Statistica Sinica **17**(2007), 761-774

ESTIMATING SPECIES ACCUMULATION CURVES AND DIVERSITY INDICES

Chang Xuan Mao

University of California, Riverside

Abstract: Diversity arises as a significant concept in many scientific fields, and particularly in ecology. The relationships between diversity indices and species accumulation functions are investigated. Nonparametric methods are illustrated with ecological examples.

Key words and phrases: Poisson mixture, rarefaction, species richness.

1. Introduction

Conservation of diversity is an important problem in biological sciences (e.g., Colwell and Coddington (1994)). Diversity also arises as a significant concept in the physical and social sciences (e.g., Patil and Taillie (1982)). Consider a community consisting of c distinct species labeled by $i = 1, \ldots, c$, with $\pi_i \in (0, .5)$ being the abundance of species i. Many diversity indices are used in ecology, paleobiology, entomology and genetics, etc. (e.g., Magurran (1988)), among which the most familiar ones are the number of species c, the Shannon index $-\sum_{i=1}^{c} \pi_i \log \pi_i$, and the Simpson index $\sum_{i=1}^{c} \pi_i^2$. For several theoretical reasons, Patil and Taillie (1982) proposed, for $h \geq -1$,

$$\Delta(h) = \sum_{i=1}^{c} \pi_i r(\pi_i; h), \tag{1}$$

where $r(\pi; h) = h^{-1}(1-\pi^h)$ for $h \neq 0$, and $r(\pi; 0) = -\log \pi$. Note that $\Delta(-1) = c-1$, $\Delta(0) = -\sum_{i=1}^{c} \pi_i \log \pi_i$, and $\Delta(1) = 1-\sum_{i=1}^{c} \pi_i^2$ are special cases of $\Delta(h)$, and the equivalent number $\mathcal{E}(h) = \{1 - h\Delta(h)\}^{-1/h}$ (e.g., MacArthur (1965)) and the entropy $\log \mathcal{E}(h)$ (Renyi (1961)) are transformations of $\Delta(h)$.

In order to quantify diversity, one needs to sample individuals from the community. The numbers of detected individuals from the species can be modeled as a multinomial sample with index c and probabilities π_i. The expected number of species being detected in a sample of m individuals is

$$A(m) = \sum_{i=1}^{c} \big\{1 - (1 - \pi_i)^m\big\}, \; m = 0, 1, \ldots. \tag{2}$$

When the number of detected individuals from species i is modeled as a Poisson process with rate λ_i (e.g., Norris and Pollock (1998) and Mao (2004)), the expected number of species detected during the interval $[0, t]$ is

$$a(t) = \sum_{i=1}^{c} \{1 - \exp(-\lambda_i t)\}, t \geq 0. \tag{3}$$

If $\Theta(\lambda) = c^{-1} \sum_{i=1}^{c} I(\lambda_i \leq \lambda)$, with $I(\cdot)$ being the indicator function, or if the λ_i are assumed to arise as a random sample from Θ, then

$$a(t) = c \int (1 - e^{-\lambda t})\, d\Theta(\lambda), t \geq 0. \tag{4}$$

Both $A(m)$ and $a(t)$ are called species accumulation functions.

The curves $(m, A(m))$ and $(t, a(t))$ provide information to assess the efficacy and completeness of sampling projects and predict the number of new species to be discovered in the future (e.g., Good and Toulmin (1956), Efron and Thisted (1976), de Caprariis, Lindemann and Collins (1976), Soberón and Llorente (1993), Colwell and Coddington (1994), Solow and Polasky (1999) and Shen, Chao and Lin (2003)).

Let Y_i denote the number of individuals from species i in a sample of size $s = \sum_{i=1}^{c} Y_i$. This sample is considered to have been taken during the interval $[0, 1]$. Let $n_x = \sum_{i=1}^{c} I(Y_i = x)$, $n = \sum_{i=1}^{c} I(Y_i > 0)$ and $D = \max\{x : n_x > 0\}$. Based on $\{n_x\}_{x=1}^{D}$, the estimation of $A(m)$ was studied by Hurlbert (1971), Smith and Grassle (1977), Solow and Polasky (1999) and Shen, Chao and Lin (2003), that of $a(t)$ by Good and Toulmin (1956), Efron and Thisted (1976) and Boneh, Boneh and Caron (1998), of $\Delta(0)$ by Zahl (1977) and Chao and Shen (2003), and of $c = \Delta(-1) + 1$ by Chao (1984) among others (see Bunge and Fitzpatrick (1993)).

It will be shown that $\Delta(h)$ is a function of the $A(m)$, leading to an approximation $\delta(h)$ to $\Delta(h)$ in the Poisson model. Besides the representations in (1), (2), (3) and (4), alternative representations of $A(m)$, $\Delta(h)$ and $a(t)$, and several representations of $\delta(h)$, will be constructed. Nonparametric estimators arise from these representations, among which some are new and others correspond to known estimators with new interpretations.

The article is organized as follows. In Section 2, new representations are introduced for the species accumulation functions and the diversity indices. Estimation methods are discussed in Section 3. Examples are investigated and a simulation study is reported in Section 4.

2. The Multinomial And Poisson Models

In the multinomial model, define a distribution $G(\nu)$ by

$$G(\nu) = c^{-1} \sum_{i=1}^{c} I(\pi_i/(1-\pi_i) \le \nu), \nu \in (0, \infty),$$

with generalized moments $\mu_x = \int \nu^x \cdot c(1+\nu)^{-s} dG(\nu)$. The distribution $G(\nu)$ is nonidentifiable because only finitely many μ_x are identifiable,

$$E(n_x) = \sum_{i=1}^{c} \binom{s}{x} \pi_i^x (1-\pi_i)^{s-x} = \binom{s}{x} \mu_x,\ x = 1, 2, \ldots, s,$$

where $\binom{\alpha}{\beta} = (\beta!)^{-1} \prod_{k=0}^{\beta-1} (\alpha - k)$ for a real α and an integer $\beta \ge 0$.

