the previous section, $(f_0, f_1, f_2, \ldots, f_t)$ is a multinomial distribution with a total S and probabilities (ρ_0, ρ_1, ρ_2, …, ρ_t) subject to $\Sigma_{k=0}^{t} \rho_k = 1$, where $\rho_k = P(X_i = k)$ in Eq. 4 for brevity. Consequently, the likelihood function could be explicitly expressed as

$$L(S, \alpha, \beta) = \frac{S!}{(S-D)!\prod_{k=1}^{t} f_k!} \rho_0^{S-D} \prod_{k=1}^{t} \rho_k^{f_k} \qquad (5)$$

where $D = \Sigma_{k=1}^{t} f_k$ denotes the number of distinct species in the sample. According to the basis of the likelihood function on the observed number of species D or on (f_1, f_2, …, f_t), the likelihood function in Eq. 5 can be decomposed into two ingredients and expressed as $L(S, \alpha, \beta) = L_b(S, \alpha, \beta) \times L_c(\alpha, \beta)$ (Sanathanan 1972, 1977), where

$$L_{\mathrm{b}}(S, \alpha, \beta) = \frac{S!}{(S-D)!D!}(1-\rho_0)^D \rho_0^{S-D}$$

and

$$L_{\mathrm{c}}(\alpha, \beta) = \frac{D!}{\prod_{k=1}^{t} f_k!} \prod_{k=1}^{t} \left(\frac{\rho_k}{1-\rho_0}\right)^{f_k}. \qquad (6)$$

$L_{\mathrm{b}}(S, \alpha, \beta)$ can be easily recognized as a binomial likelihood function with respect to D and $L_{\mathrm{c}}(\alpha, \beta)$ is a conditional likelihood function based on $(f_1, f_2, \ldots, f_t)$ only. Following Sanathanan's (1972, 1977) results, the parameters estimation could be derived from two versions of the likelihood function depicted as follows:

1) Unconditional MLE, maximum likelihood estimate (UMLE): the MLE of the parameters (S, α, β) is directly searched by maximizing Eq. 5 with respect to all parameters at once.

2) Conditional MLE (CMLE): unlike the procedure of UMLE that only has one step to evaluate, the process of the CMLE involves two steps. One needs to find the MLE (α, β) first based on $L_{\mathrm{c}}(\alpha, \beta)$, then maximize $L_{\mathrm{b}}(S, \hat{\alpha}, \hat{\beta})$ with respect to S to obtain the CMLE of S, denoted by

$$\hat{S}_{\mathrm{CMLE}} = D \left[\frac{1 - \dfrac{\Gamma(\hat{\alpha}+\hat{\beta})}{\Gamma(\hat{\beta})} \dfrac{\Gamma(T+\hat{\beta})}{\Gamma(T+\hat{\alpha}+\hat{\beta})}}{1 - \dfrac{\Gamma(\hat{\alpha}+\hat{\beta})}{\Gamma(\hat{\beta})} \dfrac{\Gamma(t+\hat{\beta})}{\Gamma(\mathrm{t}+\hat{\alpha}+\hat{\beta})}} \right]. \qquad (7)$$

If $\hat{\alpha}$ and $\hat{\beta}$ of Eq. 7 are replaced by the true values of α and β, $\hat{S}_{\mathrm{CMLE}}$ is a conditional, unbiased estimator of S in the sense that $E[\hat{S}_{\mathrm{CMLE}} \mid (\hat{\alpha}, \hat{\beta}) = (\alpha, \beta)] = S$. It is, however, worthwhile to mention that the difference between the two versions of the MLE's with respect to S is negligible based on our simulation and the results (not shown) using the two data sets in *Empirical test*. The conditional MLE (Eq. 7) is analytically and computationally simpler. For computational convenience, we will only use the conditional likelihood function $L_{\mathrm{c}}(\alpha, \beta)$ to estimate S in Eq. 7. When $\hat{\alpha}$ goes to zero, the proposed estimator (Eq. 7) is still well defined and has an alternative form of

$$\hat{S}^{*}_{\mathrm{CMLE}} = D \left[\frac{\dfrac{\Gamma'(T+\hat{\beta})}{\Gamma(T+\hat{\beta})} - \dfrac{\Gamma'(\hat{\beta})}{\Gamma(\hat{\beta})}}{\dfrac{\Gamma'(t+\hat{\beta})}{\Gamma(t+\hat{\beta})} - \dfrac{\Gamma'(\hat{\beta})}{\Gamma(\hat{\beta})}} \right]$$

where $\Gamma'(x)$ is the first derivative of the gamma function with respect to x.

