Journal of the Royal Statistical Society Series C: Applied Statistics, 2025, **74**, 884–903
https://doi.org/10.1093/jrsssc/qlaf005
Advance access publication 29 January 2025
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series C
Applied Statistics

A design-based view of species richness estimation in environmental surveys

Rosa M. Di Biase[1,2], Lorenzo Fattorini[1] and Agnese Marcelli[3]

[1]Department of Economics and Statistics, University of Siena, Piazza San Francesco 8, 53100 Siena, Italy
[2]NBFC, National Biodiversity Future Center, Piazza Marina 61, 90133 Palermo, Italy
[3]Department for Innovation in Biological, Agro-food and Forest Systems, University of Tuscia, via San Camillo de Lellis snc, 01100 Viterbo, Italy

Address for correspondence: Rosa M. Di Biase, Department of Economics and Statistics, University of Siena, Piazza San Francesco 8, 53100 Siena, Italy; NBFC, National Biodiversity Future Center, Piazza Marina 61, 90133 Palermo, Italy.
Email: rosa.dibiase@unisi.it

Abstract

In this work, the estimation of species richness is approached from a design-based perspective, considering the probabilistic sampling of species and checking the performance of the estimators automated in the SPADE software. As shown theoretically and by a simulation study, these estimators are affected by a massive negative bias. To reduce the underestimation of species richness, data integration is attempted, by exploiting the list of rare species compiled by purposive surveys. Richness estimation is then performed on the residual community of species not in the list, and a bootstrap mean squared error estimator is applied. A simulation study and the application to four case studies produce encouraging results.

Keywords: data integration, incidence data, purposive surveys, SPADE software, species loss, species sampling

1 Introduction

When communities of animals or plants contain species with very different characteristics and incomparable abundances, the number of species in the study region, usually referred to as species richness, constitutes the most straightforward method of analysing diversity (Hurlbert, 1971). In most cases, the complete list of species is an unknown characteristic of natural communities, so species richness constitutes an unknown parameter. Species lists can be compiled by purposive investigations, traditionally performed by botanists (Palmer et al., 2002), or through sample surveys of probabilistic nature (Bunge & Fitzpatrick, 1993).

When species lists are compiled by subjectively searching for species, nothing can be stated about the reliability of these investigations, or, as pointed out by Stevens (1994, Section 2.2), only model-based inference can be attempted. In these cases, the author underlies as the properties of species richness estimators are not objective because they strictly depend on the more or less realistic assumptions of the model that has been supposed to have generated the sample. Notably, this type of inference is traditionally adopted in species richness estimation, in which, as discussed later in the article, data are traditionally supposed to be generated by the so-called Bernoulli product model (Colwell et al., 2012). On this issue, Chiarucci (2007) stigmatizes naturalists 'that continue to ignore these arguments and collect their data on the basis of preferential choices', suggesting that in these cases they should avoid the term sampling and should avoid any statistical inference, simply describing their works as descriptive field recognition of natural communities. Similarly, Albert et al. (2010) note that probabilistic sampling is often difficult to perform when surveying natural communities due to logistic problems such as time, money and accessibility and that, 'unfortunately', 'this often leads

Received: June 22, 2023. **Revised:** August 31, 2024. **Accepted:** January 6, 2025

to the implementation of simplified or so-called convenience sampling designs which, like all nonprobabilistic designs, suffer from unknown biases'. In addition, Boyd et al. (2023) evidence that, in the era of big data, naturalists now have access to very large nonprobability samples obtained from digitized museum and herbarium collections or from data collected for species atlases or citizen science. Therefore, naturalists often justify the use of nonprobability samples on the basis of their large sizes. Additionally, in this case, the authors warn against the use of these nonprobability large sets of data, as 'quantity of data is no substitute for representativeness'.

Alternatively, if probabilistic sampling schemes are executed, the resulting estimators of species richness can be objectively evaluated from their design-based sampling distributions arising from the schemes actually adopted in the field rather than on the basis of some assumptions about sampling and about the communities under study (e.g. Gregoire, 1998; Särndal et al., 1992). As stated by Särndal et al. (1992, p. 21), the great appeal of a design-based approach is that 'Design-based inference is objective, nobody can challenge that the sample was really selected according to the given sampling design. The probability distribution associated with the design is real, not modelled or assumed'. Nevertheless, in their seminal paper, Palmer et al. (2002) warn botanists against the use of probabilistic sampling (what they call objective methods) because objective methods 'are likely to miss the rare or unclassifiable habitats that are likely to contribute the most to regional diversity. The error involved in extrapolation can be tremendous. Indeed, it is unlikely that such methods can outperform the guesses of experienced botanists'. The authors also emphasize that 'Experienced botanists generally have a strong intuition or 'educated guess' about where to direct one's efforts in collecting specimens. Botanists are drawn to unusual habitats, edges between vegetation types, environments suspected to be species rich, and geomorphic features such as bluffs, outcrops, streams, etc. Botanists often intensify their efforts when they detect new or scarce species, and move more quickly through areas in which the species are commonplace'. Finally, the authors conclude that 'The experienced botanist, using internal 'algorithms', will most likely outperform any objective methodology if the goal is to maximize species encountered'.

Based on these considerations, the purpose of this article is to conciliate these conflicting opinions by adopting a sort of data integration, i.e. the merging of probabilistic and opportunistic, preferential data, that has been successfully attempted in several environmental surveys (e.g. Adde et al., 2021; Bowler et al., 2019). In this framework, our goal is to exploit the valuable information acquired by nonprobabilistic surveys, especially regarding the rarest species which are difficult to sample, to modify an estimate of species richness obtained from a probabilistic sample. In this way, considering the information acquired from the nonprobabilistic survey to be fixed, the uncertainty of the resulting data integration estimator continues to stem from the sole sampling scheme adopted in the probabilistic survey, so that a rigorous and objective design-based inference can be performed.