Note that $A(m)$ with $m \in \{x\}_{x=1}^{s}$ and $\Delta(h)$ with $h \in \{x\}_{x=1}^{s-1}$ are identifiable, and that $A(m)$ with $m > s$ and $\Delta(h)$ with $h \in [-1, \infty) \backslash \{x\}_{x=1}^{s-1}$ have an additive decomposition into nonidentifiable and identifiable components.

Proposition 1. *Both $A(m)$ and $\Delta(h)$ are functions of the μ_x, with*

$$A(m) = \sum_{x=1}^{\infty} \left\{ \binom{s}{x} - \binom{s-m}{x} \right\} \mu_x, \tag{5}$$

$$\Delta(h) = \begin{cases} h^{-1} \displaystyle\sum_{m=2}^{\infty} (-1)^m \binom{h+1}{m} A(m) - 1 & (h \ne 0) \\ \displaystyle\sum_{m=2}^{\infty} \frac{A(m)}{m(m-1)} - 1 & (h = 0). \end{cases} \tag{6}$$

The Y_i can be understood as a random sample from $g_\Theta(x) = \int e^{-\lambda} \lambda^x / (x!)\, d\Theta(\lambda)$, $x = 0, 1, \ldots$. Let $dQ(\lambda) = (1 - e^{-\lambda}) d\Theta(\lambda) / \int (1 - e^{-\lambda}) d\Theta(\lambda)$. Conditioning on n, those $Y_i > 0$ arise as a random sample from $f_Q(x) = \int \lambda^x / \{(e^\lambda - 1) x!\}\, dQ(\lambda)$, $x = 1, 2, \ldots$ (e.g., Mao (2004)). $f_Q(x) = \int \lambda^x / \{(e^\lambda - 1) x!\}\, dQ(\lambda)$, $x = 1, 2, \ldots$. Note that $g_\Theta(x)$ and $f_Q(x)$ are generalized moments of Θ and Q respectively. The $f_Q(x)$ are identifiable and Q is completely determined by f_Q. Norris and Pollock (1998) considered estimating Θ and c simultaneously, which is computationally expensive. A much simpler approach is to estimate Q, e.g., by the nonparametric maximum likelihood estimator (MLE) $\widehat{Q}$ (Lindsay (1983) and Mao (2004)).

$$\widehat{Q} = \operatorname{argmax} \Big\{ \sum_{x=1}^{D} n_x \log f_Q(x) : \text{all } Q \Big\}.$$

Let $\theta = E(s)$. The expected number of individuals sampled up to the time $t = m/\theta$ is m, implying that $a(m/\theta) \approx A(m)$ and $\delta(h) \approx \Delta(h)$, where

$$\delta(h) = \begin{cases} h^{-1} \sum_{m=2}^{\infty} (-1)^m \binom{h+1}{m} a\Big(\frac{m}{\theta}\Big) - 1 & (h \neq 0) \\ \sum_{m=2}^{\infty} \frac{a(\frac{m}{\theta})}{m(m-1)} - 1 & (h = 0). \end{cases} \tag{7}$$

Proposition 2. *Both $a(t)$ and $\delta(h)$ are functions of the μ_x, $\delta(h)$ is a function in the λ_i, and $(\delta(h)+1)/a(1)$ and $a(t)/a(1)$ are functionals of Q, with*

$$a(t) = \sum_{x=1}^{\infty} \big\{1 - (1-t)^x\big\} c g_{\Theta}(x), \tag{8}$$

$$a(t) = a(1) \int \frac{1 - e^{-\lambda t}}{1 - e^{-\lambda}}\, dQ(\lambda), \tag{9}$$

$$\delta(h) = \sum_{i=1}^{c} \gamma(\lambda_i, h, \theta) - 1, \tag{10}$$

$$\delta(h) = a(1) \int \frac{\gamma(\lambda, h, \theta)}{1 - e^{-\lambda}}\, dQ(\lambda) - 1, \tag{11}$$

$$\gamma(\lambda, h, \theta) = \begin{cases} (1 + h^{-1})(1 - e^{-\frac{\lambda}{\theta}}) - h^{-1}(1 - e^{-\frac{\lambda}{\theta}})^{h+1} & (h \neq 0) \\ (1 - e^{-\frac{\lambda}{\theta}})\{1 - \log(1 - e^{-\frac{\lambda}{\theta}})\} & (h = 0). \end{cases}$$

Both $A(m)$ in (2) and $\Delta(h)$ in (1) can be written as $\sum_{i=1}^{c} \varphi(\pi_i)$, and $a(t)$ in (3) and $\delta(h)+1$ in (10) can be written as $\sum_{i=1}^{c} \psi(\lambda_i)$. Let $\Omega = \{i : Y_i > 0\}$ denote the random index set of detected species. Because $E\{I(Y_i > 0)\} = 1-(1-\pi_i)^s$ in the multinomial model, and $E\{I(Y_i > 0)\} = 1 - \exp(-\lambda_i)$ in the Poisson model, we can write $\sum_{i=1}^{c} \varphi(\pi_i)$ and $\sum_{i=1}^{c} \psi(\lambda_i)$ as

$$\sum_{i=1}^{c} \varphi(\pi_i) = \sum_{i=1}^{c} \varphi(\pi_i) \frac{E\{I(Y_i > 0)\}}{1 - (1-\pi_i)^s} = E\Big\{\sum_{i \in \Omega} \frac{\varphi(\pi_i)}{1 - (1-\pi_i)^s}\Big\}, \tag{12}$$

$$\sum_{i=1}^{c} \psi(\lambda_i) = \sum_{i=1}^{c} \psi(\lambda_i) \frac{E\{I(Y_i > 0)\}}{1 - \exp(-\lambda_i)} = E\Big\{\sum_{i \in \Omega} \frac{\psi(\lambda_i)}{1 - \exp(-\lambda_i)}\Big\}. \tag{13}$$

3. Inference

3.1. Abundance plug-in estimation

Consider estimators $\hat{\lambda}(x)$ and $\hat{\pi}(x)$ given $Y_i = x$ for λ_i and π_i, respectively. The MLEs for λ_i and π_i are $\hat{\lambda}_{ML}(Y_i)$ and $\hat{\pi}_{ML}(Y_i)$ respectively, where $\hat{\lambda}_{ML}(x) = x$, $\hat{\pi}_{ML}(x) = s^{-1}x, \quad x \geq 0$.

