Notes

Ecology, 84(3), 2003, pp. 798–804

PREDICTING THE NUMBER OF NEW SPECIES IN FURTHER TAXONOMIC SAMPLING

TSUNG-JEN SHEN, ANNE CHAO, AND CHIH-FENG LIN

Institute of Statistics, National Tsing Hua University, Hsin-Chu, 30043 Taiwan

Abstract. In evaluating the effectiveness of further sampling in species taxonomic surveys, a practical and important problem is predicting the number of new species that would be observed in a second survey, based on data from an initial survey. This problem can also be approached by estimating the corresponding expected number of new species. A. R. Solow and S. Polasky recently proposed a predictor (or estimator), with the form of a sum of many terms, that was derived under the assumption that all unobserved species in the initial sample have equal relative abundances. We show in this paper that the summation can be expressed as only one term. We provide a direct justification for the simplified estimator and connect it to an extrapolation formula based on a special type of species accumulation curve. Using the proposed justification, we show that, for large sample sizes, the estimator is also valid under an alternative condition, i.e., species that are represented the same number of times in the initial sample have equal relative abundances in the community. This condition is statistically justified from a Bayesian approach, although the estimator exhibits moderate negative bias for predicting larger samples in highly heterogeneous communities. In such situations, we recommend the use of a modified estimator that incorporates a measure of heterogeneity among species abundances. An example using field data from the extant rare vascular plant species patterns in the southern Appalachians is presented to compare the various methods.

Key words: *prediction; sample coverage; species survey.*

INTRODUCTION

Accurate assessment of species richness is essential for the effective and timely monitoring and management of biological communities. However, complete species inventories require extraordinary efforts and are an almost unattainable goal in practical applications. There are undiscovered species in almost every taxonomic survey or species inventory and the estimation of species richness when not all species are discovered in samples has been extensively discussed in ecology and conservation biology; see Bunge and Fitzpatrick (1993) or Colwell and Coddington (1994) for a review and historical development. New research work on this topic continues to appear.

In this paper, we consider a species taxonomic survey by selecting individuals independently and identifying the species of each selected individual. Suppose an initial survey has been conducted. Assessment of the effectiveness of further sampling relies on the prediction of the number of new species that will be seen in an additional survey of any given size. This topic of prediction can also be approached by estimating the corresponding expected number of new species. Relevant predictors or estimators have been proposed in the literature and can be grouped into two principal approaches. The first approach, which can be traced back to Arrhenius (1921, 1923), consists of fitting and extrapolating a parametric function to the species accumulation curve that plots the number of discovered species as a function of sample size. Various parametric forms (Soberon and Llorente 1993) have been proposed to fit accumulation curves based mainly on empirical experiences in species inventories. The other approach is based on a statistical sampling model, and was first formulated by Fisher et al. (1943) and Good and Toulmin (1956). Subsequent important work includes Efron and Thisted (1976) and Boneh et al. (1998). The researchers in these two approaches have developed models and methodologies along separate lines. Keating et al. (1998) compared 11 predictors (or estimators) including eight extrapolation methods and three sampling models and used simulations to develop some recommendations about the choice of predictors.

Manuscript received 13 November 2001; revised 11 July 2002; accepted 21 July 2002; final version received 12 August 2002. Corresponding Editor: A. R. Solow.

Recently, Solow and Polasky (1999) proposed a quick predictor, with the form of a sum of many terms. We show in this paper that the summation can be expressed as only one term. Our preliminary simulation studies have shown that their predictor is generally preferable to those of others and we focus on their approach. We also provide a direct and intuitive justification for their approach and we connect their prediction function to an extrapolation method based on a special type of species accumulation curve. Thus the approach integrates the extrapolation technique and the statistical sampling model.

