Nonparametric Prediction in Species Sampling

Anne CHAO and Tsung-Jen SHEN

Consider a continuous-time stochastic model in which species arrive in the sample according to independent Poisson processes and where the species discovery rates are heterogeneous. Based on an initial survey, we are concerned with the problem of predicting the number of new species that would be discovered by additional sampling. When the sampling time or sample size of the additional sample tends to infinity, this problem reduces to the prediction of the number of undetected species in the original sample, or equivalently, the estimation of species richness. The topic has a wide range of applications in various disciplines. We propose a simple prediction method and apply it to two datasets. One set of data deals with the capture counts of the Malayan butterfly and the other set deals with identification records of organic pollutants in a water environment. Simulation results are shown to investigate the performance of the proposed method and to compare it with the existing estimators.

Key Words: Discovery rates; Frequency counts; Species abundance; Species richness.

1. INTRODUCTION

In biological and ecological sciences, the compilation of complete species inventories often requires extraordinary effort and is an almost unattainable goal in practical applications. There are undiscovered species in almost every taxonomic survey or species inventory. Since the pioneering article by Fisher, Corbet, and Williams (1943), the estimation of species richness has been extensively discussed in the literature and applied to many disciplines. Bunge and Fitzpatrick (1993) and Colwell and Coddington (1994) provided comprehensive reviews and historical development.

We are concerned with a more general problem of predicting the number of new species that will be discovered in an additional survey given that an initial survey has already been conducted. The prediction results provide objective bases to assess the effectiveness of further surveys so that sampling efforts and funding can be allocated among sites for effective and timely management of biological communities. The results can also be used to assess survey completeness and estimate the minimum effort needed to reach a certain level

Anne Chao is Professor, and Tsung-Jen Shen is Post-Doctoral Fellow, Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan, 30043 (E-mail: chao@stat.nthu.edu.tw).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 9, Number 3, Pages 253–269
DOI: 10.1198/108571104X3262

Table 1. Frequency Counts for the Malayan Butterfly Data (Fisher et al. 1943). (f_i: the number of species represented by i individuals in the sample)

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
f_i	118	74	44	24	29	22	20	19	20	15	12	14	6	12
i	15	16	17	18	19	20	21	22	23	24	>24	D	n	
f_i	6	9	9	6	10	10	11	5	3	3	119	620	9,031	

of completeness. Therefore, the topic is especially important in species taxonomic surveys (Boulinier et al. 1998; Keating, Quinn, Ivie, and Ivie 1998) and in environmental monitoring (Janardan and Schaeffer 1981). This problem can also be approached by estimating the prediction function, that is, the corresponding expected number of new species for any further sample. When the sampling time or sample size of the prediction survey becomes infinite, the problem reduces to the prediction of the number of undetected species in the sample, or equivalently, the estimation of species richness.

Our motivations originated from analyzing the well-known Malayan butterfly data (Fisher et al. 1943) and the organic pollutant records in environmental science (Janardan and Schaeffer 1981). The former dataset includes a large collection of Malayan butterflies made by A. S. Corbet in 1942. There were 9,031 individuals classified to 620 species. Out of these species, 118 were represented by one individual, 74 by two individuals, 44 by three individuals and so on (Table 1). Detailed analysis will be given in Section 3. According to Fisher et al. (1943, p. 43) and Williams (1964, p. 20), once 24 individuals of any species had been captured, little effort was made to obtain others of that species. We were thus motivated to predict species richness based only on the exact lower order frequency counts (i.e., frequencies from rarer species). The other data compiled by Janardan and Schaeffer (1981) include 1,258 distinct organic compounds from 5,720 separate identifications in a water environment. The frequency counts (see Section 3 for details) are: 503 compounds were identified once, 238 twice, 133 three times, and so on. Janardan and Schaeffer (1981) applied several different methods to estimate how many more new compounds were expected to appear as new datasets were formed. One of our motivations was to compare the relative merits of the currently existing approaches.

Predictors or estimators based on various sampling models have been proposed. In the parametric approach, the species discovery rates are modeled as a random sample from a parametric distribution; for example, Fisher et al. (1943) adopted a gamma distribution. Other parametric models include the normal (Coull and Agresti 1999), log-normal (Bulmer 1974), inverse-Gaussian (Ord and Whitmore 1986), generalized inverse-Gaussian (Sichel 1997), and beta distribution (Lloyd and Yip 1991). Semiparametric and nonparametric approaches were discussed by Good and Toulmin (1956), Efron and Thisted (1976), Burnham and Overton (1978), Smith and van Belle (1984), Agresti (1994), Norris and Pollock (1998), Boneh, Boneh, and Caron (1998), and Solow and Polasky (1999). Keating et al. (1998) provided interesting comparisons for various approaches, including the extrapolation (curve fitting) technique and some parametric approaches.

The various approaches can be considered under a unified sampling scheme based on

the mixed Poisson processes (Bunge and Fitzpatrick 1993). In this model, species arrive in the sample according to independent Poisson processes. The commonly used multinomial model can be obtained by conditioning on the total number of individuals in a mixed Poisson sampling. In this article, we focus on the nonparametric approaches based on a continuous-time framework. See Solow and Polasky (1999) and Shen, Chao, and Lin (2003) for a discrete analogue of the model formulation.

In our model, the species discovery rates (i.e., rates of the Poisson processes) are allowed to be heterogeneous. Here the discovery rate is a combination of species abundance and individual detectability. The heterogeneity arises not only from the abundance of various species but also from many other possible factors (movement patterns, color, size, and vocalizations) that determine the chance of finding individuals.

For the general prediction problem in a continuous-time set-up, Fisher et al. (1943), Good and Toulmin (1956), and Efron and Thisted (1976) proposed several nonparametric approaches. Boneh et al. (1998) pointed out that some commonly used predictors lack certain desirable properties of the true prediction function. As will be discussed in Section 2, they indicated that three critical properties should be satisfied and that the lack of those properties would result in serious consequences. Thus, their work has established a general guideline for constructing and comparing predictors. Boneh et al. (1998) proposed a predictor that maintains those properties. However, their predictor generally exhibits bias in simulation studies. It was their comment, "finding a better bias-correction is certainly a worthy problem," that led to our work. This article presents a simple alternative predictor that satisfies the required properties of a prediction function and also improves on existing methods. The proposed predictor is derived under a weak assumption and performs satisfactorily for simulated data under a variety of models with heterogeneous discovery rates.

In Section 2, we formulate the models and present the proposed estimators. The Malayan butterfly and chemical pollution data are analyzed in Section 3 to illustrate the new method. Simulation results are reported in Section 4 to examine the performance of the proposed method. In the final section, some discussion and concluding remarks are provided.