Accordingly, the first step is to choose a sample-based estimator of species richness with appealing design-based properties to be subsequently modified based on the information acquired from the nonprobability sample. To this end, for the first time in the literature, our study views the estimation of species richness from a design-based perspective, in which species are selected by means of independent replications of a probabilistic sampling scheme (e.g. plots or transects randomly and independently located in the study region).

The familiar estimators of species richness, automated in the popular SPADE software and then widely applied in environmental studies—in most cases disregarding the sampling protocols—are checked theoretically from a design-based perspective. In addition, empirical testing of the performance of these estimators is carried out by a simulation study performed on two real communities of trees: the Barro Colorado Island (BCI) community and the Harvard Forest (HF) community. The results of our theoretical and empirical investigations are discouraging and should warn naturalists against the uncritical use of the software, with all the estimators automated in SPADE affected by large negative biases due to the missing of rarest species. Moreover, the results recognize the Lee and Chao (1994) original estimator as the best design-based performer and then as the unique candidate to be adopted in data integration.

The article is organized as follows. In Section 2, we introduce the issue of the probabilistic sampling of species, distinguishing between the sampling of plant and animal communities. In Section 3, we consider the design-based characteristics of data arising from the probabilistic sampling of species, outlining their few properties that simply and uniquely stem from the independent replications of the sampling schemes. In Section 3, we also distinguish between sampling schemes in which agglomerates

of units are selected at each replication and sampling schemes that instead select single units, indicating that schemes selecting single units are not considered in the article because they are difficult to implement in natural communities, where selections are usually performed by plots, transects, nests, or traps. In Section 4, we describe the rich sequence of results achieved in a model-based framework under the Bernoulli product model, which has generated a set of estimators of wide use, referred to as nonparametric estimators, most of which are automated in the free software SPADE (Chao et al., 2015). In Section 5, we approach these nonparametric estimators from a design-based perspective, outlining some theoretical drawbacks that may cause large negative bias and recognizing the estimator originally proposed by Lee and Chao (1994) as the most appealing from a theoretical point of view. In Section 6, we perform a design-based simulation study on BCI and HF communities that fully confirms the theoretical concerns raised about nonparametric estimators, whereas the same simulation extended in Section 7 confirms the Lee and Chao (1994) estimator to be the best performer. Therefore, in the same section, we attempt to reduce bias by using a data integration approach based on the original idea proposed by Chiarucci et al. (2018), which involves exploiting the lists of rare species usually compiled by ecologists in purposive surveys. These lists are used to recognize those species that are surely present in the community and then to estimate the richness in the residual community of species not included in the purposive lists. To this end, we adopt the Lee and Chao (1994) estimator to estimate the richness in the residual community. We check the performance of the proposed data integration estimator and a bootstrap estimator of its mean squared error via the same simulation study described in Section 6. In Section 8, the proposed method is applied to estimate the species richness in four communities of vascular plants located in nature reserves and parks in Central Italy. Concluding remarks are presented in Section 9. For brevity, details on the basic concepts of design-based inference are not reported and can be found in the first chapters of many textbooks on survey sampling (e.g. Fuller, 2009; Hedayat & Sinha, 1991; Särndal et al., 1992; Thompson, 2012), while details on sampling plant and animal communities, on BCI and HF communities and some proofs and tables are reported in the appendices of the online supplementary material file.

2 Probabilistic sampling of species

We consider a natural community (animals or plants) within a study region A. The community can be viewed as a without-frame population U of N units scattered on A. If K species are present, the community is partitioned into K subpopulations $U_1, \ldots, U_K$ of size $N_1, \ldots, N_K$, where U_l denotes the set of N_l units belonging to species l. We denote the frame of species partitioning the community as the species list L and the number of species K as the species richness.

$S \subset U$ denotes the sample of units selected from the community by means of a suitable sampling scheme that induces the inclusion probabilities of single units. Usually, schemes suitable for natural communities are such that the inclusion probabilities can be determined at least for the selected units. This feature is important because it allows for the Horvitz–Thompson estimation of the totals of some interest attributes (e.g. abundance, biomass, and basal area). However, when species rather than units are to be sampled, we view each subpopulation U_l as a unit itself and the species list L as a population. Because species are difficult to sample as unknown assemblages of units spread across the study region, the natural way to sample species is to sample units so that a species is sampled if at least one unit of that species is sampled. Accordingly, any possible sample of units $S \subset U$ univocally determines the corresponding sample of species $D \subset L$, and the scheme adopted to sample units univocally determines the species sampling design, i.e. the probability distribution over the collection $\mathfrak{D}$ of all the possible samples of species. In turn, the species sampling design determines the species inclusion probabilities $\pi_1, \ldots, \pi_L$, i.e. the probabilities that single species are selected, and the species joint inclusion probabilities $\pi_{l,h}$ for each $h > l \in L$, i.e. the probabilities that pairs of species are selected jointly. The joint inclusion probabilities strictly depend on species overlap or avoidance, i.e. species that tend to occupy similar habitats have high probabilities of being sampled jointly, as opposed to species that tend to occupy different habitats. Independence of species inclusion, i.e. $\pi_{l,h} = \pi_l \pi_h$ for each $h > l \in L$, never occurs for the sampling schemes usually adopted in natural communities.