Solow and Polasky (1999) assumed that all unobserved species in the initial sample have equal relative abundances. Because this assumption may be restrictive in practice, we show that the estimator is also valid under the alternative condition—species that are represented the same number of times in the initial sample have equal relative abundances in the community. This condition is generally satisfied from a Bayesian point of view (Good 1953). The estimator performs well with simulated data from a variety of abundance models if the prediction sample is no larger than the original one. However, it exhibits moderate negative bias especially when predicting a larger sample for highly heterogeneous communities. In such situations, we propose an improved estimator to incorporate a measure of heterogeneity.

The next section discusses our models and predictors. The third section presents an application to an extant rare vascular plant species in the southern Appalachians (Miller and Wiegert 1989). In the same section, results based on resampling from the plant community follow to compare the various methods. We close by providing some relevant discussion.

Models and Predictors

Assume that there are S species in a community and that they are labeled 1 to S. Denote the probabilities of species discovery (or relative abundance) by (p_1, p_2, . . . , p_S) with the sum of these probabilities being equal to 1. In the initial survey, assume that n individuals are selected and classified to species identity. Let f_k be the number of species with k individuals in the sample, $k = 1, 2, \ldots, n$. Let S_1 denote the number of species discovered in the initial sample. Thus, we have $f_1 + f_2 + \ldots + f_n = S_1$ and $f_1 + 2f_2 + \ldots + nf_n = n$. Let w be the number of unobserved species. It is clear that $w = S - S_1$. The goal is to predict S_2, the number of new species that will be discovered in a second survey of size m, given the information of the initial sample. This is equivalent to predicting how many of those w unobserved species would be discovered in the second survey. This prediction can also be approached by estimating $E(S_2)$, the expected number of new species as will be derived in Eq. 3.

Under the assumption of equal abundances for the unseen species, Solow and Polasky (1999) derived an estimator based on a statistical sampling theory. Their estimator is expressed as a sum of m terms as given below:

$$\hat{S}_2 = \hat{w} \sum_{k=0}^{m} \left[1 - \left(1 - \frac{1}{\hat{w}}\right)^k\right]\binom{m}{k}(1 - \hat{C})^k \hat{C}^{m-k}$$

where $\hat{w} = f_1^2/(2f_2)$ is a ''plug-in'' estimator for the number of unobserved species w and $\hat{C} = 1 - f_1/n$ is a ''plug-in'' estimator of the sample coverage C (defined in the next paragraph). Note that the above estimator can be rewritten as the difference of two summations:

$$\hat{S}_2 = \hat{w} \sum_{k=0}^{m} \binom{m}{k}(1 - \hat{C})^k \hat{C}^{m-k} - \hat{w} \sum_{k=0}^{m} \binom{m}{k}\left(1 - \hat{C} - \frac{1 - \hat{C}}{\hat{w}}\right)^k \hat{C}^{m-k}.$$

Based on the binomial expansion,

$$(a + b)^m = \sum_{k=0}^{m} \binom{m}{k} a^k b^{m-k}$$

the first summation becomes

$$[(1 - \hat{C}) + \hat{C}]^m = 1$$

and the second summation becomes

$$[1 - \hat{C} - (1 - \hat{C})/\hat{w} + \hat{C}]^m = [1 - (1 - \hat{C})/\hat{w}]^m.$$

Then the estimator is simplified to

$$\hat{S}_2 = \hat{w}\left[1 - \left(1 - \frac{1 - \hat{C}}{\hat{w}}\right)^m\right]. \quad (1)$$

From Eq. 1, it is clear that the estimation of the number of unseen species is involved in the prediction problem. The number of unseen species represents the asymptotic value of S_2 as m tends to infinity.

The coverage of a sample is defined as the total relative abundances of the species discovered in the sample. Given that there are w unseen species, we assume, without any loss of generality, that the w unseen species are indexed from 1 to w in the community and their species relative abundances are (p_1, p_2, . . . , p_w) and thus the observed species correspond to the abundances (p_{w+1}, p_{w+2}, . . . , p_S). The sample coverage is defined as $C = p_{w+1} + p_{w+2} + \ldots + p_S$. In other words, the sample coverage denotes the conditional (on data) probability of finding a species that has already been discovered in the sample if an additional observation were to be taken. A well-known estimator for the sample coverage proposed by A. M. Turing (Good 1953)

is $\hat{C} = 1 - f_1/n = (2f_2 + 3f_3 + \ldots + nf_n)/n$, which is the fraction of repeated species in the sample because $(2f_2 + 3f_3 + \ldots + nf_n)$ represents the number of individuals for species discovered at least twice. Turing's estimator generally performs well in many situations; for example, see Esty (1986).