2. MODELS AND METHODS

Assume that there are S species in a community labeled from 1 to S. In the initial sample, the community is sampled for a fixed time. Without loss of generality, assume that the initial survey corresponds to the time interval $[-1, 0]$. Individuals (or elements) of the ith species arrive in the sample according to a Poisson process with discovery rate λ_i. Let X_i be the number of times the ith species is observed in the time period $[-1, 0]$. Only those species with $X_i > 0$ are discovered in the sample. Denote the maximum frequency in the sample by r, that is, $r = \max\{X_i; 1 \le i \le S\}$. Let $n = \sum_{i=1}^{S} X_i$ be the total number of individuals found in the initial sample and $f_k, k = 0, 1, \ldots, r$, be the number of species represented by k individuals (or observed k times). Here the initial sample size n is a random

variable and f_0 denotes the number of undetected species in the initial sample. Thus, we have $n = \sum_{k=1}^{r} k f_k$ and $f_k = \sum_{i=1}^{S} I[X_i = k]$, where $I[A]$ is the usual indicator function. Let $D = \sum_{i=1}^{S} I[X_i > 0] = \sum_{k=1}^{r} f_k$ denote the number of distinct species discovered in the initial survey.

The goal is to predict $U(t)$, the number of new species that will be discovered in a second survey for time interval $[0, t]$. This is equivalent to predicting how many of those undetected species (in the initial survey) would be discovered in the second survey. The prediction can also be approached by estimating the prediction function $\theta(t) = E[U(t)]$, as will be derived below.

First, notice that for each of the undetected species with rate λ, the probability that this species will be missed in the interval $[0, t]$ is $\exp(-\lambda t)$. Therefore, we have the conditional expected value of $U(t)$ given the initial data as

$$E[U(t) \mid \text{data}] = \sum_{i=1}^{S} (1 - e^{-\lambda_i t}) I[X_i = 0] = f_0 - \sum_{i=1}^{S} e^{-\lambda_i t} I[X_i = 0]. \tag{2.1}$$

If we average out all possible initial data, then the prediction function can be expressed as

$$\theta(t) = E[E(U(t) \mid \text{data})] = \sum_{i=1}^{S} (1 - e^{-\lambda_i t}) e^{-\lambda_i}, \tag{2.2}$$

because $E\{I[X_i = 0]\} = e^{-\lambda_i}$. Expanding $1 - \exp(-\lambda_i t)$ in the above formula and using

$$E(f_k) = \sum_{i=1}^{S} \lambda_i^k \exp(-\lambda_i)/k! \tag{2.3}$$

we can write $\theta(t)$ in terms of the expected frequencies as

$$\theta(t) = \sum_{k=1}^{\infty} (-1)^{k+1} t^k E(f_k). \tag{2.4}$$

The following estimator, originally derived by Good and Toulmin (1956, p. 49) and proposed by Efron and Thisted (1976),

$$\hat{\theta}_1(t) = \sum_{k=1}^{r} (-1)^{k+1} t^k f_k, \tag{2.5}$$

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 numerical results in Sections 3 and 4. Boneh et al. (1998) suggested that any estimator $\hat{\theta}(t)$ should possess the following three properties:

1. $\hat{\theta}(0) = 0$.
2. $\hat{\theta}(t)$ has a horizontal asymptotic value as t tends to infinity.
3. For any positive integer i, $(-1)^{i+1}\hat{\theta}^{(i)}(t) > 0$ for $t \geq 0$, where $\hat{\theta}^{(i)}(t)$ denote the ith derivative of $\hat{\theta}(t)$ with respect to t.

The above properties imply that any estimator $\hat{\theta}(t)$ should be bounded, concave, and monotonically increasing with t. Boneh et al. (1998) suggested an alternative prediction function that replaces λ_i in Equation (2.2) by the observed data, resulting in the maximum likelihood estimator. They further proposed the following bias-corrected version:

$$\hat{\theta}_2(t) = \sum_{k=1}^{r} f_k e^{-k}(1 - e^{-kt}) + v\left[e^{-f_1/v} - e^{-f_1(1+t)/v}\right], \tag{2.6}$$

where v is the solution of the equation $v[1 - \exp(-f_1/v)] = \sum_{k=1}^{r} f_k \exp(-k)$ provided the condition $f_1 > \sum_{k=1}^{r} f_k \exp(-k)$ is satisfied. The first term on the right-hand side of Equation (2.6) is the maximum likelihood estimator of $\theta(t)$ and the second represents a bias correction term.

Under the assumption that species represented the same number of times in the initial sample have equal discovery rates in the community, we propose the following estimator:

$$\hat{\theta}_3(t) = \hat{f}_0\left[1 - \exp(-tf_1/\hat{f}_0)\right], \tag{2.7}$$

where $\hat{f}_0$, as will be discussed and given in Equation (2.11), denotes a predicted number of undetected species in the initial sample. It is clear that the predictor in (2.7) preserves the three properties given above. The asymptotic value of our predictor as t tends to infinity is $\hat{f}_0$. The validity of (2.7) is based on the following approximation formula:

$$\theta(t) \approx E(f_0)\left\{1 - \exp\left[-tE(f_1)/E(f_0)\right]\right\}. \tag{2.8}$$

Note that the right-hand side of (2.8) can be expanded as

$$\sum_{k=1}^{\infty}(-1)^{k+1}t^k\frac{[E(f_1)]^k}{k![E(f_0)]^{k-1}}. \tag{2.9}$$

Comparing (2.4) and (2.9), we see that the proof of (2.8) under the assumption stated above is then equivalent to showing the following: for any fixed integer $k \geq 2$, if all species that are represented j times in the initial sample have equal discovery rates in the community for each $j < k$, then $E(f_k) \approx [E(f_1)]^k/\{k![E(f_0)]^{k-1}\}$. That is, for $k = 2, 3, \ldots$, we have

$$[E(f_0)]^{k-1} \approx [E(f_1)]^k/[k!E(f_k)]. \tag{2.10}$$

The proof details are given in the Appendix.

Assuming a uniform prior in an empirical Bayes argument, Good (1953) and Engen (1978, pp. 31–32) concluded that the (posterior) discovery rate of a species, given it was observed j times in the sample, is approximately $(j+1)E(f_{j+1})/E(f_j)$. For example, all singletons have about the same discovery rate, which can be approximated by $2E(f_2)/E(f_1)$. Therefore, the assumption for our approach is satisfied from a Bayesian statistical point of view.