Notably, even if the schemes adopted to sample units allow for the quantification of their inclusion probabilities, the quantification of species inclusion probabilities is generally precluded because it would entail the knowledge of all of the units belonging to a species together with their locations in

the study region. This fact precludes the use of the Horvitz–Thompson criterion for estimating species richness. We detail schemes for sampling units (and subsequently species) in natural communities in the online supplementary material, Appendix A, where we distinguish between the sampling of plants or animals. Animals are usually sampled if they are observed from transects or points or by cameras, or if they are captured by nests or traps. Therefore, the probability of sampling an animal depends on many factors (e.g. visibility conditions and observer ability) and cannot be physically defined in terms of inclusion regions, as in the case of plant communities (online supplementary material, Appendix A). Usually, some assumptions are introduced to perform inference on animal species richness. Therefore, even if we generally refer to natural communities, it is worth noting that a genuine design-based approach to species richness estimation is possible only for plant communities in which species inclusion probabilities are well defined (even if unknown) and do not vary among sampling occasions. The same inference can also be performed to estimate the richness of animal communities but at the cost of supposing that there is a well-defined inclusion probability for each species that does not vary between different sampling occasions.

3 Design-based properties of data from species sampling

We consider a sampling scheme adopted to select units in a natural community that induces a sampling design over $\mathfrak{D}$. Then, we introduce a one-to-one mapping Z from $\mathfrak{D}$ to $\{0, 1\}^K$ such that for each $D \in \mathfrak{D}$, $\boldsymbol{Z} = Z(D)$ is a K-vector $\boldsymbol{Z} = [Z_1, \ldots, Z_K]^t$ with $Z_l = I(l \in D)$ for each $l \in L$. The discrete random vector $\boldsymbol{Z}$ has support $\{0, 1\}^K$ and probability function:

$$p(z_m) = \theta_m, \;\; z_m \in \{0, 1\}^K, \tag{1}$$

where $z_1, \ldots, z_M$ are the $M = 2^K$ vectors of $\{0, 1\}^K$ written in a lexicographic order, such as $(0, \ldots, 0)$, $\ldots$, $(1, \ldots, 1)$, and $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_M]^t$ is an M-dimensional parameter varying in the parametric space $\Theta = \{\boldsymbol{\theta} : 0 \leq \theta_j \leq 1, \;\; \sum_{j=1}^{M} \theta_j = 1\}$. In this way, the uncertainty arising from species sampling is transferred to the parameter $\boldsymbol{\theta}$. Accordingly, P_θ and E_θ denote the design-based probability measure and expectation.

According to Equation (1), each marginal variable Z_l has support $\{0, 1\}$ and $P_\theta(Z_l = 1) = \pi_l$, so that $E_\theta(Z_l) = \pi_l$ and $E_\theta(\boldsymbol{Z}) = \boldsymbol{\pi}$, where $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]^t$ is the vector of species inclusion probabilities. For some proofs in the online supplementary material file, $\pi_{(1)}$ denotes the smallest inclusion probability and $K_{(1)}$ denotes the number of species with inclusion probability $\pi_{(1)}$.

In most cases, a natural community cannot be adequately sampled by means of only one run of the sampling scheme. Thus, we suppose that the sampling scheme is independently replicated n times (e.g. n plots or transects are randomly and independently located in the study region). The replications give rise to n independent samples of species $D_1, \ldots, D_n$ that in turn give rise to n *iid* random vectors $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ from Equation (1), which are usually gathered in the $K \times n$ random matrix $\boldsymbol{Z}_{(n)} = [\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n]$ in which the (l, i)-element is the random variable $Z_{l,i,n}$, which is equal to 1 if the species l is sampled at the ith replication and is equal to 0 otherwise. Due to the independence of replications, the distribution of $\boldsymbol{Z}_{(n)}$ continues to be determined by $\boldsymbol{\theta}$, so that P_θ and E_θ also denote the probability measure and expectation induced by the n independent replications of the sampling scheme.

Now, we use $D_{(n)} = \bigcup_{i=1}^{n} D_i$ to denote the pooled sample of species and $Q_{\text{obs},n}$ to denote the total number of sampled species. Notably, only $Q_{\text{obs},n}$ rows out of the K rows of $\boldsymbol{Z}_{(n)}$ are observable (those having at least one 1), whereas the remaining $K - Q_{\text{obs},n}$ rows of 0's cannot be observed. The K-random vector $\boldsymbol{X}_n = \sum_{i=1}^{n} \boldsymbol{Z}_i$ plays an important role in the estimation of species richness, where the marginal random variable $X_{l,n}$ yields the number of replications in which species l is sampled. While the analytical formulation of the design-based joint distribution of $\boldsymbol{X}_n$ is prohibitive to derive, due to the independence of replications each marginal variable $X_{l,n}$ has a binomial distribution with parameters n and π_l. Once again, only $Q_{\text{obs},n}$ components out of the K components of $\boldsymbol{X}_n$ are actually observable (those having positive values), whereas the remaining $K - Q_{\text{obs},n}$ components equal to 0 cannot be observed. A further synthesis of the incidence matrix is provided by the $(n + 1)$ random vector $\boldsymbol{Q}_n = [Q_{0,n}, Q_{1,n}, \ldots, Q_{n,n}]^t$, where $Q_{x,n} = \sum_{l \in L} I(X_{l,n} = x)$ is the random variable that gives

the number of species detected in x replications. Obviously, the number of unsampled species $Q_{0,n}$ is not observable, whereas the n vector $\boldsymbol{Q}_{n/0} = [Q_{1,n}, \ldots, Q_{n,n}]^t$ is a genuine statistic that can be exploited in species richness estimation.