On the other hand, we can also say that $1 - C = p_1 + p_2 + \ldots + p_w$, the conditional probability of discovering a new species in an additional observation. Turing's estimator implies that an estimator for this conditional probability is the proportion of singletons in the sample. This can be intuitively understood because a new species must be a singleton in the enlarged sample that includes the additional individual.

The use of the estimator (Eq. 1) is particularly appealing because of its simplicity and the following direct justification under a special assumption that all unobserved species have identical abundances in the community. Conditional on the initial survey in that there are w species undiscovered with relative abundances $(p_1, p_2, \ldots, p_w)$, the above assumption implies that $p_1 = p_2 = \ldots = p_w = (1 - C)/w$ because $1 - C = p_1 + p_2 + \ldots + p_w$. For each of these unobserved species, the probability that this species will be again missed in the m observations is $(1 - [1 - C]/w)^m$. Therefore, the expected number of new species in the second survey is $w(1 - [1 - \{1 - C\}/w]^m)$. Plugging in estimators for w and C, we then obtain the predictor given in Eq. 1 and justify the approach of Solow and Polasky (1999).

As m and w are large enough, we have the following approximation:

$$w(1 - [1 - \{1 - C\}/w]^m) \approx w\{1 - \exp[-(1 - C)m/w]\}. \quad (2)$$

Soberon and Llorente (1993) suggested two forms of the species accumulation function. One of them is the exponential model, $w\{1 - \exp(-\alpha m/w)\}$, where α denotes a parameter. Miller and Wiegert (1989) also fitted this model to different types of data sets and obtained satisfactory results. Comparing this exponential model with Eq. 2, we see that Eq. 1 can also be regarded as an exponential extrapolation formula from the accumulation curve of the second sample. The asymptotic value of the accumulation curve when the sample size of the second sample tends to be large is the number of unobserved species in the first sample. Therefore, our justification also connects the prediction function in Eq. 1 to the well-known exponential species accumulation curve approach.

Notice that for each of the unobserved species with abundance probability p, the probability that this species will be missed in the second sample of size m is $(1 - p)^m$. Therefore, we have the conditional expected value of S_2 given the initial data as

$$E(S_2 \mid \text{data}) = \sum_{i=1}^{S} [1 - (1 - p_i)^m]\delta_i = w - \sum_{i=1}^{w} (1 - p_i)^m$$

where $\delta_i = 1$ if the ith species is undiscovered in the initial survey of size n, and $\delta_i = 0$ otherwise. (The introduction of this notation, δ_i, in the above is simply to show that only unseen species contribute to the sum.) If we average out all possible initial data information, then the expected number of new species is

$$E(S_2) \sum_{i=1}^{S} [1 - (1 - p_i)^m](1 - p_i)^n \quad (3)$$

because $E(\delta_i) = (1 - p_i)^n$. Expanding $(1 - p_i)^m$ in the above formula and letting the sample sizes m and n increase such that $m/n \to t$, $0 < t < 1$, we have

$$E(S_2) \approx \sum_{k=1}^{\infty} (-1)^{k+1} t^k E(f_k). \quad (4)$$

The estimator proposed by Efron and Thisted (1976), $\Sigma_{k=1}^{\infty} (-1)^{k+1} t^k f_k$, was obtained from the above formula by replacing $E(f_k)$ by the observed frequency f_k. However, this estimator lacks some theoretical properties of the prediction function (Boneh et al. 1998) and may take negative values or become extremely large, as will be shown in the next section.