Our proposed $\hat{f}_0$ is the sample coverage estimator derived by Chao and Lee (1992) and adapted by Chao, Hwang, Chen, and Kuo (2000) for use in the analysis of long-tailed

frequency data. In this approach, the degree of heterogeneity among the species discovery rates is characterized by the parameter called the coefficient of variation (CV or γ). The CV for fixed rates $(\lambda_1, \lambda_2, \ldots, \lambda_S)$ is defined as $\gamma = [S^{-1}\sum_{i=1}^{S}(\lambda_i - \overline{\lambda})^2]^{1/2}/\overline{\lambda}$, where $\overline{\lambda} = \sum_{i=1}^{S}\lambda_i/S$; when the rates are assumed to be a random sample from an unknown distribution with mean μ and variance σ^2, the CV is defined as σ/μ. In both cases, the CV is zero if and only if the species have equal discovery rates. The larger the CV, the greater the degree of heterogeneity among species discovery rates.

Species are divided into two groups (common and rare) by their observed frequencies. Common species are those having more than κ individuals in the sample; the observed rare species are those represented by only one, two, ..., and up to κ individuals in the sample. The prediction of the number of missing species is based entirely on the observed rare species because common species would be discovered anyway and thus they do not contain any information about the missing species. In other words, we consider a subcommunity including rare species only. Let the total number of rare species in the sample be $S_{\text{rare}} = \sum_{i=1}^{\kappa} f_i = \sum_{i=1}^{S} I[0 < X_i \le \kappa]$. Then the number of undetected species based on the estimated sample coverage $\widetilde{C}_{\text{rare}} = 1 - f_1/\sum_{i=1}^{\kappa} if_i$ is predicted by (Chao et al., 2000, sec. 2)

$$\hat{f}_0 = \frac{S_{\text{rare}}}{\widetilde{C}_{\text{rare}}} + \frac{f_1}{\widetilde{C}_{\text{rare}}}\hat{\gamma}^2_{\text{rare}} - S_{\text{rare}}, \tag{2.11}$$

where

$$\hat{\gamma}^2_{\text{rare}} = \max\left\{\frac{S_{\text{rare}}}{\widetilde{C}_{\text{rare}}}\frac{\sum_{i=1}^{\kappa} i(i-1)f_i}{\left(\sum_{i=1}^{\kappa} if_i\right)^2} - 1, 0\right\} \tag{2.12}$$

denotes the estimated squared CV for those species in the subcommunity. We remark that in a multinomial sampling for which the sample size is fixed by design, the estimation procedure is identical except that the estimator for the squared CV is slightly different, that is, the term $(\sum_{i=1}^{\kappa} if_i)^2$ in Equation (2.12) is replaced by $(\sum_{i=1}^{\kappa} if_i)(\sum_{i=1}^{\kappa} if_i - 1)$. It is clearly seen that all frequency counts are needed to compute the estimates proposed by Efron and Thisted (1976) and Boneh et al. (1998). However, as Boneh et al. (1998) indicated, species appearing many times (i.e., high frequencies) have almost no effect on the prediction. Our proposed estimator based on Equations (2.7), (2.11), and (2.12) only uses the first κ frequency counts $(f_1, f_2, \ldots, f_\kappa)$. Chao, Ma, and Yang (1993) suggested a fixed cut-off $\kappa = 10$, based on empirical evidence. The effect of the cut-off point on the proposed estimates will be discussed in the examples and simulations.

As the prediction time tends to infinity, it follows from Equation (2.7) that our predictor tends to $\hat{f}_0$. Consequently, our approach also offers an estimate of the total species richness by adding in the number of discovered species. As shown in the following, this estimate of species richness is needed in computing an estimated standard error of $\hat{\theta}_3(t)$.

Note that the frequencies $f_0, f_1, f_2, \ldots, f_r$ approximately follow a multinomial distribution with an estimated number of species $\hat{S} = D + \hat{f}_0$ and cell probabilities $(\hat{f}_0/\hat{S}, f_1/\hat{S}, f_2/\hat{S}, \ldots, f_r/\hat{S})$. Therefore, a variance estimator of the proposed estimator $\hat{\theta}_3(t)$ can be obtained by using a standard asymptotic approach. That is, we have the following variance

estimator

$$\widehat{\text{var}}(\hat{\theta}_3(t)) = \sum_{i=1}^{\kappa}\sum_{j=1}^{\kappa} \frac{\partial\hat{\theta}_3(t)}{\partial f_i}\frac{\partial\hat{\theta}_3(t)}{\partial f_j}\widehat{\text{cov}}(f_i, f_j), \tag{2.13}$$

where $\widehat{\text{cov}}(f_i, f_j) = f_i(1-f_i/\hat{S})$ for $i = j$ and $\widehat{\text{cov}}(f_i, f_j) = -f_i f_j/\hat{S}$ for $i \neq j$. A similar approach can be used to derive a variance estimator for the estimated number of species and other estimators. The performance of this variance estimator will be investigated in the simulation section.

3. REAL DATA EXAMPLES

3.1 BUTTERFLY DATA

We first consider the Malayan butterfly data analyzed by Fisher et al. (1943) in which 620 species were observed out of 9,031 individuals. All frequencies were given by Williams (1964, p. 19). As described in Section 1, only the first 24 frequencies (given in Table 1) are exact counts, so our method is based on these lower frequency counts.

This dataset has been analyzed by several authors for estimating species richness: Bulmer (1974) obtained an estimate of 815 (SE 43) species using a Poisson-lognormal model; Ord and Whitmore (1986) estimated 719 species based on a Poisson-inverse Gaussian model (but no SE); and Sichel (1997) fitted a Poisson-generalized inverse Gaussian model, obtaining an estimate of 1,000 (but no SE was given).

Under a continuous-time model, the sample time of the initial survey is assumed to be 1. Our goal is to estimate the number of new species expected in an additional time interval with $t = 1/3$, 1/2, 1, and 2, which roughly correspond to the prediction sizes of 3,010, 4,515, 9,030, and 18,060. The following three predictors were calculated:

- The predictor suggested by Efron and Thisted (1976), $\hat{\theta}_1(t)$ (see Equation (2.5));
- Boneh et al. (1998) predictor, $\hat{\theta}_2(t)$ (see Equation (2.6));
- The proposed predictor, $\hat{\theta}_3(t)$, for the cut-off point $\kappa = 5, 6, \ldots, 20$ (see Equation (2.7)).

The results are shown in Table 2, where our estimates are given only for four cut-off values ($\kappa = 5$, 10, 15, and 20). The estimated CV based on Equation (2.12) is also shown for each cut-off. The asymptotic method as derived in Equation (2.13) was used to obtain the estimated SE for each estimator.