To derive the variance of the proposed estimator (Eq. 7), the variance decomposition formula can be applied to $\hat{S}_{\mathrm{CMLE}}$ with respect to the observed number of species D:

$$\begin{aligned} \mathrm{Var}(\hat{S}_{\mathrm{CMLE}}) &= E[\mathrm{Var}(\hat{S}_{\mathrm{CMLE}}|D)] + \mathrm{Var}[E(\hat{S}_{\mathrm{CMLE}}|D)] \\ &\approx E(D^2)\mathrm{Var}[g(\hat{\alpha}, \hat{\beta})|D] + \mathrm{Var}(D)(E[g(\hat{\alpha}, \hat{\beta})|D])^2 \end{aligned}$$

where $g(\hat{\alpha}, \hat{\beta}) = \hat{S}_{\mathrm{CMLE}}/D$ is a function of $\hat{\alpha}$ and $\hat{\beta}$ only. In addition to directly estimating $E(D^2)$ and $E[g(\hat{\alpha}, \hat{\beta}) \mid D]$ by the method of moment, there are two more terms which need to be estimated in the approximate formula of Eq. 8. Since D is approximately distributed from a binomial distribution with parameters S and $E(D)/S$, an approximate variance of D is accordingly given by $\mathrm{Var}(D) \approx E[D](1 - E[D]/S)$. On the other hand, $\mathrm{Var}[g(\hat{\alpha}, \hat{\beta}) \mid D]$ has an approximate formula:

$$\mathrm{Var}[g(\hat{\alpha}, \hat{\beta})|D] \approx \left(\frac{\partial g}{\partial \alpha}, \frac{\partial g}{\partial \beta}\right)[I(\alpha, \beta)]^{-1}\left(\frac{\partial g}{\partial \alpha}, \frac{\partial g}{\partial \beta}\right)^{\top} \qquad (9)$$

where $I(\alpha, \beta)$ is the Fisher information matrix of (α, β) with respect to the likelihood function $L_{\mathrm{c}}(\alpha, \beta)$. The variance estimator of $\hat{S}_{\mathrm{CMLE}}$ is then completed as follows:

$$\begin{aligned} \widehat{\mathrm{Var}}(\hat{S}_{\mathrm{CMLE}}) &= D^2 \widehat{\mathrm{Var}}[g(\hat{\alpha}, \hat{\beta})|D] \\ &\quad + D(1 - D/\hat{S}_{\mathrm{CMLE}})g^2(\hat{\alpha}, \hat{\beta}) \qquad (10) \end{aligned}$$

where $\widehat{\mathrm{Var}}[g(\hat{\alpha}, \hat{\beta}) \mid D]$ is obtained by substituting the CMLE $(\hat{\alpha}, \hat{\beta})$ into Eq. 9. As will be seen in the empirical test below, the performance of the variance estimator (Eq. 10) is satisfactory for most scenarios considered there.

Evaluation of the Estimator

Other incidence-based estimators

We now test and compare the performance of our estimator (Eq. 7) against three major incidence-based estimators and their corrected forms.

1) The first-order jackknife estimator (Heltshe and Forrester 1983):

$$\tilde{S}_{\mathrm{jack1}} = D + \left(\frac{t-1}{t}\right) f_1.$$

2) The bootstrap estimator (Smith and van Belle 1984):

$$\tilde{S}_{\mathrm{boot}} = D + \sum_{k=1}^{t} f_k \left(1 - \frac{k}{t}\right)^t.$$

3) The Chao2 estimator (Chao 1987, Colwell and Coddington 1994):

$$\tilde{S}_{\mathrm{Chao2}} = D + \left(\frac{t-1}{t}\right) \frac{f_1^2}{2f_2}.$$

As stated above, for a random sample of quadrats taken without replacement, these estimators will overestimate species richness when the sampling proportion is large. For example, when $t = T$, none of the three estimators will attain the true species richness S unless

$$\sum_{i=1}^{S} I(\Phi_i = 1) = 0$$

with respect to $\tilde{S}_{\mathrm{jack1}}$ and $\tilde{S}_{\mathrm{Chao2}}$, and

$$\sum_{i=1}^{S} I(\Phi_i = k) = 0$$

for all k with respect to $\tilde{S}_{\text{boot}}$. The magnitude of the bias of $\tilde{S}_{\text{jack1}}$, $\tilde{S}_{\text{boot}}$, and $\tilde{S}_{\text{Chao2}}$ are, respectively,

$$[(T-1)/T]\sum_{i=1}^{S} I(\Phi_i = 1)$$

$$\sum_{k=1}^{T}\sum_{i=1}^{S} I(\Phi_i = k)(1-k/T)^T$$

and

$$[(T-1)/T]\left[\sum_{i=1}^{S} I(\Phi_i = 1)\right]^2 \Big/ \left[2\sum_{i=1}^{S} I(\Phi_i = 2)\right]$$

at $t = T$. This explains why $\tilde{S}_{\text{jack1}}$ and $\tilde{S}_{\text{boot}}$ overestimate the true species richness when the sampling percentages are more than 70% in Hellmann and Fowler (1999). The resulting positive biases for the most estimators in the study of Hellmann and Fowler (1999) are due to the sampling device (sampling quadrats without replacement [Chiarucci et al. 2003:292]) or due to the assumption of infinite population (Schreuder et al. 1999).

To overcome this problem, Schreuder et al. (1999) use a finite population correction term to adjust those estimators, say $\tilde{S}_{\text{any}}$, whose derivations are based on sampling with replacement. The corrected formula is

$$\hat{S}_{\text{any}} = D + (1 - t/T)(\tilde{S}_{\text{any}} - D). \tag{11}$$

This will lower the number of unseen species so that the corrected estimators attain the true richness when $t = T$. We denote the corrected formula of the above three estimators as $\hat{S}_{\text{jack1}}$, $\hat{S}_{\text{boot}}$, and $\hat{S}_{\text{Chao2}}$.