When Θ is treated as a prior, the posterior mean of λ given x is

$$\lambda(x) \equiv E\{\lambda|x\} = (x+1)g_\Theta(x+1)g_\Theta^{-1}(x) = (x+1)f_Q(x+1)f_Q^{-1}(x),$$

where $f_Q(0) = g_\Theta(0)/\{1-g_\Theta(0)\}$ is the odds of the probability of a species being undetected. Given the nonparametic MLE $\widehat{Q}$ for Q, as $\pi_i = \lambda_i/\sum_{k=1}^c \lambda_k$ and s estimates $\sum_{k=1}^c \lambda_k$, the nonparametric empirical Bayes estimators (EBEs) for λ_i and π_i are $\hat{\lambda}_{EB}(Y_i)$ and $\hat{\pi}_{EB}(Y_i)$ respectively, where

$$\hat{\lambda}_{EB}(x) = \frac{(x+1)f_{\widehat{Q}}(x+1)}{f_{\widehat{Q}}(x)}, \quad \hat{\pi}_{EB}(x) = \frac{(x+1)f_{\widehat{Q}}(x+1)}{sf_{\widehat{Q}}(x)}, \quad x \geq 0.$$

In terms of risk, EBEs are better than MLEs in a setting of estimating a collection of parameters simultaneously (Lehmann and Casella (1998, p.272)).

The sample coverage is $\sum_{i=1}^c \pi_i I(Y_i > 0)$, for which a nonparametric EBE is $1 - n_1/s$ (Good (1953)). Chao and Shen (2003) considered an estimator $\hat{\pi}_{SC}(Y_i)$ for π_i, when $Y_i > 0$, as a hybrid of EBE and MLE, where $\hat{\pi}_{SC}(x) = (1 - n_1/s)\hat{\pi}_{ML}(x) = (1 - n_1/s)x/s, \quad x \geq 1$.

One might consider plugging estimators for π_i and λ_i into $A(m)$ in (2), $\Delta(h)$ in (1), $a(t)$ in (3) and $\delta(h)$ in (10), e.g., an estimator for $\Delta(0)$ given by $\widehat{\Delta}_{ML}(0) = -\sum_{i=1}^c \hat{\pi}_{ML}(x)\log\hat{\pi}_{ML}(x) = -\sum_{x=1}^D n_x x/s \log(x/s)$. If $\hat{\pi}_{ML}(x)$ and $\hat{\lambda}_{ML}(x)$ are used, then the estimators for $\Delta(h)$ and $\delta(h)$ are infinity for $h < 0$ and a substantial bias can exist in the estimators for $A(m)$, $a(t)$, $\Delta(h)$ and $\delta(h)$ with $h \geq 0$. Zahl (1977) gave a jackknifed version $\widehat{\Delta}_{JM}(0)$ of $\widehat{\Delta}_{ML}(0)$. There are other bias-corrected versions that depend on an estimator for c (e.g., Boneh, Boneh and Caron (1998)) and should not be used, because they can vary dramatically when different estimators for c are used.

Because of (12) and (13), it is unnecessary to consider estimators for π_i or λ_i when $Y_i = 0$. Given appropriate estimators $\hat{\pi}(Y_i)$ and $\hat{\lambda}(Y_i)$ for π_i and λ_i, respectively, with $i \in \Omega$, the Horvitz-Thompson plug-in estimators are

$$\sum_{i\in\Omega} \frac{\varphi(\hat{\pi}(Y_i))}{1-(1-\hat{\pi}(Y_i))^s}, \quad \sum_{i\in\Omega} \frac{\psi(\hat{\lambda}(Y_i))}{1-\exp(-\hat{\lambda}(Y_i))}.$$

Specifically, the following are estimators for $A(m)$, $\Delta(h)$, $a(t)$ and $\delta(h)$,

$$\widehat{A}(m) = \sum_{x=1}^D n_x \frac{1-(1-\hat{\pi}(x))^m}{1-(1-\hat{\pi}(x))^s}, \qquad \widehat{\Delta}(h) = \sum_{x=1}^D n_x \frac{\hat{\pi}(x)r(\hat{\pi}(x);h)}{1-(1-\hat{\pi}(x))^s},$$

$$\hat{a}(t) = \sum_{x=1}^D n_x \frac{1-\exp(-\hat{\lambda}(x)t)}{1-\exp(-\hat{\lambda}(x))}, \qquad \hat{\delta}(h) = \sum_{x=1}^D n_x \frac{\gamma(\hat{\lambda}(x),h,s)}{1-\exp(-\hat{\lambda}(x))} - 1.$$

We write $\widehat{A}_{EB}(m)$ and $\widehat{\Delta}_{EB}(h)$ if $\hat{\pi}(x) = \hat{\pi}_{EB}(x)$, $\widehat{A}_{SC}(m)$ and $\widehat{\Delta}_{SC}(h)$ if $\hat{\pi}(x) = \hat{\pi}_{SC}(x)$, and $\hat{a}_{EB}(t)$ and $\hat{\delta}_{EB}(h)$ if $\hat{\lambda}(x) = \hat{\lambda}_{EB}(x)$. Note that $\widehat{\Delta}_{SC}(0)$ was proposed in Chao and Shen (2003), while Shen, Chao and Lin (2003) gave an estimator for $A(m)$ that is fully derived from the sample coverage approach.