Table 2 shows that the estimates proposed by Boneh et al. (1998) are generally lower than the other two. This is consistent with the findings in the simulations (Section 4) that it has severe negative bias in most cases. When predicting a sample with a size less than the original ($t = 1/3$ and 1/2), our estimates are little affected as the cut-off point is increased

Table 2. Comparison of Three Methods for Four Prediction Times for the Butterfly Data. (Numbers in parentheses denote the estimated standard error.)

	Prediction time			
Methods	*t = 1/3*	*t = 1/2*	*t = 1*	*t = 2*
Efron and Thisted (1976)	32.5 (3.6)	45.2 (5.7)	78.0 (24.7)	Overflow
Boneh et al. (1998)	24.5 (1.1)	32.8 (1.5)	49.0 (2.3)	62.2 (3.1)
Proposed in Equation (2.7)				
$\kappa = 5$ ($\hat{\gamma}_{\text{rare}} = 0.33$)	31.2 (3.2)	42.0 (4.7)	62.4 (8.3)	77.0 (12.2)
$\kappa = 10$ ($\hat{\gamma}_{\text{rare}} = 0.66$)	32.0 (3.1)	43.6 (4.5)	66.5 (7.8)	85.0 (11.6)
$\kappa = 15$ ($\hat{\gamma}_{\text{rare}} = 0.77$)	32.6 (3.1)	44.7 (4.4)	69.6 (7.8)	91.2 (11.7)
$\kappa = 20$ ($\hat{\gamma}_{\text{rare}} = 0.86$)	33.2 (3.1)	45.9 (4.4)	73.1 (7.8)	98.8 (12.0)

from 5 to 20. In these cases, our estimates are very close to those of Efron and Thisted (1976). As reflected by the estimated SE in Table 2 and will be verified in the simulations, an advantage of our method over theirs is that our estimator has smaller variation in all cases. Another advantage is that for $t > 1$ our estimate is always obtainable whereas Efron and Thisted (1976) estimate may become extremely large such as in this example for $t = 2$. However, for $t \geq 1$, our estimates become sensitive to the value of the cut-off point. See the discussion in Section 4.

We illustrate our method for the cut-off $\kappa = 10$. In the subcommunity including only the rare species, the number of observed rare species is $S_{\text{rare}} = \sum_{i=1}^{10} f_i = 385$ and there are $\sum_{i=1}^{10} i f_i = 1393$ individuals. In this subcommunity, we have the estimated sample coverage $\widetilde{C}_{\text{rare}} = 1 - f_1 / \sum_{i=1}^{10} i f_i = 91.5\%$ and the CV is estimated by $\hat{\gamma}_{\text{rare}} = 0.66$ based on (2.12). The relatively large estimated CV indicates the existence of heterogeneity in species discovery rates. It follows from (2.11) that the predicted number of unseen species is $\hat{f}_0 = 92$. Adding in the observed species, we obtain an estimate for the species richness $\hat{S} = D + \hat{f}_0 = 620 + 92 = 712$ with an estimated SE of 17.3. If additional samples with subsequently increasing sizes of 3,010, 4,515, 9,030, and 18,060 are conducted, then the corresponding predicted numbers of new species using (2.7) become 32, 44, 67, and 85; and these estimates are increased to $\hat{f}_0 = 92$ if the prediction size tends to be very large.

3.2 POLLUTANTS DATA

Janardan and Schaeffer (1981) presented all frequency counts for 5,720 identifications of chemical compounds in an aquatic environment. There were 1,258 different distinct compounds and the first 24 frequencies are given in Table 3. The prediction results for three different approaches are presented in Table 4 for $t = 1/3$, 1/2, 1, and 2. Our proposed predictor along with the estimated CV is shown for $\kappa = 5, 10, 15$, and 20.

As in the butterfly example, the estimates by Boneh et al. (1998) in Table 4 are the lowest for each fixed prediction time. For $t < 1$, the relative differences in our estimates, as the cut-off point is increased from 5 to 20, are limited, but for $t \geq 1$, the influence of the cut-off point becomes substantial. The estimated CV value for $\kappa = 10$ is $\hat{\gamma}_{\text{rare}} = 0.75$, which indicates evidence of strong heterogeneity in the identification rates of polluants. Our

Table 3. Frequency Counts for the Pollutants Data (Janardan and Schaeffer 1981). (f_i: the number of species represented by i individuals in the sample).

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
f_i	503	238	133	80	56	46	20	14	15	18	15	16	10	10
i	15	16	17	18	19	20	21	22	23	24	>24	D	n	
f_i	9	4	12	6	7	4	4	1	4	0	33	1,258	5,720	

method with $\kappa = 10$ yields comparable estimates to those of Efron and Thisted (except for the case of $t = 2$) but the estimated SE for their estimator is larger in all cases.

Suppose another dataset with a similar size was formed $(t = 1)$; the method of Efron and Thisted (1976) predicts that there would be 346 (SE 34.5) new compounds in the set and our method for $\kappa = 10$ give a very close estimate of 337 (SE 18.0). If two sets of datasets (each of size 5,720) were formed, the additional number of new compounds would be increased to 480 (SE 31.3) using our method for $\kappa = 10$ whereas the estimate by Efron and Thisted (1976) is not obtainable. As t tends to infinity, our estimates of new pollutants approach an asymptotic value of $\hat{f}_0 = 584$ (SE 55.8).

4. SIMULATION STUDY

A simulation study was carried out to examine the performance of the proposed estimator. The number of species in our simulation scenarios was fixed at 1,000. A variety of theoretical models were considered but we only report some representative results. The Poisson process rates $(\lambda_1, \lambda_2, \ldots, \lambda_{1000})$ of these five models, their CV and the expected initial sample sizes are given below. The first model, the homogeneous model, is presented for comparison although the model is rarely adopted in practice. The second model was considered by Boneh et al. (1998). The third model, the Poisson-gamma or negative binomial model, was first proposed by Fisher et al. (1943) and has been used by many others. Models 4 and 5 are the Zipf and Zipf-Mandelbrot models (Zipf 1965; Mandelbrot 1977), which are prevalent in natural sciences and linguistics for modeling frequency data. All the estimation results are shown in Tables 5 to 9.

Model 1: (Homogenous model). $\lambda_1 = \lambda_2 = \cdots = \lambda_{1000} = 1.5$ (CV $= 0$, $E(n) = 1{,}500$).

Table 4. Comparison of Three Methods for Four Prediction Times for the Pollutants Data. (Numbers in parentheses denote the estimated standard error)

.