Mingoti and Meeden (1992) assume that the number of quadrats with species i in the sample, X_i, is from a beta-binomial distribution with parameters $a > 0$ and $b > 0$ for the beta distribution. Based on this parametric mixture model, they proposed a richness estimator:

$$\hat{S}_{\text{MM}} = D + \frac{f_1}{t\hat{a}}(t + \hat{b} - 1)\left[1 - \frac{\Gamma(t+\hat{a}+\hat{b})}{\Gamma(t+\hat{b})}\frac{\Gamma(T+\hat{b})}{\Gamma(T+\hat{a}+\hat{b})}\right]$$

where $\hat{a}$ and $\hat{b}$ are the maximum likelihood estimators of a and b, respectively. When $\hat{a}$ goes to zero, the estimator is still well-defined and given by

$$\hat{S}^*_{\text{MM}} = D + \frac{f_1}{t}(t + \hat{b} - 1)\sum_{j=t}^{T-1}\frac{1}{j+\hat{b}}.$$

Haas et al. (2006) proposed a series of nonparametric estimators based on the generalized jackknife procedure. However, their basic model is different from ours in the way that they assumed X_i comes from a binomial distribution with parameters $\Phi_i > 0$ and $q = t/T$. Therefore, samples from their distribution are approximate to a simple random sample when T is large enough. In contrast, ours is an exact model rather than an approximation. Under the assumption that all $\mathbf{\Phi} = (\Phi_1, \Phi_2, \ldots, \Phi_S)$ are equal (very unlikely in reality), they presented an estimator:

$$\hat{S}_{j1} = D\left[1 - \frac{(1-q)f_1}{\sum_{i=1}^{t} if_i}\right]^{-1} \tag{12}$$

which is referred to as the first-order jackknife estimator in their paper. If $\mathbf{\Phi}$ are heterogeneous, they added one more term, involving the coefficient of variation (CV) of $\mathbf{\Phi}$, to Eq. 12 in order to reduce the bias of $\hat{S}_{j1}$. The corrected estimator is

$$\hat{S}_{j2} = \frac{D - (1-q)\ln(1-q)\hat{\gamma}^2 f_1/q}{1 - (1-q)f_1/\sum_{i=1}^{t} if_i}$$

where

$$\hat{\gamma}^2 = \max\left[0, \hat{S}_{j1}\frac{\sum_{i=1}^{t} i(i-1)f_i}{\left(\sum_{i=1}^{t} if_i\right)^2} + \frac{q\hat{S}_{j1}}{\sum_{i=1}^{t} if_i} - 1\right]$$

is the estimate of squared CV of $\mathbf{\Phi}$. This estimator is called as the second-order jackknife estimator in Haas et al. (2006). As will be seen below, the estimator $\hat{S}_{j1}$ systematically underestimates the true species richness for the two test data sets. $\hat{S}_{j2}$ improves the estimation but considerable bias still remains.

Empirical test

Two data sets of large forest plots are used to test and compare our estimator with those introduced previously. The data are census data and are common in their field survey protocol. In each plot, all free-standing trees and shrubs ≥1 cm diameter at breast height were enumerated, individually located on a reference map, and identified to species.

The Barro Colorado Island (BCI) plot is 50 ha (1000 × 500 m), located in Barro Colorado Island, Panama (Condit et al. 1996). Six censuses have so far been surveyed. The 1985 census is used here. There are 299 species and 238 018 individuals. The Pasoh plot is 50 ha (1000 × 500 m), locating in Pasoh Forest Reserve, Malaysia (Kochummen et al. 1990, Abdul Rahim et al. 2004). The first census conducted during 1985–1987 was used in this study. There are 817 species and 320 904 individuals.

The performance of the estimators can be tested by taking quadrat samples without replacement from each plot. Four sizes of quadrats are used: 5 × 5 m, 10 × 10 m,