Unfortunately, even the analytical formulation of the design-based joint distribution of $\boldsymbol{Q}_n$ is prohibitive to derive, and only the expectations of each single component can be analytically derived. Indeed, from the binomial distribution of each $X_{l,n}$, it follows that:

$$E_\theta(Q_{x,n}) = \binom{n}{x} \sum_{l \in L} \pi_l^x (1 - \pi_l)^{n-x}, \quad x = 0, 1, \ldots, n. \tag{2}$$

Equation (2) provides the unique theoretical support for the design-based estimation of species richness, which is the price that must be paid to work from a design-based perspective, without assuming models.

Notably, the literature on species sampling traditionally distinguishes two kinds of schemes for selecting units from communities: schemes that, at each replication, select agglomerates of units by means of devices such as plots, transects, nests, or traps, and schemes that, at each replication, randomly select single units (Chao, 2005; Chao & Chiu, 2016a, 2016b). In the first case, the resulting data matrices $\boldsymbol{Z}_{(n)}$ are referred to as incidence data, and they are the types of data previously considered in this section. This data format is particularly convenient in natural communities because only the presence or absence of a species is recorded at each replication, neglecting the number of individuals whose quantification may be problematic in the presence of highly abundant species, such as insects, grass species, or species exhibiting clonal reproduction (Palmer, 1990). However, the literature on species richness estimation has traditionally considered schemes in which a single individual is randomly selected at each replication with replacement. In this case, the resulting data are referred to as abundance data because the random variable $X_{l,n}$ now provides the abundance of units of species l that are present in the final sample of n individuals. Many estimators of species richness were originally proposed for abundance data and subsequently adapted to incidence data. Familiar examples are the Chao1 estimator (Chao, 1984), followed by the Chao2 estimator (Chao, 1987), and the ACE estimator (Chao & Lee, 1992), followed by the ICE estimator (Lee & Chao, 1994). However, in natural communities, in which species exhibit spatial aggregation, the random selection of single units with replacement, similar to balls from an urn, is difficult to implement (e.g. Chiu, 2022). As noted by Wang (2010), abundance data are realistic in other fields of application, such as the analysis of gene expression data. From a design-based point of view, abundance data are generated by designs of much lower entropy with respect to those generating incidence data. Indeed, in the case of abundance data, the support of the discrete random vector $\boldsymbol{Z}$ reduces to the standard base of R^K, and the possible samples are constituted by single species, so that their probabilities coincide with the species inclusion probabilities that sum to 1 and are proportional to the species abundances in the whole community. In addition, the joint inclusion probabilities are invariably zero because only one species is selected at each replication. In practice, abundance data constitute particular cases of incidence data and are collected by sampling schemes that are difficult to implement in natural communities. For these reasons, we avoid considering estimators based on abundance data.

4 Model-based view of species sampling and richness estimation

The Bernoulli product model (BPM) was originally introduced by Burnham and Overton (1978) to estimate animal population sizes from capture–recapture experiments. The analogy between capture–recapture population size estimation and species richness estimation from incidence data has been well recognized (e.g. Chao & Chiu, 2016b). In typical capture–recapture experiments, data consist of an individual-by-trapping sample matrix with rows that correspond to individuals, columns that correspond to trapping occasions and elements that correspond to either the capture (1) or noncapture (0) of individuals. Thus, an individual in capture–recapture studies corresponds to a species in species richness estimation, and a trapping occasion corresponds to a replication.

As stated in the Section 1, since the seminal paper by Burnham and Overton (1978), and without any reference to the sampling scheme adopted to select species, the matrix of incidence data is traditionally supposed to be generated by the BPM

$$\Pr(\boldsymbol{Z}_{(n)};\boldsymbol{\pi}) = \prod_{i=1}^{n}\prod_{l\in L}\pi_l^{Z_{l,i}}(1-\pi_l)^{1-Z_{l,i}} = \prod_{l\in L}\pi_l^{X_{l,n}}(1-\pi_l)^{n-X_{l,n}} \tag{3}$$

The BPM is based on three assumptions: (i) the n sampling occasions are independent; (ii) species inclusion probabilities are constant between occasions, and (iii) species selections at each sampling occasion are independent events. The BPM is presented in the literature as a general, unavoidable framework for performing species richness estimation (e.g. Chao & Colwell, 2017; Chao et al., 2014; Colwell et al., 2012). Despite its wide use, it should be noted that while Assumption (i) is ensured by independently replicating the sampling scheme (e.g. independent location of plots or transects) and Assumption (ii) is ensured by the type of community (e.g. it surely holds in plant communities as outlined in the online supplementary material, Appendix A), Assumption (iii) is unrealistic in the presence of incidence data due to the tendency of species to aggregate or to avoid each other.

Moreover, considering Equation (3) and supposing that $\pi_1, \ldots, \pi_K$ are *iid* random variables from a distribution F, Burnham and Overton (1978) demonstrated that $\boldsymbol{Q}_n$ has a multinomial distribution with parameters K and $\boldsymbol{g}_F = [g_F(0), g_F(1)\ldots, g_F(n)]^t$, where:

$$g_F(x) = \binom{n}{x}\int_0^1 \pi^x(1-\pi)^{n-x}\,dF(\pi) \quad \text{for } x = 0, 1, \ldots, n \tag{4}$$

so that the vector $\boldsymbol{Q}_{n/0} = [Q_{n,1}, \ldots, Q_{n,n}]^t$ is the minimal sufficient statistic irrespective of F, which allows for the estimation of species richness by exploiting the best synthesis of the incidence data.