	Prediction time			
Methods	*t = 1/3*	*t = 1/2*	*t = 1*	*t = 2*
Efron and Thisted (1976)	145.3 (6.9)	204.8 (11.0)	346.0 (34.5)	overflow
Boneh et al. (1998)	92.6 (1.9)	124.6 (2.7)	187.2 (4.3)	239.4 (5.8)
Proposed in Equation (2.7)				
$\kappa = 5$ ($\hat{\gamma}_{\text{rare}} = 0.40$)	140.1 (6.5)	193.0 (9.7)	303.8 (18.3)	403.9 (30.2)
$\kappa = 10$ ($\hat{\gamma}_{\text{rare}} = 0.75$)	145.7 (6.3)	204.3 (9.4)	337.2 (18.0)	479.6 (31.3)
$\kappa = 15$ ($\hat{\gamma}_{\text{rare}} = 0.97$)	150.1 (6.3)	213.4 (9.3)	365.4 (18.0)	551.1 (32.2)
$\kappa = 20$ ($\hat{\gamma}_{\text{rare}} = 1.11$)	152.5 (6.3)	218.4 (9.3)	382.0 (18.0)	596.2 (32.8)

Table 5. Simulation Comparison of Three Methods for Four Prediction Times. (Homogeneous Model, 5,000 Trials, $\overline{D} = 777.0$)

Prediction time	*True value/methods*	*Average estimate*	*Average estimated SE*	*Sample SE*	*Sample RMSE*
$t = 0.5$	True value $\theta(t) = 117.7$				
	Efron and Thisted (1976)	117.7	9.4	9.3	9.3
	Boneh et al. (1998)	97.7	1.8	2.1	20.1
	Proposed in Equation (2.7) $\kappa = 5$	118.6	6.5	6.6	6.6
	$\kappa = 10$	118.5	7.0	7.0	7.0
	$\kappa = 15$	118.5	7.0	7.0	7.0
$t = 1.0$	True value $\theta(t) = 173.3$				
	Efron and Thisted (1976)	173.1	27.3	27.2	27.2
	Boneh et al. (1998)	146.0	3.0	3.2	27.6
	Proposed in Equation (2.7) $\kappa = 5$	175.5	11.1	11.2	11.4
	$\kappa = 10$	175.4	12.2	12.3	12.5
	$\kappa = 15$	175.4	12.2	12.3	12.5
$t = 2.0$	True value $\theta(t)$ = 212.0				
	Efron and Thisted (1976)	202.3	284.9	306.7	306.9
	Boneh et al. (1998)	185.1	4.1	4.2	27.3
	Proposed in Equation (2.7) $\kappa = 5$	216.0	15.7	15.8	16.3
	$\kappa = 10$	215.9	17.9	18.0	18.4
	$\kappa = 15$	215.9	17.9	18.0	18.4
$t = 3.0$	True value $\theta(t) = 220.7$				
	Efron and Thisted (1976)	−12.9	3,933.7	7,738.3	7,741.8
	Boneh et al. (1998)	197.4	4.6	4.5	23.7
	Proposed in Equation (2.7) $\kappa = 5$	225.4	17.2	17.4	18.0
	$\kappa = 10$	225.3	19.9	20.0	20.5
	$\kappa = 15$	225.3	19.9	20.0	20.5

Model 2: (Two-class model). There are two classes for $(\lambda_1, \lambda_2, \ldots, \lambda_{1000})$. Specifically, $\lambda_i = 2.0$ for $i = 1, 2, \ldots, 900$, and $\lambda_i = 0.2$ for $i = 901, 902, \ldots, 1{,}000$ (CV = 0.30, $E(n) = 1{,}820$).

Model 3: (Gamma model). $\lambda_i = 1.2u_i$, where $(u_1, u_2, \ldots, u_{1000})$ are a random sample from a gamma density $f(u) = \beta^\alpha u^{\alpha-1} \exp(-\beta u)/\Gamma(\alpha)$ with $\alpha = 0.5$ and $\beta = 1.0$ (CV = 1.41, $E(n) = 600$).

Model 4: (Zipf model). $\lambda_i = 400/(i + 10), i = 1, 2, \ldots, 1{,}000$ (CV = 1.88, $E(n) = 1{,}842$).

Model 5: (Zipf-Mandelbrot model). $\lambda_i = 7200/(i + 20)^{1.3}, i = 1, 2, \ldots, 1000$ (CV = 2.18, $E(n) = 6{,}695$).

For each model, we considered four prediction times ($t = 0.5$, 1, 2, and 3). For each prediction time, 5,000 simulated datasets (each included an original sample and four prediction samples) were generated. We recorded D (the number of distinct species discovered in the initial sample) and calculated $U(t)$ (the number of new species discovered in the second interval with length t) for each generated dataset. The average of the number of distinct species found in the original sample, $\overline{D}$, is given in the table title for each model. The average value of $U(t)$ over 5,000 trials is almost identical to the theoretical value (i.e., $\theta(t)$); thus only the latter is shown. The following three predictors and their asymptotic SE were computed: the predictors suggested by Efron and Thisted (1976), Boneh et al. (1998),

Table 6. Simulation Comparison of Three Methods for Four Prediction Times. (Two-Class Model, 5,000 Trials, $\overline{D} = 796.3$)

Prediction time	*True value/methods*	*Average estimate*	*Average estimated SE*	*Sample SE*	*Sample RMSE*
$t = 0.5$	True value $\theta(t) = 84.8$				
	Efron and Thisted (1976)	85.0	8.7	8.7	8.7
	Boneh et al. (1998)	88.8	1.7	2.0	4.5
	Proposed in Equation (2.7) $\kappa = 5$	84.5	5.3	5.4	5.4
	$\kappa = 10$	84.1	6.0	6.0	6.1
	$\kappa = 15$	84.1	6.0	6.0	6.1
$t = 1.0$	True value $\theta(t) = 120.2$				
	Efron and Thisted (1976)	120.8	27.9	27.8	27.9
	Boneh et al. (1998)	132.0	2.7	3.0	12.2
	Proposed in Equation (2.7) $\kappa = 5$	117.6	8.2	8.3	8.7
	$\kappa = 10$	116.9	9.8	9.9	10.4
	$\kappa = 15$	116.9	9.8	9.9	10.4
$t = 2.0$	True value $\theta(t) = 146.6$				
	Efron and Thisted (1976)	149.6	534.3	576.4	576.4
	Boneh et al. (1998)	166.6	3.8	3.8	20.4
	Proposed in Equation (2.7) $\kappa = 5$	135.7	10.4	10.5	15.1
	$\kappa = 10$	134.6	13.0	13.1	17.7
	$\kappa = 15$	134.6	13.0	13.1	17.7
$t = 3.0$	True value $\theta(t) = 158.4$				
	Efron and Thisted (1976)	−82.4	12,831.6	28,582.6	28,583.6
	Boneh et al. (1998)	177.3	4.2	4.0	19.3
	Proposed in Equation (2.7) $\kappa = 5$	138.5	10.8	10.9	22.7
	$\kappa = 10$	137.3	13.7	13.8	25.2
	$\kappa = 15$	137.4	13.8	13.9	25.2

and the proposed predictor ($\kappa = 5$, 10, and 15). The resulting 5,000 estimates as well as their estimated SE were averaged to give the results "Average estimate" and "Average estimated SE" in Tables 5–9. Also, the sample SE and sample root mean squared errors (RMSE) for each estimator are shown.