We now explain that the predictor (Eq. 1) is also valid under the condition that the species that are represented the same number of times in the initial sample have equal relative abundances in the community. The validity is based on

$$E(S_2) \approx E(w)\left[1 - \left(1 - \frac{E(1 - \hat{C})}{E(w)}\right)^m\right]. \quad (5)$$

Substituting $\hat{C}$ and using a binomial expansion, the right hand side of the above as m and n increase such that $m/n \to t$, $0 < t < 1$, reduces to

$$\sum_{k=1}^{\infty} (-1)^{k+1} t^k \frac{[E(f_1)]^k}{k![E(w)]^{k-1}}. \quad (6)$$

Based on Eqs. 4 and 6, the proof of Eq. 5 is equivalent to showing that for any integer $k \geq 2$, if all species that are represented j times in the initial sample have equal abundances in the community for each $j < k$, then $E(f_k) \approx [E(f_1)]^k/\{k![E(w)]^{k-1}\}$. That is,

$$[E(w)]^{k-1} \approx [E(f_1)]^k/[k!E(f_k)]. \quad (7)$$

The proof for $k = 2$ is briefly sketched here. Note that

$$E(w) = \sum_{i=1}^{S} (1 - p_i)^n = \sum_{i=1}^{S} \frac{1 - p_i}{np_i}[np_i(1 - p_i)^{n-1}]$$

$$= E \sum_{i=1}^{S} \frac{1 - p_i}{np_i}\Delta_i$$

where $\Delta_i = 1$ if the ith species is a singleton in the initial sample, and $\Delta_i = 0$ otherwise. Hence, $\Delta_1 + \Delta_2 + \ldots + \Delta_S = f_1$. Only the abundances of singletons contribute to the sum $\Sigma_{i=1}^{S} (1 - p_i)\Delta_i/(np_i)$. Under the assumption that all singletons have identical relative abundances, denote this common abundance value by h_1 and let $\theta = h_1/(1 - h_1)$. It follows from the above equation that $E(w) \approx E[f_1/(n\theta)]$. An approximation of θ can be obtained from the following:

$$E(f_1\theta) = E \sum_{i=1}^{S} \frac{p_i}{1 - p_i}\Delta_i$$

$$= \sum_{i=1}^{S} np_i^2(1 - p_i)^{n-2} \approx E\left(\frac{2f_2}{n}\right)$$

which implies $\theta \approx 2E(f_2)/[nE(f_1)]$. Consequently, Eq. 7 is valid for $k = 2$. The proof for a general k (by a mathematical induction) is omitted here.

Notice that for $k = 2$, Eq. 7 implies that if all singletons (species appearing only once in sample) have identical relative abundances, then $E(w) \approx [E(f_1)]^2/[2E(f_2)]$, which subsequently yields the ''plug-in'' estimator $\hat{w} = f_1^2/(2f_2)$ in Solow and Polasky's procedure. This estimator is originally proposed by Chao (1984) as a lower bound of the number of unseen species. Eq. 7 shows that the estimator is also valid for extreme heterogeneous cases as long as the singletons in the sample have the same relative abundances.

Good (1953) and Engen (1978:31–32) concluded that the (posterior) relative abundance of a species, given it was observed j times in the sample, is approximately $(j + 1) f_{j+1}/(nf_j)$. For example, all singletons have about the same abundances and can be estimated by $2f_2/(nf_1)$. Therefore, the assumption for the equal abundance of species represented the same number of times is satisfied from a Bayesian point of view.