In this framework, estimation can be performed on the basis that $Q_{\text{obs},n}$ is binomial with parameters K and $1-g_F(0)$. Accordingly, a pseudo-maximum likelihood estimator (Gong & Samaniego, 1981) for K can be obtained by $\widehat{K}_{\text{PMLE},n} = Q_{\text{obs},n}(1+\widehat{\varphi})$, where $\varphi = g_F(0)/\{1-g_F(0)\}$ is the odd of missing a species. In practice, the problem of estimating K reduces to that of estimating φ in a parametric way, by supposing a parametric model for F (e.g. Chao, 2005, pp. 7908–7909 and many references therein), or in a nonparametric way, leaving F unspecified (e.g. Norris & Pollock, 1998; Wang, 2010; Wang & Lindsay, 2005). Chao (2005, p. 7909) criticizes the parametric approach, emphasizing the difficulty in specifying the mixing distribution F. The author noted that 'Two models with different mixing distributions may fit the data equally well, but they yield widely different estimates'. In addition, Mao and Lindsay (2007) showed that if the support of the mixing distribution F is arbitrarily near 0 (i.e. if it does not exist a π_0 such that $\pi > \pi_0$), no unbiased estimator for K and no genuine two-sided confidence interval exist. Even if this result is originally achieved by supposing a Poisson distribution for $X_{l,n}$s, it also holds in the binomial case. Therefore, considering that the distributions of species inclusion probabilities in natural communities are in most cases highly spiked at 0 (e.g. Figures B2 and B3 in the online supplementary material, Appendix B), it seems quite unrealistic to presume a lower threshold π_0 in real cases.

The above concerns have led to the wide use of so-called nonparametric approaches, which avoid the maximum likelihood criterion and for which the species richness is simply estimated by suitable functions of the components of the minimal sufficient statistics $\boldsymbol{Q}_{n/0}$. In addition, due to the convergence of $\boldsymbol{Q}_{n/0}$ to normality that holds under the BPM for large K, variance estimation and the construction of confidence intervals are performed by standard methods.

Most nonparametric estimators have long been automated in the free software SPADE (Chao et al., 2003). Due to the possibility of computing estimates, variance estimates and confidence intervals by means of a well-known, easily accessible software—continuously improved in an R-based version referred to as SpadeR (Chao et al., 2015; Chao & Shen, 2010)—SPADE software is widely applied by botanists, foresters, and zoologists. For these reasons, we focus our attention on the nonparametric estimators implemented in SPADE.

5 Design-based view of nonparametric estimators

In this section, we show that all of the nonparametric estimators automated in SPADE have a design-based nature because they can be derived simply from the independence of the replications that in turn determines the basic Equation (2) while disregarding the BPM. From the same equation, we also derive the design-based properties of these estimators. In addition, we consider two real communities of trees to check the estimator performance in real cases: the BCI community, which is a portion of a tropical forest on Barro Colorado Island (Panama) within a rectangular stand of 50 ha composed of $N = 221{,}758$ trees and $K = 302$ species, and the HF community, which is a portion of a temperate forest in Massachusetts within a rectangular stand of 35 ha composed of $N = 77{,}536$ trees and $K = 55$ species (see the online supplementary material, Appendix B for details). We sampled both communities by means of $n = 10$, 20, 50 circular plots of 10 m radius, which were randomly and independently located in the stands, and we determined the species inclusion probabilities of the species encountered during plot sampling (see the online supplementary material, Appendix A). As it is customary in natural communities, the resulting sets of species inclusion probabilities give rise to heavy-tailed distributions clumped near 0, with frequencies that slowly decrease toward 1 (see Figures B3 and B4 in the online supplementary material, Appendix B).

According to Equation (2), the design-based expectation of $Q_{\text{obs},n}$ is given by

$$E_\theta(Q_{\text{obs},n}) = K - \sum_{l \in L} (1 - \pi_l)^n. \tag{5}$$

Equation (5) is well known as the species accumulation curve, as it describes the increase in the expected number of sampled species as the number of replications increases (Colwell & Coddington, 1994). For finite n, $Q_{\text{obs},n}$ is a design-biased estimator of K with a negative bias that approaches 0 as n increases. However, the bias is relevant, and it slowly approaches 0 when there are many species with small inclusion probabilities. This condition is apparent in the BCI and HF communities (see Figure 1), where the expectations of $Q_{\text{obs},n}$ are plotted for $n = 10, 20, 50$ (black line) and from the percentages of the relative bias reported in Table 1. In both communities, the species accumulation curves slowly approach the horizontal black lines representing the true species richness, with approximately 30% of the species remaining undetected notwithstanding the sampling effort of 50 plots, which means there was approximately one plot per ha, field work that is hardly sustainable in real surveys.

In this framework, nonparametric estimators can be viewed as criteria to reduce the bias of $Q_{\text{obs},n}$ to achieve approximately unbiased estimators of K. Based on the criteria adopted, we grouped the nonparametric estimators automated in SPADE into jackknife estimators, CHAO2-type estimators, and ICE-type estimators.