We first note the effect of the cut-off point on our prediction results. Based on Tables 5–9 and other (unreported) simulation results for κ between 5 and 15, the effect is small when the CV is relatively low as in the first two models (Tables 5 and 6). A low CV means smaller variation among discovery rates, so all the expected frequencies tend to be concentrated in a narrow range. Thus, almost no difficulty arises in separating common and rare species. For the other three models (Tables 7–9), a relatively large CV implies long-tail frequency data and some species are hardly detectable; thus our method becomes sensitive to the choice of the cut-off point. The effect is limited when the prediction interval is shorter than the original one $(t < 1)$, as also noted in the data analyses. However, the cut-off point has substantial effect for predicting a longer interval $(t \geq 1)$ in the case of strong heterogeneity. In the Zipf and Zipf-Mandelbrot models, our method underestimates for $\kappa = 5$ whereas it overestimates for $\kappa = 15$. This is also a reason for our proposal of setting $\kappa = 10$.

For comparisons of the three approaches, we have concluded the following separately for two cases:

Table 7. Simulation Comparison of Three Methods for Four Prediction Times. (Gamma Model, 5,000 Trials, $\overline{D} = 325.9$)

Prediction time	*True value/methods*	*Average estimate*	*Average estimated SE*	*Sample SE*	*Sample RMSE*
$t = 0.5$	True value $\theta(t) = 76.6$				
	Efron and Thisted (1976)	76.8	6.4	6.7	6.7
	Boneh et al. (1998)	42.7	1.5	2.1	33.9
	Proposed in Equation (2.7) $\kappa = 5$	74.6	5.8	6.2	6.5
	$\kappa = 10$	76.8	5.8	6.2	6.2
	$\kappa = 15$	77.1	5.8	6.2	6.2
$t = 1.0$	True value $\theta(t) = 131.9$				
	Efron and Thisted (1976)	132.4	17.2	17.6	17.6
	Boneh et al. (1998)	64.6	2.5	3.3	67.4
	Proposed in Equation (2.7) $\kappa = 5$	122.6	11.6	12.0	15.2
	$\kappa = 10$	129.5	11.6	12.2	12.5
	$\kappa = 15$	130.5	11.7	12.3	12.4
$t = 2.0$	True value $\theta(t) = 208.0$				
	Efron and Thisted (1976)	547.2	5,297.3	48,365.2	48,366.3
	Boneh et al. (1998)	83.1	3.4	4.3	124.9
	Proposed in Equation (2.7) $\kappa = 5$	173.8	20.8	21.5	40.4
	$\kappa = 10$	190.7	21.8	22.5	28.4
	$\kappa = 15$	193.5	22.2	23.1	27.2
$t = 3.0$	True value $\theta(t) = 258.9$				
	Efron and Thisted (1976)	-4.3×10^7	9.2×10^7	4.2×10^9	4.2×10^9
	Boneh et al. (1998)	89.3	3.7	4.7	169.7
	Proposed in Equation (2.7) $\kappa = 5$	194.8	26.6	27.4	69.7
	$\kappa = 10$	219.2	28.9	29.8	49.6
	$\kappa = 15$	223.6	29.8	31.1	47.1

1. When the prediction sample is not larger than the original one (i.e., $t \leq 1$): In this situation, the method proposed by Efron and Thisted (1976) works well. Our prediction results for $\kappa = 10$ are very close to their estimates. The proposed estimator is slightly biased but has less variation whereas Efron and Thisted's estimator is unbiased but has larger variation. Overall, our predictor has smaller RMSE in all cases. Both methods are generally preferable to the method of Boneh et al. (1998) except for the two-class model.

2. When the prediction sample is larger than the original one (i.e., $t > 1$): In the prediction for a larger prediction sample, the estimator suggested by Efron and Thisted (1976) could become either negative or extremely large. In such cases, the standard errors are unboundedly large. The other two estimators are free of the above erratic behavior associated with Efron and Thisted's estimator. Although the predictor suggested by Boneh et al. (1998) performs satisfactorily in the two-class model, it is severely biased downwards in the other models. Comparing all the estimates with the true values, we see that the estimator of Boneh et al. (1998) for our cases does not seem to increase at an anticipated rate as t is increased. The proposed new method appears to be promisingly useful for predicting a larger survey as well.

As expected, the performance of our estimator deteriorates as the prediction time interval is increased except for a homogeneous population. In some models (the homogeneous,

Table 8. Simulation Comparison of Three Methods for Four Prediction Times. (Zipf's Model, 5,000 Trials, $\overline{D} = 603.3$)

Prediction time	*True value/methods*	*Average estimate*	*Average estimated SE*	*Sample SE*	*Sample RMSE*
$t = 0.5$	True value $\theta(t) = 114.5$				
	Efron and Thisted (1976)	114.5	8.2	8.0	8.0
	Boneh et al. (1998)	69.2	2.0	2.2	45.3
	Proposed in Equation (2.7) $\kappa = 5$	111.2	7.3	7.0	7.7
	$\kappa = 10$	118.0	7.1	6.8	7.7
	$\kappa = 15$	122.4	7.1	6.8	10.4
$t = 1.0$	True value $\theta(t) = 190.8$				
	Efron and Thisted (1976)	190.5	23.8	23.5	23.5
	Boneh et al. (1998)	104.2	3.1	3.4	86.7
	Proposed in Equation (2.7) $\kappa = 5$	178.3	14.1	13.6	18.5
	$\kappa = 10$	199.1	13.9	13.3	15.7
	$\kappa = 15$	212.9	13.9	13.3	25.7
$t = 2.0$	True value $\theta(t) = 282.1$				
	Efron and Thisted (1976)	2.8×10^{14}	4.8×10^{14}	1.2×10^{16}	1.2×10^{16}
	Boneh et al. (1998)	133.5	4.3	4.5	148.7
	Proposed in Equation (2.7) $\kappa = 5$	242.9	24.0	23.4	45.7
	$\kappa = 10$	292.3	25.1	24.2	26.3
	$\kappa = 15$	328.9	26.2	24.9	53.0
$t = 3.0$	True value $\theta(t) = 330.5$				
	Efron and Thisted (1976)	1.6×10^{25}	3.6×10^{25}	1.9×10^{27}	1.9×10^{27}
	Boneh et al. (1998)	143.2	4.7	4.9	187.3
	Proposed in Equation (2.7) $\kappa = 5$	266.5	29.4	28.7	70.2
	$\kappa = 10$	335.8	32.7	31.3	31.8
	$\kappa = 15$	392.1	35.6	33.3	70.1