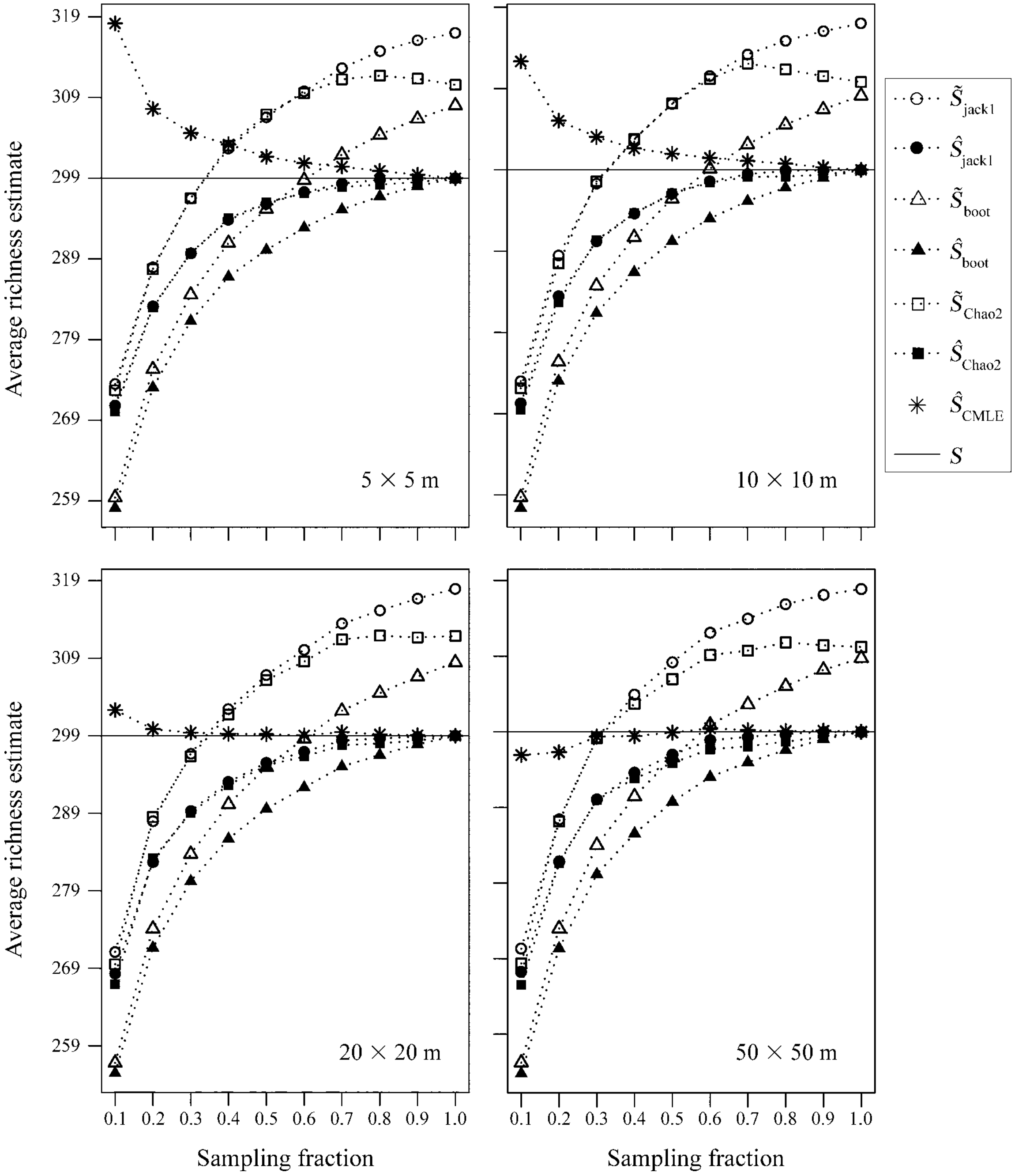

FIG. 1. The Barro Colorado Island (BCI), Panama plot with species richness in relation to four quadrat sizes (range 5 × 5 m to 50 × 50 m). Sampling fraction is the proportion of quadrats sampled.

20 × 20 m, and 50 × 50 m. For each of the four quadrat sizes, six sampling proportions ($q = 0.015$, 0.05, 0.1, 0.2, 0.3, and 0.5) are implemented to investigate the effect of sample sizes on the richness estimators. For a given quadrat size and sampling proportion, 500 replicates are generated. We then compute the mean of observed numbers of species, denoted by $\bar{D}$, the means of the estimates of our proposed estimator ($\hat{\bar{S}}_{\text{CMLE}}$) and other estimators. In total, we compare our method against nine competitors. Six of them are shown in Figs. 1 and 2 for assessing the effect of sampling without replacement on estimators $\tilde{S}_{\text{jack1}}$, $\tilde{S}_{\text{boot}}$, and $\tilde{S}_{\text{Chao2}}$ and their corrected forms $\hat{S}_{\text{jack1}}$, $\hat{S}_{\text{boot}}$, and $\hat{S}_{\text{Chao2}}$. The other three estimators, $\hat{\bar{S}}_{\text{MM}}$, $\hat{\bar{S}}_{j1}$, and $\hat{\bar{S}}_{j2}$, have taken the sample size into account and are compared in Tables 1 and 2.

The results for four quadrat sizes plotted against the sampling proportions are shown in Figs. 1 and 2. It is clear that the three unadjusted estimators, $\tilde{S}_{\text{jack1}}$, $\tilde{S}_{\text{boot}}$,