3.2. Moment plug-in estimation

In the multinomial model, when $1 \le m \le s$, since $E(n_x) = \binom{s}{x}\mu_x$ one can replace μ_x with its estimator $\binom{s}{x}^{-1} n_x$ to obtain an estimator $\check{A}(m)$ for $A(m)$,

$$\check{A}(m) = n - \sum_{x=1}^{s-m} \binom{s-m}{x}\binom{s}{x}^{-1} n_x.$$

This estimator was originally found in Hurlbert (1971). It is a minimum variance unbiased estimator for $A(m)$ (Smith and Grassle (1977)) and is a U-statistic (Mao, Colwell and Chang (2005)). There is an unbiased estimator for $\Delta(h)$ when $h \in \{x\}_{x=1}^{s-1}$, e.g., $\sum_{i=1}^{c} \pi_i^2 = 1 - \Delta(1)$ estimated by $\sum_{x=1}^{s} x(x-1)n_x/\{s(s-1)\}$. There is neither an unbiased estimator for $A(m)$ with $m > s$, nor an unbiased estimator for $\Delta(h)$ with $h \in [-1, \infty)\backslash\{x\}_{x=1}^{s-1}$. For $m > s$, truncating the series in (5) at $x = s$ yields a biased estimator $\check{A}(m)$ for $A(m)$,

$$\check{A}(m) = \sum_{x=1}^{s}\Big\{1 - \binom{s-m}{x}\binom{s}{x}^{-1}\Big\}n_x = \sum_{x=1}^{D}\Big\{1 - \binom{s-m}{x}\binom{s}{x}^{-1}\Big\}n_x.$$

In the Poisson model, one can replace $cg_\Theta(x)$ in (8) with n_x to obtain

$$\check{a}(t) = \sum_{x=1}^{D}\big\{1 - (1-t)^x\big\}n_x = \sum_{x=1}^{D}\big\{1 - (1-t)^x\big\}n_x.$$

The estimator $\check{a}(t)$ for $t > 1$ was proposed in Good and Toulmin (1956).

Although both $A(m)$ and $a(t)$ are monotonic in m or t, neither $\check{A}(m)$ nor $\check{a}(t)$ is necessarily monotonic. The estimator $\check{A}(m)$ for $m \le 2s$, and approaching $2s$, can be huge if D and s are comparable. Neither $\check{A}(m)$ for $m > 2s$ nor $\check{a}(t)$ for $t > 2$ is reliable. When m and t go to infinity, $\check{A}(m)$ and $\check{a}(t)$ must diverge to infinity or minus infinity, depending on D being odd or even.

3.3. Distribution plug-in estimation

The estimators $\tilde{a}(t)$ for $a(t)$ in (9) and $\tilde{\delta}(h)$ for $\delta(h)$ in (11) are defined by

$$\tilde{a}(t) = n\int (1 - e^{-\lambda t})/(1 - e^{-\lambda})\, d\widehat{Q}(\lambda),$$
$$\tilde{\delta}(h) = n\int \gamma(\lambda, h, s)/(1 - e^{-\lambda})\, d\widehat{Q}(\lambda) - 1.$$

Although Q is over $\lambda \in (0, \infty)$, $\widehat{Q}$ might have a support point at or close to zero. When $Q = \omega Q_\xi + (1-\omega)Q_{-\xi}$ with Q_ξ degenerate at ξ, $a(t)$ is approximately linearly increasing in t as ξ approaches zero. When Q is degenerate at λ, $\lim_{\lambda\to 0} \delta(h)/a(1) = \infty$ for $h \in (-1, 0]$. If $\widehat{Q}$ has a support point near zero, then $\tilde{a}(t)$ for large t or $\tilde{\delta}(h)$ for $h \le 0$ can be unreliable. The quality of the distribution plug-in estimators can be improved if extra information about Q is available, e.g., $\lambda \in [\epsilon, \infty)$ for some $\epsilon > 0$, or one uses a penalized nonparametric MLE for Q (e.g., the minimum AIC estimator).

3.4. Confidence intervals

A bootstrap method can be used to construct confidence intervals following Mao et al. (2005). The variance of $\check{A}(m)$ can be calculated (Good and Toulmin (1956)), and with the normality assumption, confidence intervals can be constructed for $A(m)$ and $\Delta(h)$. The variance of the estimator $\check{a}(t)$ is

$$\varrho^2(t) = \sum_{x=1}^{\infty} \{1-(1-t)^x\}^2 c g_\Theta(x) - c^{-1}a^2(t) < \sum_{x=1}^{\infty} \{1-(1-t)^x\}^2 c g_\Theta(x).$$

Asymptotically, $\check{a}(t)$ is a normal random variable with mean $a(t)$ and variance $\varrho^2(t)$. Constructing asymptotic confidence intervals for $a(t)$ requires an estimator for c. A conservative choice uses the upper bound to $\varrho^2(t)$.

4. Numerical studies

4.1. Examples

Table 1 presents two datasets: `bird` concerning $n = 72$ bird species among $s = 645$ birds (Norris and Pollock (1998)), and `bivalve` concerning $n = 102$ bivalve families among $s = 748$ species (Siegel and German (1982)). The estimates $\hat{\pi}_{SC}(x)$ and $\hat{\pi}_{EB}(x)$ are shown in Figure 1. The estimates $\widehat{\Delta}_{SC}(-1)+1$, $\widehat{\Delta}_{EB}(-1)+1$, $\hat{\delta}_{EB}(-1)+1$ and $\tilde{\delta}(-1)+1$ are given, respectively, by 81.21, 75.96, 75.98, 77.24 in `bird` and 120.16, 114.75, 114.77, 123.17 in `bivalve`.