Zipf, and Zipf-Mandelbrot models), our method for a cut-off value of 10 produces reasonable results even for the prediction of a long interval three times the original sample. For the other two models, however, when the prediction sample time is longer than the initial sample, our approach might result in severe bias, although it is still superior to the others. This is a common limitation for making inferences over a very long prospective time.

The variance formula in Equation (2.13) produces satisfactory estimates for our proposed estimator because in all cases the estimated standard error is very close to the sample standard error. The estimated standard error can be subsequently used to construct an associated confidence interval.

5. CONCLUSION AND DISCUSSION

This article proposes a new approach to predicting the number of new species in further sampling. The new predictor maintains the desirable properties of the theoretical prediction function. Our predictor generally improves the existing methods based on a simulation study for various models with heterogeneous species discovery rates.

Our proposed method only requires the lower order frequency counts $(f_1, f_2, \ldots, f_\kappa)$, where κ is a cut-off point dividing the species into two groups (common and rare). Other

Table 9. Simulation Comparison of Three Methods for Four Prediction Times. (Zipf-Mandelbrot Model, 5,000 Trials, $\overline{D} = 848.7$)

Prediction time	*True value/methods*	*Average estimate*	*Average estimated SE*	*Sample SE*	*Sample RMSE*
$t = 0.5$	True value $\theta(t) = 72.8$				
	Efron and Thisted (1976)	72.8	7.7	7.6	7.6
	Boneh et al. (1998)	66.0	1.8	1.8	7.0
	Proposed in Equation (2.7) $\kappa = 5$	67.7	5.7	5.4	7.4
	$\kappa = 10$	75.5	5.7	5.4	6.0
	$\kappa = 15$	80.7	5.8	5.4	9.5
$t = 1.0$	True value $\theta(t) = 108.7$				
	Efron and Thisted (1976)	108.1	28.9	29.2	29.3
	Boneh et al. (1998)	98.0	2.9	2.8	11.1
	Proposed in Equation (2.7) $\kappa = 5$	94.2	9.4	8.8	17.0
	$\kappa = 10$	113.3	10.0	9.2	10.3
	$\kappa = 15$	127.7	10.3	9.4	21.2
$t = 2.0$	True value $\theta(t) = 137.8$				
	Efron and Thisted (1976)	4.1×10^{52}	4.1×10^{52}	2.8×10^{54}	2.8×10^{54}
	Boneh et al. (1998)	123.9	3.9	3.5	14.4
	Proposed in Equation (2.7) $\kappa = 5$	109.0	12.5	12.0	31.2
	$\kappa = 10$	142.4	14.5	13.5	14.3
	$\kappa = 15$	171.8	16.1	14.6	37.0
$t = 3.0$	True value $\theta(t) = 146.8$				
	Efron and Thisted (1976)	-1.5×10^{84}	1.5×10^{84}	8.8×10^{85}	8.8×10^{85}
	Boneh et al. (1998)	131.8	4.3	3.7	15.5
	Proposed in Equation (2.7) $\kappa = 5$	111.5	13.3	12.6	37.5
	$\kappa = 10$	150.0	16.3	14.9	15.2
	$\kappa = 15$	187.2	18.8	16.8	43.8

estimators require all frequencies. As discussed in Section 2 and simulations, high order frequencies $(f_{\kappa+1}, f_{\kappa+2}, \ldots)$ have almost no effect on the predicted number of new species in a second survey. Hence an advantage of our method is that the exact frequencies for relatively common species are not necessary; this implies that ecologists and biologists do not need to expend efforts on recording the exact frequencies of those relatively common species once they have been encountered a sufficient number of times in the sample. However, a practical data-dependent guideline about how large is "sufficient," that is, the choice of the cut-off, is still unclear to us and should be studied more. From the numerical results in Sections 3 and 4, we see that this cut-off problem arises mainly in predicting a longer interval for highly heterogeneous communities. In such cases, it is suggested that several cut-off values should be considered and the behavior of the estimates be examined.

A current area of research involves extending the method to the prediction based on presence/absence data. In some biological surveys, counting the exact number of individuals up to the cut-off value may become impossible (e.g., in plant communities where trees are densely distributed), presence/absence data are often collected for each species over several quadrats or sampling points. It would be useful if a similar predictor can be derived based on incidence counts.

Because of the analogy between populations and communities, the proposed method can also be applied to sampling individuals of a single species. In population-level stud-

ies, each individual can be regarded as a species and data consist of individual capture frequencies. Tagging or marking is generally needed to distinguish individuals caught in the sampling process. Using the proposed method, we can predict the number of new individuals that will be found in additional sampling occasions. Similarly, we can apply the proposed method to estimate the number of undiscovered genes or alleles based on frequency counts (Huang and Weir 2001) and to predict undercount in surveillance records (Bunge and Fitzpatrick 1993).

The computer program SPADE (Species Prediction And Diversity Estimation), written in C language, that calculates all estimates discussed in this article and some diversity indices is available upon request and is freely downloadable at the first author's Web site at http://chao.stat.nthu.edu.tw/.

APPENDIX: PROOF OF EQUATION (2.10)

Based on Equation (2.3), we can write

$$E(f_0) = \sum_{i=1}^{S} e^{-\lambda_i} = \sum_{i=1}^{S} (\lambda_i e^{-\lambda_i})/\lambda_i = E \sum_{i=1}^{S} I[X_i = 1]/\lambda_i. \tag{A.1}$$

Note that the last summation in (A.1) represents the total of the inverse rates associated with the species which appeared only once in the sample (i.e., singletons). If all singletons have identical discovery rates (Poisson rates) in the community and we denote this common rate by α_1, then $E(f_0) \approx E(f_1/\alpha_1)$ as a result of (A.1). Assuming a uniform prior, Good (1953) and Engen (1978, p. 31–32) used an empirical Bayes argument and concluded that the (posterior) rate of a species, given it was encountered k times in the sample, is approximately equal to $(k+1)E(f_{k+1})/E(f_k)$. In particular, all singletons have an approximate discovery rate of $2E(f_2)/E(f_1)$. Substituting $\alpha_1 \approx 2E(f_2)/E(f_1)$ into the formula $E(f_0) \approx E(f_1/\alpha_1)$, we obtain $E(f_0) \approx [E(f_1)]^2/[2E(f_2)]$. Consequently, Equation (2.10) is valid for $k = 2$.