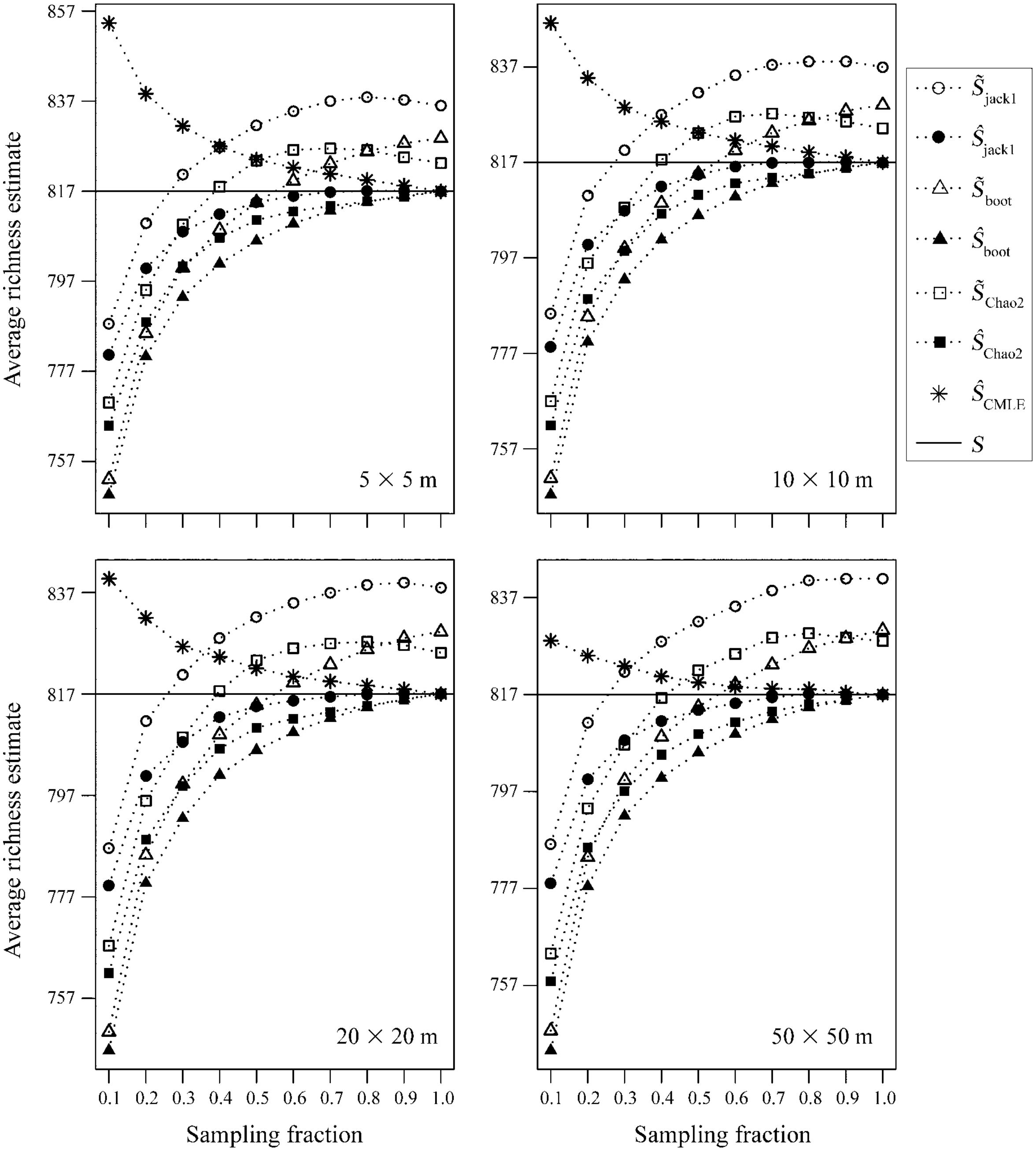

FIG. 2. The Pasoh, Malaysia plot with species richness in relation to four quadrat sizes (range 5 × 5 m to 50 × 50 m). Sampling fraction is the proportion of quadrats sampled.

and $\tilde{S}_{\text{Chao2}}$, quickly overestimate the true richness as the sample proportions increase. The adjusted estimators, $\hat{S}_{\text{jack1}}$, $\hat{S}_{\text{boot}}$, and $\hat{S}_{\text{Chao2}}$, provide the lower mean estimates and approach true richness as sampling intensity increases. However, they are quite sensitive to sampling intensity, especially at small sampling proportions. It appears that $\hat{S}_{\text{CMLE}}$ and $\hat{S}_{\text{Chao2}}$ are the two estimators least sensitive to the effect of sampling intensity.

The comparison of our estimator ($\bar{\hat{S}}_{\text{CMLE}}$) against the methods ($\bar{\hat{S}}_{\text{MM}}$, $\bar{\hat{S}}_{j1}$, and $\bar{\hat{S}}_{j2}$) of Mingoti and Meeden (1992) and Haas et al. (2006) is given in Tables 1 and 2. The estimators are evaluated by the sample root mean squared errors (RMSE):

$$\sqrt{\sum_{i=1}^{500}(\hat{S}_i - S)^2/500}$$

where $\hat{S}_i$ is the computed estimate for a given estimator for the ith replicate and S is the true species richness of a

TABLE 1. For the Barro Colorado Island (BCI), Panama plot, a comparison of the performance of $\hat{S}_{\text{CMLE}}$, $\hat{S}_{\text{MM}}$, $\hat{S}_{j1}$, and $\hat{S}_{j2}$.