It follows from Eq. 5 that a general class of predictors of S_2 has the form of Eq. 1, but $\hat{w}$ could be any estimator of the unobserved species. In the next section, we show that the estimator (Eq. 1) performs satisfactorily by sampling from a real heterogeneous species-abundance distribution, but it has modest negative bias when $m > n$. A simple way to remove the bias is to replace the estimator $\hat{w}$ by an estimator which incorporates a measure of the heterogeneity of species abundances. We propose the use of an estimator derived in Chao and Lee (1992) and Chao et al. (2000). In their approach, abundant and rare species are treated separately. Abundant species are those having more than k individuals in the sample; the observed rare species are those represented by only one, two, . . . , and up to k individuals in the sample. The estimation of the number of missing species is based entirely on the observed rare species because abundant species would be discovered in any sample anyway and thus they do not contain any information about the missing species. Let the total number of rare species in the sample be $S_{\text{rare}} = \Sigma_{i=1}^{k} f_i$. Then the estimator of the unobserved species based on the estimated sample coverage $\tilde{C} = 1 - f_1/\Sigma_{i=1}^{k} if_i$ is the following (Chao et al. 2000, Section 2):

$$\tilde{w} = \frac{S_{\text{rare}}}{\tilde{C}} + \frac{f_1}{\tilde{C}}\hat{\gamma}^2 - S_{\text{rare}} \tag{8}$$

where

$$\hat{\gamma}^2 = \max\left\{\frac{S_{\text{rare}}}{\tilde{C}} \frac{\sum_{i=1}^{k} i(i - 1)f_i}{\left(\sum_{i=1}^{k} if_i\right)\left(\sum_{i=1}^{k} if_i - 1\right)} - 1, 0\right\} \tag{9}$$

denotes the estimated squared coefficient of variation (CV) of the species abundances. The value of CV characterizes the degree of heterogeneity among the species abundances. The CV is zero if and only if the species have equal abundance. The larger the CV, the greater the degree of heterogeneity. A value of the cutoff point, $k = 10$, is adopted throughout the paper based on empirical experiences (Chao et al. 1993). Our modified estimator is thus

$$\tilde{S}_2 = \tilde{w}\left[1 - \left(1 - \frac{1 - \hat{C}}{\tilde{w}}\right)^m\right]. \tag{10}$$

The proposed estimator is a function of the frequencies $(\tilde{w}, f_1, f_2, \ldots, f_n)$, which is approximately a multinomial distribution with an estimated number of species $\tilde{S} = S_1 + \tilde{w}$ and cell probabilities $(\tilde{w}/\tilde{S}, f_1/\tilde{S}, f_2/\tilde{S}, \ldots, f_n/\tilde{S})$. Therefore, a variance estimator of the proposed estimator $\tilde{S}_2$ can be obtained by using a standard asymptotic approach. That is, we have the following variance estimator:

$$\widehat{\text{var}}(\tilde{S}_2) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial \tilde{S}_2}{\partial f_i} \frac{\partial \tilde{S}_2}{\partial f_j} \widehat{\text{cov}}(f_i, f_j) \tag{11}$$

where $\widehat{\text{cov}}(f_i, f_j) = f_i(1 - f_i/\tilde{S})$ for $i = j$ and $\widehat{\text{cov}}(f_i, f_j) = -f_i f_j/\tilde{S}$ for $i \neq j$. The performance of this variance estimator is investigated in the next section.

Real Data Example

Miller and Wiegert (1989) documented the species-abundance distribution of endangered and rare vascular plant species in the central portion of the southern Appalachian region. A total of $S_1 = 188$ species were recorded out of $n = 1008$ individuals compiled over a

TABLE 1. Frequency counts of the extant rare plant species (S_1 = 188 out of 1008 individuals counted) in the southern Appalachians (Miller and Wiegert 1989).

No. individuals, i	No. species, f_i	No. individuals, i	No. species, f_i
1	61	14	2
2	35	15	3
3	18	16	2
4	12	19	1
5	15	20	2
6	4	22	1
7	8	29	1
8	4	32	1
9	5	40	1
10	5	43	1
11	1	48	1
12	2	67	1
13	1		

150-yr period. The species-abundance distribution is reproduced in Table 1.