We now prove that Equation (2.10) is also valid for any $k > 2$ by a mathematical induction. A generalized form for Equation (A.1) is given by

$$E(f_0) = \sum_{i=1}^{S} e^{-\lambda_i} = \sum_{i=1}^{S} \frac{k!}{\lambda_i^k} \frac{\lambda_i^k e^{-\lambda_i}}{k!} = E \sum_{i=1}^{S} \frac{k!}{\lambda_i^k} I[X_i = k]. \tag{A.2}$$

A similar derivation parallel to that for $k = 2$ follows. Assume those species represented k times have equal rates and denote this common rate by α_k. Then Equation (A.2) implies that $E(f_0) \approx E(k! f_k/\alpha_k^k)$. Using Good's (1953) result, as discussed earlier, we have $\alpha_k \approx (k+1)E(f_{k+1})/E(f_k)$ and the following formula

$$E(f_0) \approx k! E(f_k)[E(f_k)]^k / \{(k+1)^k [E(f_{k+1})]^k\}. \tag{A.3}$$

If Equation (2.10) is valid for k, then $E(f_k)$ in (A.3) can be replaced by $[E(f_1)]^k / \{k![E(f_0)]^{k-1}\}$. Then it is seen that Equation (2.10) also holds for $k+1$.

ACKNOWLEDGMENTS

The authors thank the editor, an associate editor, and the reviewers for providing thoughtful suggestions and improving the exposition. The research was supported by the National Science Council of Taiwan.

[Received October 2002. Revised August 2003.]

REFERENCES

Agresti, A. (1994), "Simple Capture-Recapture Models Permitting Unequal Catchability and Variable Sampling Effort," *Biometrics*, 50, 494–500.

Boneh, S., Boneh, A., and Caron, R. J. (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., Nichols, J. D., Sauer, J. R., Hines, J. E., and Pollock, K. H. (1998), "Estimating Species Richness: the Importance of Heterogeneity in Species Detectability," *Ecology*, 79, 1018–1028.

Bulmer, M. G. (1974), "On Fitting the Poisson Lognormal Distribution to Species Abundance Data," *Biometrics*, 30, 101–110.

Bunge, J., and Fitzpatrick, M. (1993), "Estimating the Number of Species: A Review," *Journal of the American Statistical Association*, 88, 364–373.

Burnham, K. P., and Overton, W. S. (1978), "Estimating the Size of a Closed Population When Capture Probabilities Vary Among Animals," *Biometrika*, 65, 625–633.

Chao, A., and Lee, S.-M. (1992), "Estimating the Number of Classes via Sample Coverage," *Journal of the American Statistical Association*, 87, 210–217.

Chao, A., Hwang, W.-H., Chen, Y.-C., and Kuo, C.-Y. (2000), "Estimating the Number of Shared Species in Two Communities," *Statistica Sinica*, 10, 227–246.

Chao, A., Ma, M.-C., and Yang, M. C. K. (1993), "Stopping Rules and Estimation for Recapture Debugging With Unequal Failure Rates," *Biometrika*, 80, 193–201.

Colwell, R. K., and Coddington, J. A. (1994), "Estimating Terrestrial Biodiversity Through Extrapolation," *Philosophical Transactions of the Royal Society*, London B, 345, 101–118.

Coull, B. A., and Agresti, A. (1999), "The Use of Mixed Logit Models to Reflect Heterogeneity in Capture-Recapture Studies," *Biometrics*, 55, 294–301.

Efron, B., and Thisted, R. (1976), "Estimating the Number of Unseen Species: How Many Words Did Shakespeare Know?" *Biometrika*, 63, 435–447.

Engen, S. (1978), *Stochastic Abundance Models*, London: Chapman and Hall.

Fisher, R. A., Corbet, A. S., and Williams, C. B. (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 Toulmin, G. H. (1956), "The Number of New Species, and the Increase in Population Coverage, When a Sample is Increased," *Biometrika*, 43, 45–63.

Huang, S. P., and Weir, B. S. (2001), "Estimating the Total Number of Alleles Using a Sample Coverage Method," *Genetics*, 159, 1365–1373.

Janardan, K. G., and Schaeffer, D. J. (1981), "Methods for Estimating the Number of Identifiable Organic Pollutants in the Aquatic Environment," *Water Resources Research*, 17, 243–249.

Keating, K. A., Quinn, J. F., Ivie, M. A., and Ivie, L. L. (1998), "Estimating the Effectiveness of Further Sampling in Species Inventories," *Ecological Applications*, 8, 1239–1249.

Lloyd, C. J., and Yip, P. (1991), "A Unification of Inference for Capture-Recapture Studies Through Martingale Estimating Functions," in *Estimating Equations*, ed. V. P. Godambe, Oxford: Clarendon Press, pp. 65–88.

Mandelbrot, B. (1977), *Fractals, Form, Chance and Dimension*, San Francisco: Freeman.

Norris III, J. L., and Pollock, K. H. (1998), "Non-Parametric MLE for Poisson Species Abundance Models Allowing for Heterogeneity Between Species," *Environmental and Ecological Statistics*, 5, 391–402.

Ord, J. K., and Whitmore, G. A. (1986), "The Poisson-Inverse Gaussian Distribution as a Model for Species Abundance," *Communications in Statistics, Part A—Theory and Methods*, 15, 853–871.

Shen, T.-J., Chao, A., and Lin, C.-F. (2003), "Predicting the Number of New Species in Further Taxonomic Sampling," *Ecology*, 84, 798–804.

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.

Smith, E. P., and van Belle, G. (1984), "Nonparametric Estimation of Species Richness," *Biometrics*, 40, 119–129.

Solow, A. R., and Polasky, S. (1999), "A Quick Estimator for Taxonomic Surveys," *Ecology*, 80, 2799–2803.

Williams, C. B. (1964), *Patterns in the Balance of Nature*, London: Academic Press.

Zipf, G. K. (1965), *Human Behavior and Principle of Least Effort*, New York: Addison-Wesley.