5.1 Jackknife estimators

As jackknife techniques constitute very common tools for reducing bias, Heltshe and Forrester (1983) proposed estimating K via a first-order jackknife on $Q_{\text{obs},n}$. Deleting one replication at a time, the first-order jackknife estimator of K is given by

$$\widehat{K}_{\text{jack},n} = Q_{\text{obs},n} + \frac{n-1}{n} Q_{1,n}. \tag{6}$$

Based on Equation (2), the design-based expectation of Equation (6) is given by

$$E_\theta(\widehat{K}_{\text{jack},n}) = K - \sum_{l \in L} (1 - \pi_l)^{n-1}(1 - n\pi_l). \tag{7}$$

Subsequently, Smith and Van Belle (1984) proposed the use of a second-order jackknife. Deleting a pair of replications at a time, the second-order jackknife estimator of K is given by

$$\widehat{K}_{\text{jack2},n} = Q_{\text{obs},n} + \frac{2n-3}{n} Q_{1,n} - \frac{(n-2)^2}{n(n-1)} Q_{2,n} \tag{8}$$

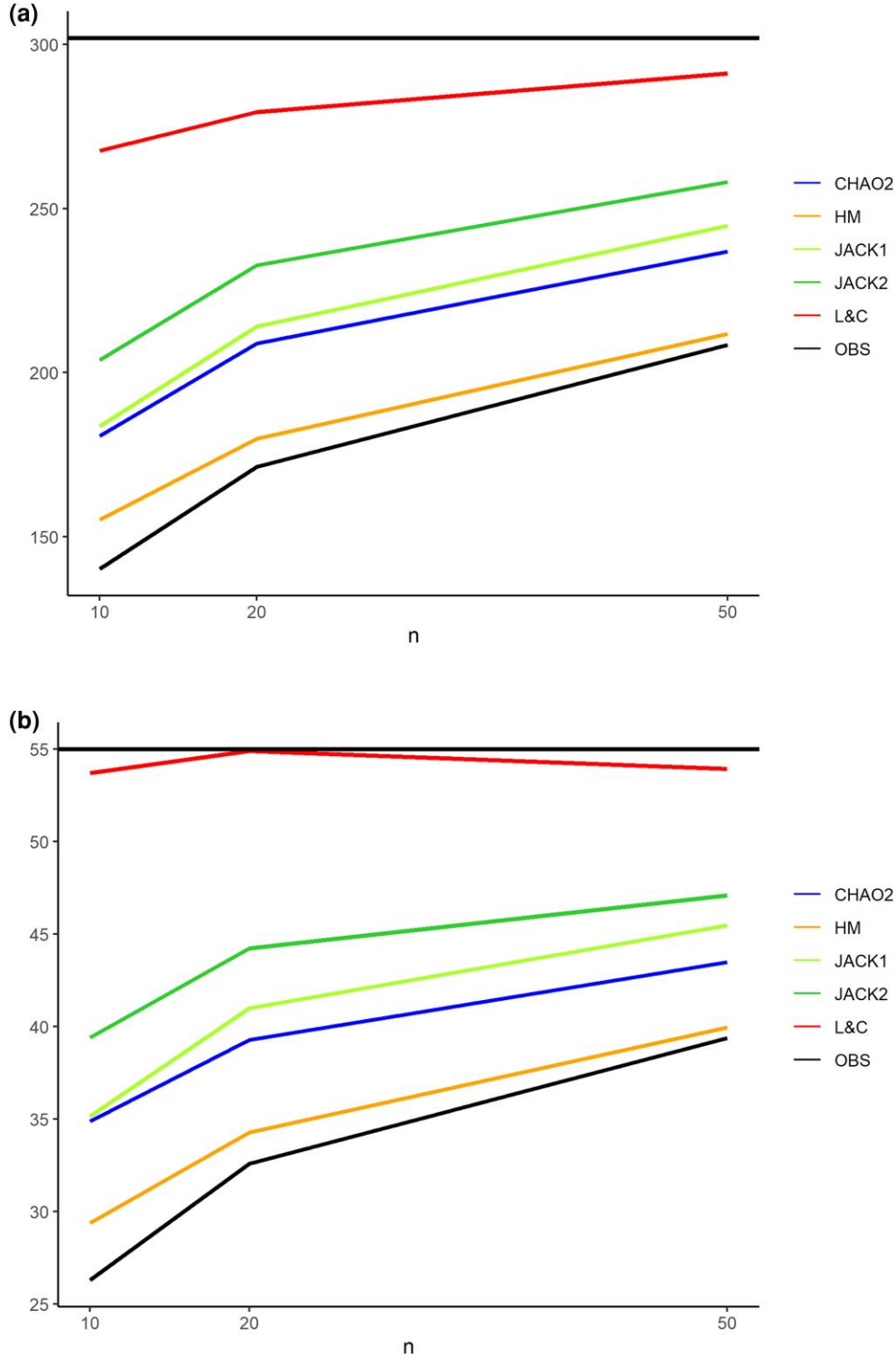

Figure 1. Design-based expectations of observed species (OBS, first line from the bottom) and jackknife estimators (JACK1, JACK2; fourth and fifth lines from the bottom) and values of the Chao lower bounds (CHAO2, third line from the bottom) and of the Lee and Chao (L&C, sixth line from the bottom) and homogeneous model (HM, second line from the bottom) approximators achieved with $n = 10, 20, 50$ circular plots of 10 m radius, which were randomly and independently located in the stands of the Barro Colorado Island (a) and Harvard Forest (b) communities compared with their true species richness represented by the black horizontal lines.