Table 1. The nonzero known n_x in two examples.

`bird`	x	1	2	3	4	5	6	7	8	9	10	12	13
	n_x	11	12	10	6	2	5	1	3	2	4	1	1
	x	14	15	16	18	25	29	30	32	39	44	53	54
	n_x	1	2	1	2	1	1	1	1	1	1	1	1
`bivalve`	x	1	2	3	4	5	6	7	8	9	12	13	14
	n_x	24	16	9	9	6	6	6	5	2	1	4	2
	x	15	16	17	20	22	29	35	55	99			
	n_x	1	1	3	1	1	2	1	1	1			

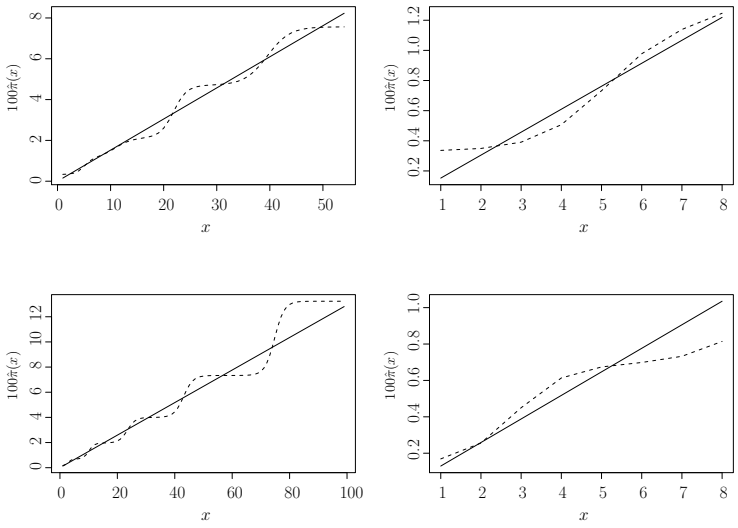

Figure 1. The estimates $\hat{\pi}_{SC}(x)$ (solid) and $\hat{\pi}_{EB}(x)$ (dashed) in `bird` (top) and `bivalve` (bottom). The right panels are close-ups of the left ones.

Let $N(m) = A(m+1) - A(m)$ with $\check{N}(m) = \check{A}(m+1) - \check{A}(m)$. Although $N(m) > 0$ and $N(m+1)/N(m) < 1$, and $2s = 1,290$ in `bird` and $2s = 1,496$ in `bivalve`, $\check{N}(1,041) < 0$ in `bird` and $\check{N}(1,322)/\check{N}(1,321) > 1$ in `bivalve`. Let $\epsilon(m)$ be the relative difference between $\tilde{a}(m/s)$ and $\widehat{A}_{EB}(m)$, $\hat{a}_{EB}(m/s)$, $\check{a}(m/s)$ and $\check{A}(m)$ (e.g., $\epsilon(m) = \check{A}(m)/\tilde{a}(m/s) - 1$), which are given by, respectively,

$$100 \times \max\{|\epsilon(m)| : m \in [1, 1041]\} = 2.04, 1.94, 0.14, 1.82 \qquad (\texttt{bird}),$$
$$100 \times \max\{|\epsilon(m)| : m \in [1, 1322]\} = 3.39, 3.38, 1.28, 1.78 \qquad (\texttt{bivalve}).$$

Let $\eta(h)$ be the relative difference between $\tilde{\delta}(h)$ and $\widehat{\Delta}_{SC}(h)$, $\widehat{\Delta}_{EB}(h)$ or $\hat{\delta}_{EB}(h)$ (e.g., $\eta(h) = \hat{\delta}_{EB}(h)/\tilde{\delta}(h) - 1$). Figure 2 presents $\tilde{\delta}(h)$ and $\eta(h)$.

The 95% asymptotic confidence intervals from $\check{a}(m/s)$ and 95% bootstrap confidence intervals for $A(m)$ from $\tilde{a}(m/s)$ are shown in Figure 3, where c is replaced with $\hat{c} = n + n_1^2/(2n_2)$ (Chao (1984)). Figure 4 presents six estimates with 95% bootstrap confidence intervals for $\Delta(0)$. The 95% bootstrap confidence limits are the 2.5% and 97.5% quantiles of 200 estimates.

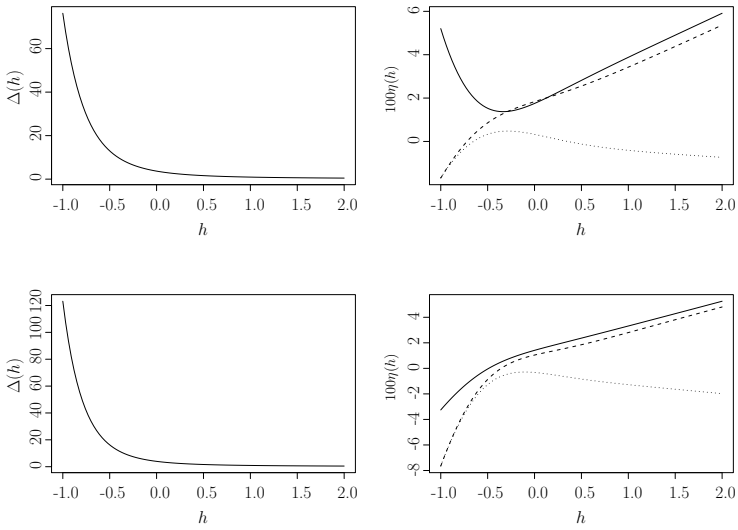

Figure 2. The estimates $\tilde{\delta}(h)$ (left) and relative differences $\eta(h) = \widehat{\Delta}_{SC}(h)/\tilde{\delta}(h) - 1$ (solid), $\widehat{\Delta}_{EB}(h)/\tilde{\delta}(h) - 1$ (dashed), $\hat{\delta}_{EB}(h)/\tilde{\delta}(h) - 1$ (dotted) (right) in `bird` (top) and `bivalve` (bottom).