Assume that the data in Table 1 can be regarded as a multinomial sample from the community and we are interested in estimating the number of new species expected in the next m additional individuals. The following four estimators are shown in Table 2 for m = 500, 1000, 2000, 3000, and 4000: Solow and Polasky (1999) estimator $\hat{S}_2$ (Eq. 1); the proposed estimator $\tilde{S}_2$ (Eq. 10); the estimator suggested by Boneh et al. (1998); and the estimator suggested by Efron and Thisted (1976) based on Eq. 4. The estimated standard error for each estimator was computed from an asymptotic approach as described earlier.

The method suggested by Boneh et al. in all cases produces the lowest estimate. When predicting a smaller sample (m = 500 and 1000), except for Boneh et al.'s estimator, the other three estimates are generally comparable. When predicting a larger sample (m = 2000, 3000, and 4000), Efron and Thisted's (1976) estimate overflows. In these cases, the proposed method yields slightly higher estimate than the one suggested by Solow and Polasky (1999).

When m tends to infinity, it follows from Eq. 10 and the first 10 frequencies in Table 1 that the proposed estimates converge stably to $\tilde{w}$ = 58. This asymptotic value is needed in calculating our estimate for any finite value of m. For example, the estimate for m = 500 in Table 2 is obtained by $\tilde{S}_2 = 58[1 - (1 - 0.0605/58)^{500}] = 23.6$, where $0.0605 = 1 - \hat{C} = f_1/n = 61/1008$. An estimate of species richness thus becomes $S_1 + \tilde{w} = 188 + 58 = 246$, with an asymptotic standard error of 15.2. Similarly, Solow and Polasky's estimates converge to $\hat{w} = f_1^2/(2f_2) = 53$, which subsequently gives a species richness estimate of 241 with a standard error of 17.9. We remark that Sichel (1997) analyzed these data by fitting a parametric function (generalized inverse Gaussian distribution) to the frequency counts and obtained an estimate of 283 for the total number of species.

We also conducted a simulation experiment by selecting individuals with replacement from this species-abundance distribution. That is, we consider a community with 188 species and there are 61 species with a relative abundance of 1/1008, 35 species with an abundance of 2/1008, . . . etc. The CV value for this distribution is 1.56, which represents a measure of the degree of heterogeneity. In our experiment, 2000 simulated samples with four initial sizes (n = 100, 200, 300, and 400) were generated from the extant plant species distribution. For each value of additional sample size (m = 100, 200, 300, and 400), we calculated the true value S_2 (the number of new species discovered in the second sample of size m) and four estimates.

The resulting 2000 true values and estimates were averaged to give the final prediction results in Table 3. The mean true values are nearly identical to $E(S_2)$ given in Eq. 3. The averaged values for all estimators as well as the sample standard error for each estimator are shown in the same table. For the proposed estimator, the average of the estimated standard error obtained from Eq. 11 is also given.

Based on Table 3 and other unreported simulation results for various abundance models, we summarize the following findings:

1) The method proposed by Efron and Thisted (1976) works well when the prediction size (i.e., m) is not greater than the initial sample size n. In the prediction of a larger prospective sample, the estimate could become either negative as in the cases of (n = 100, m = 300) and (n = 300, m = 400) or extremely large as in the cases of (n = 100, m = 400) and (n = 200, m = 400). In these cases, the sample standard errors are unboundedly large. Therefore, we have a similar conclusion as in Boneh et al. (1998) that the method is useful only in the prediction of a future sample with a size not greater than the original.

2) The other three estimators are free of the erratic behavior associated with Efron and Thisted's estimator. The estimator suggested by Boneh et al. (1998) is generally biased downwards and the bias is often substan-

TABLE 2. Estimation results for the plant community.

Second sample size, m	Solow and Polasky (1999)	Proposed in Eq. 10	Boneh et al. (1998)	Efron and Thisted (1976)
500	23.1 (4.4)	23.6 (4.2)	16.1 (1.6)	23.6 (3.9)
1000	36.1 (7.7)	37.5 (6.9)	24.1 (2.2)	43.8 (12.9)
2000	47.7 (12.1)	50.7 (9.9)	30.7 (2.4)	overflow
3000	51.4 (14.3)	55.3 (11.4)	32.9 (2.4)	overflow
4000	52.6 (15.3)	57.0 (12.1)	33.6 (2.4)	overflow

Note: Values in parentheses are 1 SE.