Table 1. Percentages of the relative bias of observed species (OBS) and jackknife estimators (JACK1, JACK2) and percentages of the approximation errors of the Chao lower bounds (CHAO2) and the Lee and Chao (L&C) and homogeneous model (HM) approximators achieved with $n = 10, 20, 50$ circular plots of 10 m radius, which were randomly and independently located in the stands of the Barro Colorado Island (BCI) and Harvard Forest (HF) communities

STAND	n	OBS	JACK1	JACK2	CHAO2	L&C	HM
BCI	10	−54	−39	−33	−40	−11	−49
	20	−43	−29	−23	−31	−8	−40
	50	−31	−19	−15	−22	−4	−30
HF	10	−52	−36	−28	−37	−2	−47
	20	−41	−25	−19	−29	0	−38
	50	−28	−17	−14	−21	−2	−27

Based on Equation (2), the design-based expectation of Equation (8) is given by

$$E_\theta(\widehat{K}_{\text{jack2},n}) = K - \sum_{l \in L} (1-\pi_l)^{n-2}\left(\frac{n^2\pi_l^2}{2} - 2n\pi_l + \pi_l + 1\right). \tag{9}$$

D'Alessandro and Fattorini (2002) provided theoretical justifications for the design-based inadequacy of jackknife estimators to reduce the bias of $Q_{\text{obs},n}$, on the basis of the failure of the $Q_{\text{obs},n}$ statistic to satisfy some standard conditions required by the jackknife to be effective (see, e.g. Shao & Tu, 1995, Section 2.4.2). This situation is apparent in the BCI and HF communities (see Figure 1), where the expectations of the jackknife estimators are plotted for $n = 10, 20, 50$ (light and dark green lines), and from Table 1, which reports the percentages of the relative bias. Despite a bias reduction, the convergence of expectations toward the black horizontal lines of true species richness is similar to that provided by $Q_{\text{obs},n}$, and the bias is approximately −15%, notwithstanding the great sampling effort of 50 plots.

5.2 Chao2-type estimators

These estimators are based on the Chao inequality, which was originally derived in capture–recapture experiments (Chao, 1987). The same inequality can be derived from Equation (2), exploiting the Cauchy–Schwarz inequality (see, e.g. Chao & Colwell, 2017), so that it also holds in a design-based framework. From Equation (2), it follows that:

$$K_{\text{Chao2},n} = E_\theta(Q_{\text{obs},n}) + \frac{n-1}{2n}\frac{E_\theta(Q_{1,n})^2}{E_\theta(Q_{2,n})} \tag{10}$$

is a lower bound for K. Chao (1987) proposes to estimate $K_{\text{Chao2},n}$ instead of K, giving rise to the following estimators, which were implemented in SPADE:

$$\widehat{K}_{\text{Chao2},n} = Q_{\text{obs},n} + \frac{n-1}{2n}\frac{Q_{1,n}^2}{Q_{2,n}} \tag{11}$$

$$\widehat{K}_{\text{Chao2}-bc,n} = Q_{\text{obs},n} + \frac{n-1}{2n}\frac{Q_{1,n}(Q_{1,n}-1)}{Q_{2,n}+1} \tag{12}$$

to be used instead of $\widehat{K}_{\text{Chao2},n}$ when $Q_{2,n} = 0$ (Chao et al., 2015) and

$$\widehat{K}_{\text{iChao2},n} = \widehat{K}_{\text{Chao2},n} + \frac{(n-3)Q_{3,n}}{4nQ_{4,n}}\max\left[0, Q_{1,n} - \frac{(n-3)Q_{2,n}Q_{3,n}}{2(n-1)Q_{4,n}}\right] \tag{13}$$

which exploits the additional information provided by $Q_{3,n}$ and $Q_{4,n}$ to obtain an 'improved' reduced-bias estimator (Chiu et al., 2014).

Even if $K_{\text{Chao2},n}$ converges to K as n increases (online supplementary material, Appendix C), and despite the emphasis given to the inequality (e.g. Chao & Colwell, 2017), the approximation to K is accurate in unrealistic situations in which species inclusion probabilities are quite homogeneous. Otherwise, when rare and common species are present, resulting in high variability in the odds of missing species, $K_{\text{Chao2},n}$ approaches K very slowly (online supplementary material, Appendix C). This situation is apparent in the BCI and HF communities (see Figure 1), in which the lower bounds are plotted for $n = 10, 20, 50$ (blue line), and from Table 1, which reports the percentage approximation errors. In both communities, $K_{\text{Chao2},n}$ converges very slowly to the black horizontal lines representing the true species richness, with approximation errors of approximately 20% even with 50 plots. In these situations, the CHAO2-type estimators, estimating these bounds, are likely to heavily underestimate the actual richness.

5.3 ICE-type estimators

These estimators are based on the following approximation to K, which was originally derived in capture–recapture experiments by Lee and Chao (1994):

$$K_{\text{L\&C},n} = \frac{E_\theta(Q_{\text{obs},n})}{E_\theta(C_n)} + \frac{E_\theta(Q_{1,n})}{E_\theta(C_n)}\gamma^2, \tag{14}$$

where C_n is the sample coverage and γ^2 is the squared coefficient of variation of the species inclusion probabilities. In the case of incidence data, C_n and γ^2 are respectively given by

$$C_n = \frac{\sum_{l\in L}\pi_l I(X_{l,n} > 0)}{\sum_{l\in L}\pi_l}$$

and

$$\gamma^2 = K\frac{\sum_{l\in L}\pi_l^2}{\left(\sum_{l\in L}\pi_l\right)^2} - 1 \tag{15}$$

while from Equation (2) C_n has design-based expectation

$$E_\theta(C_n) = 1 - \frac{\sum_{l\in L}\pi_l(1-\pi_l)^n}{\sum_{l\in L}\pi_l} \tag{16}$$

Based on Equation (14), Lee and Chao (1994) propose estimating $K_{\text{L\&C},n}$ instead of K. In the online supplementary material, Appendix D, we prove that the approximator (14), henceforth referred to as the L&C approximator, continues to hold in a design-based framework, so that the estimation criterion proposed by the authors continues to hold in this framework.