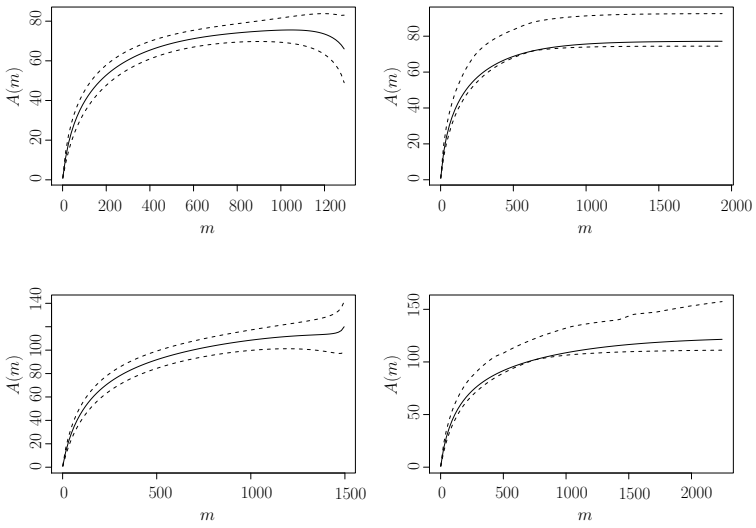

Figure 3. Confidence intervals for $A(m)$ in `bird` (top) and `bivalve` (bottom). The left panels show $\breve{a}(m/s)$ (solid) and point-wise asymptotic confidence bands (dashed) for $m \in [1, 2s]$. The right panels show $\tilde{a}(m/s)$ (solid) and point-wise bootstrap confidence bands (dashed) for $m \in [1, 3s]$.

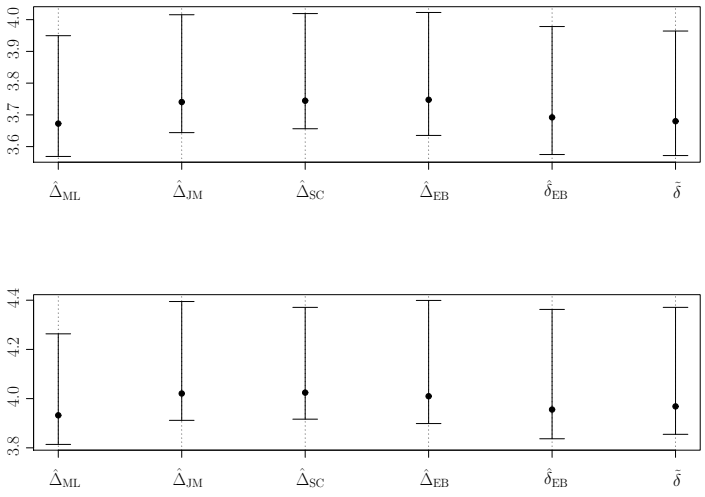

Figure 4. Six estimates (dots) and 95% bootstrap confidence intervals (segments) for $\Delta(0)$ in `bird` (top) and `bivalve` (bottom).

4.2. A simulation study

Consider three synthetic abundance models M_1, M_2 and M_3. In M_1 and M_2, $\pi_i = 1/(i\sum_{k=1}^{c} 1/k)$ for $i = 1, 2, \ldots, c$, with $c = 100$ (M_1) or $c = 200$ (M_2), and sample sizes of $s_1 = 100$ and $s_2 = 200$. In M_3, $c = 7,407$, $\pi_i = \pi^{(k)}$ with $10^5 \times \pi^{(k)} = 9.40$, 76.76, 254.09, 421.33 and 925.80, and $\sum_{i=1}^{c} I(\pi_i = \pi^{(k)}) = 7,058$, 329, 6, 12 and 2 for $k = 1, 2, 3, 4, 5$, with sample size of $s_1 = 1,000$ and $s_2 = 2,500$. Results concerning $\Delta(0)$ and $A(m)$ at $m = 2s$, $m = 2.5s$ and $m = 3s$ are reported. Given M_j and s_i, $B = 200$ samples are taken. Tables 2 and 3 present β (bias) and σ (root mean square error) for estimators of $A(m)$ and $\Delta(0)$, respectively, where for a parameter ϕ and its estimates $\hat{\phi}_k$,

$$\beta = B^{-1}\sum_{k=1}^{B}(\hat{\phi}_k - \phi), \quad \sigma = \Big\{B^{-1}\sum_{k=1}^{B}(\hat{\phi}_k - \phi)^2\Big\}^{\frac{1}{2}}.$$

The estimator $\tilde{a}(m/s)$ has the smallest $|\beta|$ and is better than, or at least comparable to, other estimators for $A(m)$ in terms of σ. While $\tilde{\delta}(0)$, $\widehat{\Delta}_{ML}(0)$ and $\widehat{\Delta}_{JM}(0)$ might have a substantial bias, $\widehat{\Delta}_{SC}(0)$, $\widehat{\Delta}_{EB}(0)$ and $\hat{\delta}_{EB}(0)$ have comparable performance. Bias is dominant for $\widehat{\Delta}_{ML}(0)$ and $\widehat{\Delta}_{JM}(0)$ in M_3. Note that $\widehat{\Delta}_{SC}(0)$ is the best in M_1 and M_2 and $\widehat{\Delta}_{EB}(0)$ is the best in M_3.

Table 2. The bias β and root mean square error σ of $\tilde{a}(m/s)$, $\hat{a}_{EB}(m/s)$ and $\widehat{A}_{EB}(m)$ for $A(m)$ in three synthetic models for $m = 2s$, $m = 2.5s$ and $m = 3s$.