Quadrat size (m) and number	q	$\bar{D}$	$\hat{S}_{\text{CMLE}}$ Mean	RMSE	σ	$\bar{\hat{\sigma}}$	$\hat{S}_{\text{MM}}$ Mean	RMSE	$\hat{S}_{j1}$ Mean	RMSE	$\hat{S}_{j2}$ Mean	RMSE
5 × 5, 20 000	0.015	167.2	367.1	69.7	14.9	23.7	363.1	70.8	171.5	127.6	323.9	39.3
	0.05	226.9	337.0	40.7	14.6	21.1	315.2	25.4	227.7	71.4	380.9	87.3
	0.10	246.6	318.2	21.4	9.5	14.8	304.1	13.9	247.0	52.2	389.9	94.9
	0.20	264.0	307.6	10.9	6.7	9.5	299.9	9.7	264.1	35.1	393.8	98.6
	0.30	273.8	304.6	7.6	5.2	7.3	299.7	7.3	273.8	25.4	393.7	97.7
	0.50	285.0	301.7	4.6	3.7	4.9	299.4	4.7	285.1	14.2	383.6	86.6
10 × 10, 5000	0.015	165.6	370.0	74.9	23.9	36.4	361.2	71.2	171.9	127.2	2412	60.5
	0.05	225.2	325.9	31.0	15.4	22.5	310.5	22.5	226.4	72.8	301.0	17.0
	0.10	245.7	312.3	16.8	10.2	14.1	302.1	14.0	246.2	53.0	317.1	24.4
	0.20	263.6	305.0	9.1	6.8	9.2	300.1	9.5	263.9	35.3	330.3	34.6
	0.30	273.5	303.1	6.8	5.4	7.0	300.2	7.6	273.7	25.6	336.0	39.6
	0.50	285.0	301.0	4.2	3.7	4.6	299.7	4.8	285.1	14.2	335.9	38.4
20 × 20, 1250	0.015	162.6	338.7	60.9	46.3	55.5	334.7	60.7	174.1	125.2	200.1	99.8
	0.05	223.2	311.9	20.2	15.5	21.1	302.9	19.5	225.4	73.8	260.9	39.6
	0.10	243.7	302.3	10.3	9.8	13.4	296.9	12.8	244.6	54.6	279.2	22.0
	0.20	261.9	299.9	7.4	7.4	8.9	298.9	10.0	262.3	36.9	297.3	10.0
	0.30	272.1	299.4	5.9	5.9	6.8	299.0	7.8	272.4	26.9	305.0	11.0
	0.50	284.2	299.2	4.0	4.0	4.5	299.0	5.0	284.3	15.0	311.1	14.0
50 × 50, 200	0.015	168.6	308.7	61.0	60.3	59.9	311.3	64.3	201.2	98.7	201.2	98.7
	0.05	221.3	301.1	18.0	17.9	22.9	297.5	19.8	227.4	71.9	237.2	62.3
	0.10	241.4	295.9	12.7	12.3	14.1	294.6	15.4	244.1	55.1	256.6	43.0
	0.20	260.2	296.3	9.1	8.7	9.1	297.5	11.2	261.4	37.9	275.7	24.5
	0.30	271.4	298.4	6.4	6.3	7.0	300.1	8.1	272.1	27.2	286.8	13.7
	0.50	284.0	298.9	4.3	4.3	4.5	299.6	5.2	284.3	15.1	296.6	5.5

Notes: There are 299 observed tree species. Quadrat number is the total number of quadrats in the BCI plot for the given quadrat size, RMSE is root mean squared error, q is the proportion of quadrats sampled, and $\bar{D}$ is the richness averaged over the sampled quadrats; $\hat{S}_{\text{CMLE}}$, proposed richness estimator; $\hat{S}_{\text{MM}}$, Mingoti and Meeden's estimator; $\hat{S}_{j1}$, the first-order jackknife estimator by Haas et al.; $\hat{S}_{j2}$, the second-order jackknife estimator by Haas et al.; σ, sample standard error; $\bar{\hat{\sigma}}$, average estimated standard error.

given plot (e.g., $S = 299$ for the BCI plot). The sample standard errors σ and the mean of the estimated standard errors $\bar{\hat{\sigma}}$ for the variance estimator (Eq. 10) are also included in the tables.

As the sampling proportions increase, $\bar{\hat{S}}_{\text{CMLE}}$ and $\bar{\hat{S}}_{\text{MM}}$ consistently approach the true richness and their RMSE correspondingly decreases with q (Tables 1 and 2). The performance of the two nonparametric estima-

TABLE 2. For the Pasoh, Malaysia plot, a comparison of the performance of $\hat{S}_{\text{CMLE}}$, $\hat{S}_{\text{MM}}$, $\hat{S}_{j1}$, and $\hat{S}_{j2}$.

Quadrat size (m) and number	q	$\bar{D}$	$\hat{S}_{\text{CMLE}}$ Mean	RMSE	σ	$\bar{\hat{\sigma}}$	$\hat{S}_{\text{MM}}$ Mean	RMSE	$\hat{S}_{j1}$ Mean	RMSE	$\hat{S}_{j2}$ Mean	RMSE
5 × 5, 20 000	0.015	488.7	1025.5	223.1	79.5	104.5	992.3	193.7	512.9	304.3	768.9	60.7
	0.05	665.2	875.8	62.6	21.4	30.1	855.2	46.5	669.1	148.1	899.0	86.9
	0.10	718.8	854.3	39.9	14.0	19.2	839.3	29.1	720.4	96.9	915.8	102.3
	0.20	759.8	838.6	23.3	8.5	12.2	825.0	14.5	760.3	56.9	904.1	89.7
	0.30	778.5	831.5	16.0	6.7	9.1	821.4	10.3	778.8	38.5	895.1	80.4
	0.50	797.5	824.1	8.4	4.5	5.8	818.7	6.1	797.5	19.8	877.8	62.3
10 × 10, 5000	0.015	483.5	978.3	181.5	83.3	100.4	955.7	164.0	513.5	303.8	683.0	137.5
	0.05	661.9	861.9	50.0	22.0	29.2	846.2	39.7	666.7	150.5	826.9	25.0
	0.10	716.3	846.2	32.1	13.3	18.8	833.9	24.4	718.3	99.0	855.3	42.9
	0.20	758.4	834.7	20.0	9.4	11.9	824.6	14.8	759.1	58.2	863.8	50.0
	0.30	777.3	828.5	13.2	6.4	8.9	819.7	9.4	777.7	39.6	860.1	45.4
	0.50	797.5	823.4	7.9	4.6	5.7	819.0	6.4	797.6	19.7	855.8	40.4
20 × 20, 1250	0.015	472.7	916.3	134.1	90.2	101.9	907.4	129.0	519.9	297.7	598.9	219.8
	0.05	657.1	849.4	40.5	24.3	29.2	840.8	37.5	664.6	152.7	759.4	60.5
	0.10	714.2	839.7	27.2	15.0	18.7	833.0	25.0	717.2	100.1	801.2	22.1
	0.20	757.8	832.0	17.3	8.7	11.9	824.9	14.3	758.8	58.5	823.9	13.7
	0.30	776.7	826.4	11.5	6.7	8.8	820.2	9.9	777.2	40.1	829.0	16.1
	0.50	797.0	822.0	6.7	4.5	5.6	818.4	6.0	797.2	20.2	832.8	17.6
50 × 50, 200	0.015	492.8	837.8	139.6	138.1	130.1	838.2	140.6	608.6	211.7	608.6	211.7
	0.05	651.5	824.8	31.4	30.4	30.8	825.0	33.7	670.4	147.3	697.0	121.2
	0.10	709.0	828.1	20.1	16.7	19.4	827.6	22.0	716.7	100.8	748.0	70.0
	0.20	754.2	825.0	12.8	10.0	12.0	821.9	13.0	756.9	60.5	783.9	34.2
	0.30	775.3	822.8	9.4	7.3	8.8	819.2	9.4	776.6	40.8	798.9	19.5
	0.50	795.8	819.4	5.4	4.8	5.5	817.6	5.8	796.2	21.2	812.0	7.4