TABLE 3. Comparison of various estimators from resamples from the plant community.

Initial size	True value	Solow and Polasky (1999)	Proposed in Eq. 10	SE from Eq. 11	Boneh et al. (1998)	Efron and Thisted (1976)
100						
$m = 100$	29.0	28.2 (5.9)	29.2 (5.5)	6.3	12.0 (1.2)	29.0 (7.3)
$m = 200$	48.0	44.2 (11.9)	46.9 (11.1)	11.6	15.6 (1.7)	28.8 (1859)
$m = 300$	61.8	53.6 (16.3)	57.6 (15.1)	15.9	16.8 (1.8)	$-22\ 075$ ($>1 \times 10^6$)
$m = 400$	72.0	58.7 (20.2)	64.0 (18.5)	19.3	17.3 (1.8)	$>1 \times 10^6$ ($>1 \times 10^6$)
200						
$m = 100$	19.0	18.9 (3.0)	19.4 (2.8)	3.8	10.6 (0.9)	19.1 (3.2)
$m = 200$	32.6	31.7 (6.2)	33.4 (5.6)	6.8	16.1 (1.4)	33.1 (8.7)
$m = 300$	43.0	39.8 (9.5)	42.7 (8.5)	9.3	18.9 (1.7)	26.5 (2018)
$m = 400$	51.5	45.3 (11.8)	49.6 (10.3)	11.4	20.6 (1.8)	$>1 \times 10^5$ ($>1 \times 10^5$)
300						
$m = 100$	13.7	13.7 (2.0)	13.9 (1.9)	2.7	8.8 (0.7)	13.8 (2.1)
$m = 200$	24.1	23.5 (4.2)	24.2 (3.8)	4.8	14.2 (1.1)	24.2 (5.0)
$m = 300$	32.3	30.6 (6.3)	32.0 (5.5)	6.6	17.7 (1.4)	32.3 (10.3)
$m = 400$	39.0	36.4 (8.2)	38.3 (7.0)	8.1	20.1 (1.6)	-133 (7220)
400						
$m = 100$	10.5	10.4 (1.5)	10.5 (1.4)	2.0	7.2 (0.5)	10.4 (1.5)
$m = 200$	18.7	18.4 (3.2)	18.8 (2.8)	3.7	12.3 (0.9)	18.7 (3.5)
$m = 300$	25.4	24.7 (4.6)	25.3 (4.0)	5.1	15.8 (1.2)	25.7 (5.8)
$m = 400$	30.9	29.3 (6.2)	30.4 (5.2)	6.3	18.4 (1.4)	31.2 (10.6)

Note: Numbers in parentheses denote the sample standard error over 2000 trials.

tial. Although their estimator has the smallest sample standard error compared with other estimators, it seems for our cases that the estimator does not increase accordingly in an anticipated rate as m is increased.

3) It is clear that the performance of all estimators deteriorates as the prediction sample size is increased. When the prediction size m is no larger than n, all methods except the one suggested by Boneh et al. (1998) yield quite satisfactory estimates. Our proposed estimator generally has the smallest sample standard error. When m is larger than n, as described earlier, Efron and Thisted's estimator becomes unstable and the method of Boneh et al. yields a value far below the true value. Solow and Polasky's approach and our modification are generally preferable. The former tends to be biased downwards for predicting larger samples, and the magnitude of the bias increases with the prediction size. Our modified estimator can reduce the bias and exhibits less variation, and thus is recommended for practical use. The estimated standard errors using an asymptotic approach, although positively biased, are generally satisfactory compared with the sample standard errors.