			$m = 2s$		$m = 2.5s$		$m = 3s$	
			β	σ	β	σ	β	σ
M_1	s_1	$\tilde{a}$	0.3	7.7	0.9	10.3	1.8	13.2
		$\hat{a}_{EB}$	−2.3	7.6	−3.8	9.8	−5.3	12.0
		$\widehat{A}_{EB}$	−2.4	7.6	−3.8	9.8	−5.3	12.0
	s_2	$\tilde{a}$	0.3	8.1	1.3	10.8	2.5	13.7
		$\hat{a}_{EB}$	−3.4	8.1	−5.0	10.1	−6.3	11.9
		$\widehat{A}_{EB}$	−3.5	8.1	−5.0	10.1	−6.4	11.9
M_2	s_1	$\tilde{a}$	0.5	8.8	0.9	11.4	1.5	14.4
		$\hat{a}_{EB}$	−2.6	9.1	−4.7	11.9	−7.2	14.9
		$\widehat{A}_{EB}$	−2.6	9.1	−4.8	11.9	−7.3	15.0
	s_2	$\tilde{a}$	1.2	9.9	2.0	13.2	3.2	17.1
		$\hat{a}_{EB}$	−3.9	10.0	−7.1	13.6	−10.5	17.3
		$\widehat{A}_{EB}$	−4.0	10.1	−7.2	13.6	−10.6	17.4
M_3	s_1	$\tilde{a}$	−4.1	33.3	−7.4	48.5	−11.6	67.6
		$\hat{a}_{EB}$	−29.7	44.7	−58.7	75.0	−97.1	114.8
		$\widehat{A}_{EB}$	−29.7	44.7	−58.8	75.0	−97.2	114.8
	s_2	$\tilde{a}$	1.6	52.5	9.6	77.2	24.1	111.5
		$\hat{a}_{EB}$	−96.3	109.1	−177.9	190.8	−273.3	286.9
		$\widehat{A}_{EB}$	−96.4	109.1	−178.0	190.9	−273.4	287.0

Table 3. The bias β and root mean square error σ of $\widehat{\Delta}_{ML}(0)$, $\widehat{\Delta}_{JM}(0)$, $\widehat{\Delta}_{SC}(0)$, $\widehat{\Delta}_{EB}(0)$, $\widehat{\delta}_{EB}(0)$, and $\tilde{\delta}(0)$ for $\Delta(0)$ in three synthetic models.

			$\widehat{\Delta}_{ML}(0)$	$\widehat{\Delta}_{JM}(0)$	$\widehat{\Delta}_{SC}(0)$	$\widehat{\Delta}_{EB}(0)$	$\hat{\delta}_{EB}(0)$	$\tilde{\delta}(0)$
M_1	s_1	β	−0.462	−0.127	−0.063	−0.076	−0.144	0.412
		σ	0.484	0.216	0.181	0.207	0.245	1.048
	s_2	β	−0.274	−0.047	0.010	−0.060	−0.148	0.205
		σ	0.292	0.125	0.113	0.133	0.200	0.569
M_2	s_1	β	−0.702	−0.279	−0.194	−0.141	−0.201	0.674
		σ	0.718	0.339	0.278	0.261	0.303	1.432
	s_2	β	−0.436	−0.117	−0.045	−0.074	−0.135	0.442
		σ	0.450	0.172	0.128	0.153	0.193	0.904
M_3	s_1	β	−1.806	−1.021	−0.553	−0.377	−0.387	0.846
		σ	1.806	1.022	0.560	0.389	0.398	1.843
	s_2	β	−1.146	−0.497	−0.372	−0.226	−0.241	0.400
		σ	1.146	0.497	0.374	0.230	0.246	1.032

5. Discussion

Several methods can provide useful estimators for $A(m)$ with $m \leq s$ and

$a(t)$ with $t \le 1$. Because it is easy to calculate $\check{a}(t)$ and its variance, $\check{a}(t)$ is recommended. For small or moderate t (e.g., $t \in [0,3]$), $\tilde{a}(t)$ is a useful estimator for $a(t)$. While $\widehat{\Delta}_{SC}(0)$ (Chao and Shen (2003)) is a useful estimator for $\Delta(0)$, $\widehat{\Delta}_{EB}(0)$ and $\hat{\delta}_{EB}(0)$ are new competitors. In addition, the number of species, the Shannon index, and the Simpson index can be derived from the inter-specific encounter theory (Hurlbert (1971)), which also yields $1+\sum_{i=1}^{c} \pi_i^2 \log(\pi_i)/(1-\pi_i)$ (Patil and Taillie (1982)). Good (1953) considered $D(u,v) = \sum_{i=1}^{c} \pi_i^u (-\log \pi_i)^v$ for $u, v = 0, 1, \ldots$, whose domain is extended by Baczkowski, Joanes and Shamia (1998). The proposed methods can be applied to these indices.

Appendix. Proofs of Propositions

To show Proposition 1, write $A(m) = \sum_{i=1}^{c} \{1-(1-\pi_i)^m\}$ in (2) as

$$\begin{aligned}\sum_{i=1}^{c}\Big\{1-\Big(1+\frac{\pi_i}{1-\pi_i}\Big)^{-m}\Big\} &= \int \{(1+\nu)^s-(1+\nu)^{s-m}\}c(1+\nu)^{-s}dG(\nu)\\ &= \int \sum_{x=0}^{\infty}\left\{\binom{s}{x}-\binom{s-m}{x}\right\}\nu^x c(1+\nu)^{-s}dG(\nu)\\ &= \sum_{x=1}^{\infty}\left\{\binom{s}{x}-\binom{s-m}{x}\right\}\mu_x.\end{aligned}$$