Notes: There are 817 observed tree species. See Table 1 for definitions of parameters.

PLATE 1. Gutianshan nature Reserve (81 km^2), Zhejiang, mainland China. A 24-ha stem-mapped plot is established to address questions including the number of different tree species present. Photo credit: F. He.

tors, $\hat{S}_{j1}$ and $\hat{S}_{j2}$, is inferior to $\tilde{\hat{S}}_{\mathrm{CMLE}}$ and $\tilde{\hat{S}}_{\mathrm{MM}}$ in terms of RMSE and bias except for small quadrat size along with the smallest sampling proportion. As expected, $\hat{S}_{j1}$ systematically underestimates the true species richness for all plots (Tables 1 and 2). Overestimation is also observed for the bias corrected $\hat{S}_{j2}$ of Haas et al. (2006).

$\hat{S}_{\mathrm{MM}}$ is the only method that is comparable to our estimator in both RMSE and bias (Tables 1 and 2). The relative small bias of $\hat{S}_{\mathrm{MM}}$ is partially due to its derivation by the method of moment. Our proposed estimator, $\hat{S}_{\mathrm{CMLE}}$, is a likelihood-based estimator; it has asymptotic properties. Our method consistently outperforms other estimators when quadrat size is relatively large. This provides an advantage in field sampling because it is easier to set up and survey a few larger quadrats than many small ones.

DISCUSSION

Counting species presence or absence in quadrats is a practical and convenient sampling scheme. The majority of richness estimators in the literature are derived from probabilistic models considering sampling with replacement or assuming infinite population size (Bunge and Fitzpatrick 1993, Chao 2005). Such a sampling scheme, however, is seldom applied to sampling sessile organisms such as plants where sampling without replacement, instead, is used. It has been thought that an estimator deriving from a model of sample-without-replacement would usually involve combinations (or factorials) like those in Eq. 2, thus making the derivation mathematically cumbersome. But the new model proposed in this study has got over this awkward problem.

Empirical tests using the two real data sets have shown the superior performance of the proposed estimator (Eq. 7) to other existing estimators (Tables 1 and 2, Figs. 1 and 2). The superiority is particularly noticeable when quadrat size is large. For example, at 50 × 50 m, $\hat{S}_{\mathrm{CMLE}}$ very quickly approaches true number of species for the two data sets and in most cases its RMSE is smallest among all other compared estimators. This feature suggests that a few large quadrats should be preferred over many small quadrats. This is clearly a welcoming advantage in field sampling.

Most richness estimators, such as $\tilde{S}_{\mathrm{boot}}$ and $\tilde{S}_{\mathrm{Chao2}}$, were derived from models of sampling with replacement or from the assumption of infinite population size. As shown in this study, if such estimators are unconditionally applied to data of sample-without-replacement, they will overestimate the true richness when the proportion of sampled quadrats is large. This violates the basic property of that an estimator should approach the true value when sample sizes approach to the entire study region. Although Schreuder et al. (1999) attempt to correct such estimators by adjusting the number of unseen species with a finite population correction term like Eq. 11, the performance of the corrected estimators still do not work as well as our method (Figs. 1 and 2).

Mingoti and Meeden (1992) propose a richness estimator based on the beta-binomial model. Their model allows species, present in the sampled quadrats, to have probabilities to be unseen. Based on their beta-binomial model, it does not require all species to appear when all quadrats are surveyed. Although this assumption may be reasonable for animal species like birds (birds may not be seen even entire area is surveyed because of their mobility), it may not be appropriate for sessile species. Haas et al. (2006) also devote to the same topic and propose a range of estimators derived from the generalized jackknife method, but their model was based an approximation in order to avoid combinations like those in Eq. 2. Our results in almost all cases of Tables 1 and 2 show that $\hat{S}_{\mathrm{CMLE}}$ outperforms $\hat{S}_{j1}$ and $\hat{S}_{j2}$. While

$\hat{S}_{\text{CMLE}}$ is inferior to $\hat{S}_{\text{MM}}$ at small quadrat sizes, it is indistinguishable from it at intermediate quadrat sizes but superior at large sizes. Another advantage of our estimator is that the variance for $\hat{S}_{\text{CMLE}}$ is available. This is not the case in other estimators in Tables 1 and 2.