DISCUSSION

We have simplified and provided an intuitive justification for the Solow and Polasky (1999) approach to predicting the number of new species in further taxonomic sampling. We have also proposed a modification to the approach to adapt for use in highly heterogeneous communities. An additional advantage for Solow and Polasky's approach, in Eq. 1, is that only two frequencies are used, that is, the number of singletons and doubletons; i.e., (f_1, f_2). Therefore, the approach is simple and the implementation is direct. Our modification extends the method and uses more frequencies; see Eq. 10. That is, only the statistics $(f_1, f_2, \ldots, f_k)$ where k is usually selected as 10, are used. The frequencies for relatively common species are not needed in these methods. Thus, ecologists do not need to obtain the exact frequencies of relatively abundant species once there are enough representatives in the sample.

Keating et al. (1998) made a valuable comparison of various estimators. Although a limited simulation study in this paper has shown the advantage of Solow and Polasky's original and modified approach, extensive comparison is still needed to investigate the relative merits of the proposed and the other existing methods.

This paper deals with the sampling scheme in which individuals are selected independently with replacement from the community. If communities are sampled by randomly selected quadrats, a common approach to estimating the number of unseen species in quadrat-sampling is the capture–recapture model with presence–absence data as discussed by Boulinier et al. (1998). Whether the methods discussed in this paper can be modified to a prediction problem based on capture–recapture framework would be an interesting topic for future research.

A computer program SPADE (Species Prediction And Diversity Estimation), written in C language, that

calculates all estimates discussed in this paper and some diversity indices, will be soon available.[1]

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for helpful comments. This research is supported by the National Science Council of Taiwan.

LITERATURE CITED

Arrhenius, O. 1921. Species and area. Journal of Ecology **9**: 95–99.

Arrhenius, O. 1923. Statistical investigations in the constitution of plant associations. Ecology **4**:68–73.

Boneh, S., A. Boneh, and R. J. Caron. 1998. Estimating the prediction function and the number of unseen species in sampling with replacement. Journal of the American Statistical Association **93**:372–379.

Boulinier, T., J. D. Nichols, J. R. Sauer, J. E. Hines, and K. H. Pollock. 1998. Estimating species richness: the importance of heterogeneity in species detectability. Ecology **79**: 1018–1028.

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

Chao, A. 1984. Nonparametric estimation of the number of classes in a population. Scandinavian Journal of Statistics **11**:265–270.

Chao, A., W.-H. Hwang, Y.-C. Chen, and C.-Y. Kuo. 2000. Estimating the number of shared species in two communities. Statistica Sinica **10**:227–246.

Chao, A., and S.-M. Lee. 1992. Estimating the number of classes via sample coverage. Journal of the American Statistical Association **87**:210–217.

[1] URL: ⟨http://chao.stat.nthu.edu.tw⟩

Chao, A., M.-C. Ma, and M. C. K. Yang. 1993. Stopping rules and estimation for recapture debugging with unequal failure rates. Biometrika **80**:193–201.

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

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

Engen, S. 1978. Stochastic abundance models. Chapman and Hall, London, UK.

Esty, W. W. 1986. The efficiency of Good's nonparametric coverage estimator. Annals of Statistics **14**:1257–1260.

Fisher, R. A., A. S. Corbet, and C. B. Williams. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. Journal of Animal Ecology **12**:42–58.

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

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

Keating, K. A., J. F. Quinn, M. A. Ivie, and L. L. Ivie. 1998. Estimating the effectiveness of further sampling in species inventories. Ecological Applications **8**:1239–1249.

Miller, R. I., and R. G. Wiegert. 1989. Documenting completeness, species-area relations, and the species-abundance distribution of a regional flora. Ecology **70**:16–22.

Sichel, H. S. 1997. Modelling species-abundance frequencies and species-individual functions with the generalized inverse Gaussian-Poisson distribution. South African Statistical Journal **31**:13–37.

Soberon, M. J., and B. J. Llorente. 1993. The use of species accumulation functions for the prediction of species richness. Conservation Biology **7**:480–488.

Solow, A. R., and S. Polasky. 1999. A quick estimator for taxonomic surveys. Ecology **80**:2799–2803.