For $h \notin \{0,-1\}$, as $\sum_{m=0}^{\infty}(-1)^m\binom{h+1}{m} = 0$, write $1-h\Delta(h) = \sum_{i=1}^{c}\pi_i^{h+1}$ as

$$\begin{aligned}\sum_{i=1}^{c}\{1-(1-\pi_i)\}^{h+1} &= \sum_{i=1}^{c}\sum_{m=0}^{\infty}(-1)^m\binom{h+1}{m}(1-\pi_i)^m\\ &= c\sum_{m=0}^{\infty}(-1)^m\binom{h+1}{m}-\sum_{m=0}^{\infty}(-1)^m\binom{h+1}{m}\sum_{i=1}^{c}\{1-(1-\pi_i)^m\}\\ &= 1-h\Delta(h) = -\sum_{m=0}^{\infty}(-1)^m\binom{h+1}{m}A(m)\\ &= h+1-\sum_{m=2}^{\infty}(-1)^m\binom{h+1}{m}A(m).\end{aligned}$$

For $h=0$, as $\lim_{h\downarrow 0} h^{-1}\binom{h+1}{m} = (-1)^{m-2}/\{m(m-1)\}$, write

$$\begin{aligned}\Delta(0)+1 &= \lim_{h\downarrow 0}\Delta(h)+1 = \lim_{h\downarrow 0}\sum_{m=2}^{\infty}(-1)^m h^{-1}\binom{h+1}{m}A(m)\\ &= \sum_{m=2}^{\infty}\frac{A(m)}{m(m-1)}.\end{aligned}$$

For Proposition 2, it is clear that the conclusions that concern $a(t)$ hold. To show (11) for $h \neq 0$, write $h(\delta(h)+1)/a(1)$ as

$$h\frac{\delta(h)+1}{a(1)} = \sum_{m=2}^{\infty}(-1)^m\binom{h+1}{m}\int(1-e^{-\lambda m/\theta})(1-e^{-\lambda})^{-1}\,dQ(\lambda)$$
$$= \int \frac{\sum_{m=0}^{\infty}\binom{h+1}{m}(-1)^m - \sum_{m=0}^{\infty}\binom{h+1}{m}(-e^{-\frac{\lambda}{\theta}})^m - \binom{h+1}{1}(-1) + \binom{h+1}{1}(-e^{-\frac{\lambda}{\theta}})}{1-e^{-\lambda}}\,dQ(\lambda)$$
$$= \int \frac{(h+1)(1-e^{-\frac{\lambda}{\theta}}) - (1-e^{-\frac{\lambda}{\theta}})^{h+1}}{1-e^{-\lambda}}dQ(\lambda).$$

The case with $h = 0$ in (11) holds by letting h go to zero, and (10) holds due to the relationship between Q and Θ, and $\Theta(\lambda) = c^{-1}\sum_{i=1}^{c} I(\lambda_i \leq \lambda)$.

Acknowledgements

The author thanks the Editor, an associate editor and the referee for their helpful comments that improved this article.

References

Baczkowski, A. J., Joanes, D. N. and Shamia, G. M. (1998). Range of validity of α and β for a generalized diversity index $H(\alpha, \beta)$ due to Good. *Math. Biosci.* **148**, 115-128.

Boneh, S., Boneh, A. and Caron, R. J. (1998). Estimating the prediction function and the number of unseen species in sampling with replacement. *J. Amer. Statist. Assoc.* **93**, 372-379.

Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *J. Amer. Statist. Assoc.* **88**, 364-373.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scand. J. Statist.* **11**, 265-270.

Chao, A. and Shen, T. J. (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environ. Ecol. Stat.* **10**, 429-443.

Colwell, R. K. and Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philos. Trans. R. Soc.: Biolog. Sci.* **345**, 101-118.

de Caprariis, P., Lindemann, R. H. and Collins, C. M. (1976). A method for determining optimal sample size in species diversity studies. *J. Math. Geology* **8**, 575-581.

Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: how many words did Shakespeare know? *Biometrika* **63**, 435-447.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237–264.

Good, I. J. and Toulmin, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43**, 45-63.

Hurlbert, S. H. (1971). The non-concept of species diversity: a critique and alternative parameters. *Ecology* **52**, 577–586.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. 2nd edition. Springer, New York.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: a general theory. *Ann. Statist.* **11**, 86-94.

MacArthur, R. H. (1965). Patterns of species diversity. *Biological Reviews of the Cambridge Philosophical Society* **40**, 510-533.

Magurran, A. E. (1988). *Ecological Diversity and its Measurement*. Princeton University Press.

Mao, C. X. (2004). Predicting the conditional probability of discovering a new class. *J. Amer. Statist. Assoc.* **99**, 1108-1118.

Mao, C. X., Colwell, R. K. and Chang, J. (2005). Estimating species accumulation curves using mixture. *Biometrics* **61**, 433-441.

Norris, J. L. I. and Pollock, K. H. (1998). Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species. *Envir. Ecol. Statist.* **5**, 391-402.

Patil, G. P. and Taillie, C. (1982). Diversity as a concept and its measurement. *J. Amer. Statist. Assoc.* **77**, 548-561.

Renyi, A. (1961). On measure of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Volume I, 547-561.

Shen, T. J., Chao, A. and Lin, C. F. (2003). Predicting the number of new species in taxonomic sampling. *Ecology* **84**, 798–804.

Siegel, A. F. and German, R. Z. (1982). Rarefaction and taxonomic diversity. *Biometrics* **38**, 235-241.

Smith, W. and Grassle, J. F. (1977). Sampling properties of a family of diversity measures. *Biometrics* **33**, 283-292.

Soberón, M. J. and Llorente, B. J. (1993). The use of species accumulation functions for the prediction of species richness. *Conservation Biology* **7**, 480-488.

Solow, A. R. and Polasky, S. (1999). A quick estimator for taxonomic surveys. *Ecology* **80**, 2799-2803.

Zahl, S. (1977). Jackknifing an index of diversity. *Ecology* **58**, 907-913.

Department of Statistics, University of California, Riverside, CA 92521, U.S.A.

E-mail: cmao@stat.ucr.edu

(Received September 2004; accepted September 2005)