In practice, two crucial issues often arise: (1) How large should the sampling quadrat be? and (2) How many quadrats should be taken? Based on our results (Tables 1 and 2 and Figs. 1 and 2), we would recommend using at least 20 × 20 m, preferably 50 × 50 m, quadrat size if the purpose is to estimate richness. In terms of the number of quadrats to be sampled, the minimum is three (otherwise, the model becomes statistically unidentifiable). The price to pay for taking a small sample size is the inflation of the estimated standard error. It is clear from Tables 1 and 2 and Figs. 1 and 2 that quadrat size and sample size compensate each other. If quadrat size is large, we can use relatively small sample size to attain the same level of accuracy and precision. As a thumb of rule, based on our results, we would recommend a minimum sampling scheme: 20 × 20 m quadrat size with 10% sampling intensity.

Acknowledgments

The authors thank the Center for Tropical Forest Science for generously providing the BCI and Pasoh data. The large-scale forest plot at Pasoh Forest Reserve is an ongoing project of the Malaysian Government, initiated by the Forest Research Institute Malaysia through its Director General, and under the leadership of N. Manokaran, P. S. Ashton, and S. P. Hubbell. The authors thank M. N. Nur Supardi for helpful comments on an early draft. The work was supported by grants from the National Science Council of Taiwan to T. J. Shen, and by the Sustainable Forest Management Network of Canada and the NSERC (Canada) to F. He.

Literature Cited

Abdul Rahim, N., M. N. Nur Supardi, N. Manokaran, S. J. Davies, J. V. La Frankie, P. S. Ashton, and T. Okuda. 2004. Demographic tree data from the 50-ha Pasoh forest dynamics plot. [CD-ROM] CTFS Forest Dynamics Plot Data Series, Kepong, Malaysia.

Bunge, J., and M. Fitzpatrick. 1993. Estimating the number of species: a review. Journal of the American Statistical Association 88:364–373.

Chao, A. 1987. Estimating the population size for capture–recapture data with unequal catchability. Biometrics 43: 783–791.

Chao, A. 2005. Species richness estimation. Pages 7907–7916 *in* N. Balakrishnan, C. B. Read, and B. Vidakovic, editors. Encyclopedia of statistical sciences. Second edition, volume 12. Wiley, New York, New York, USA.

Chazdon, R. L., R. K. Colwell, J. S. Denslow, and M. Guariguata. 1998. Statistical estimation of species richness of woody regeneration in primary and secondary rainforests of NE Costa Rica. Pages 285–309 *in* F. Dallmeier and J. Comisky, editors. Forest biodiversity in North, Central, and South America and the Caribbean: research and monitoring. Parthenon Press, Paris, France.

Chiarucci, A., N. J. Enright, G. L. W. Perry, and B. P. Miller. 2003. Performance of nonparametric species richness estimators in a high diversity plant community. Diversity and Distributions 9:283–295.

Colwell, R. K., and J. A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society: Biological Sciences 345: 101–118.

Condit, R., S. P. Hubbell, and R. B. Foster. 1996. Changes in a tropical forest with a shifting climate: results from a 50-ha permanent census plot in Panama. Journal of Tropical Ecology 12:231–256.

Haas, P. J., Y. Liu, and L. Stokes. 2006. An estimator of number of species from quadrat sampling. Biometrics 62: 135–141.

Hellmann, J. J., and G. W. Fowler. 1999. Bias, precision, and accuracy of four measures of species richness. Ecological Applications 9:824–834.

Heltshe, J. F., and N. E. Forrester. 1983. Estimating species richness using the jackknife procedure. Biometrics 39:1–2.

Kochummen, K. M., J. V. LaFrankie, and N. Manokaran. 1990. Floristic composition of Pasoh Forest Reserve, a lowland rain forest in Peninsular Malaysia. Journal of Tropical Forest Science 3:1–13.

Magnussen, S., R. Pelissier, F. He, and B. R. Ramesh. 2006. An assessment of sample-based estimators of tree species richness in two wet tropical forest compartments in Panama and India. International Forestry Review 8:417–431.

Mingoti, S. A., and G. Meeden. 1992. Estimating the total number of distinct species using presence and absence data. Biometrics 48:863–875.

Palmer, M. W. 1990. The estimation of species richness by extrapolation. Ecology 71:1195–1198.

Sanathanan, L. 1972. Estimating the size of a multinomial population. Annals of Mathematical Statistics 42:58–69.

Sanathanan, L. 1977. Estimating the size of a truncated sample. Journal of the American Statistical Association 72:669–672.

Schreuder, H. T., M. S. Williams, and R. M. Reich. 1999. Estimating the number of tree species in a forest community using survey data. Environmental Monitoring and Assessment 56:293–303.

Smith, E. P., and G. V. van Belle. 1984. Nonparametric estimation of species richness. Biometrics 40:119–129.