If the variability of species inclusion probabilities is neglected, i.e. $\gamma = 0$, the approximator defined in Equation (14) reduces to the so-called homogeneous model (HM) approximator

$$K_{\text{HM},n} = \frac{E_\theta(Q_{\text{obs},n})}{E_\theta(C_n)}. \tag{17}$$

Both approximators converge to K as n increases (online supplementary material, Appendix E), but while $K_{\text{HM},n}$ has an approximation error asymptotically equivalent to the bias provided by $Q_{\text{obs},n}$ (online supplementary material, Appendix E, Equation E2) so that any estimator based on it is destined to fail, the idea of considering the heterogeneity of species inclusion probabilities by $K_{\text{L\&C},n}$ is effective. This situation is apparent from Figure 1, where the approximators $K_{\text{L\&C},n}$ and $K_{\text{HM},n}$ in the BCI and HF communities are plotted for $n = 10, 20, 50$ (red and orange lines), and from Table 1, which reports the percent approximation errors of the two approximators.

In both communities, the values of $K_{L\&C,n}$ are the closest to the black horizontal lines representing the true species richness, and the relative errors are close to 10% or smaller, even for sampling efforts of 10 and 20 plots. The effectiveness of the L&C approximator with respect to the Chao2 approximator is apparent: while $K_{Chao2,n}$ is a lower bound achieved from the Cauchy–Schwarz inequality so that it always under-evaluates K being accurate only in unrealistic situations in which the inclusion probabilities are quite homogeneous, $K_{L\&C,n}$ is equivalent to K as n increases and coincides with K up to an infinitesimal of higher order than the variance of the inclusion probabilities (Supplementary Material Appendix D, Equation D4).

Notwithstanding the theoretical appeal of the L&C approximator, the SPADE implementation of the estimators based on $K_{L\&C,n}$ is questionable. This implementation stems from the idea of Chao et al. (1993), which, when estimating the number of different bugs in software from abundance data, suggest choosing a cutoff value x_0 so that bugs are partitioned into those detected more than x_0 times, referred to as frequent bugs, and those detected no more than x_0 times, referred to as infrequent bugs. Since frequent bugs are likely to be detected, the authors suggest ignoring them and applying the Chao and Lee (1992) estimator only to estimate infrequent bugs. The number of bugs detected more than x_0 times is then added to the resulting estimate. To our knowledge, the performance of this proposal has been checked only once: in the original paper by Chao et al. (1993), which describe their simulation study in which the abundance data were generated from several artificial sets of detection probabilities. Nevertheless, the splitting procedure had already been implemented at the time of the first version of SPADE and had also been extended to incidence data (Chao et al., 2003). In practice, the $K_{L\&C,n}$ and $K_{HM,n}$ approximators (Equations (14) and (17)) are exploited to estimate the number of infrequent species with a default value $x_0 = 10$, giving rise to the ICE, ICE1, and HM estimators. For brevity, we do not report the huge expressions of the ICE estimators, which are reported in the online supplementary material, Appendix F. Rather, we outline that the proposal of splitting natural communities into frequent and infrequent species, depending on the random vector $\boldsymbol{Q}_x$, is likely to involve an additional source of uncertainty over that involved by the sole sampling scheme. This issue, together with the unsuitable use of the HM estimator as an initial estimate of K, which is necessary in the estimation of γ^2, is likely to deteriorate the performance of ICE and ICE1.

5.4 Variance estimation

While the nonparametric estimators automated in SPADE, i.e. jackknife estimators, CHAO2-type estimators and ICE-type estimators, hold in a design-based scenario due to Equation (2) and irrespective of the BPM, their variances and the subsequent variance estimators adopted in the literature and implemented in SPADE are strictly of model-based nature being invariably based on the asymptotic normality of $\boldsymbol{Q}_n$ derived from the BPM. Therefore, we do not consider the variance estimators adopted in SPADE because they are completely outside the design-based perspective. Moreover, given the presence of a large negative bias that, as theoretically argued in this section, is likely to affect nonparametric estimators in real situations, the estimation of variance seems to be a quite useless task. Rather, reliable evaluations of the actual precision should be based on mean squared error estimators.

6 Simulation studies

We empirically checked the performance of the nonparametric estimators adopted in SPADE by conducting a design-based simulation study performed on the BCI and HF communities. As in theoretical investigations, for both communities we considered $n = 10, 20, 50$ plots of 10 m radius, and we conducted $R = 10{,}000$ simulation runs. At the rth run, we randomly and independently locate n plots recording the species contained within them, achieving an incidence matrix of n columns and as many rows as the number of sampled species. The matrix was subsequently passed as input in the R-based version of SPADE (Chao et al., 2015).

Based on the Monte Carlo distributions achieved from the R runs, for both communities, for each estimator and for each n, we empirically computed the expectation of estimators E_n and their mean squared errors MSE_n, from which the values of the relative bias $\mathrm{RB}_n = (E_n - K)/K$, the standard errors $\mathrm{SE}_n = \sqrt{\mathrm{MSE}_n - E_n^2}$ and the relative root mean squared errors $\mathrm{RRMSE}_n = \sqrt{\mathrm{MSE}_n}/K$ were derived. Regarding the performance of the standard error estimators and the

Table 2. Monte Carlo percentages of the relative bias (RB), relative root mean squared errors (RRMSE), coverage of the 0.95 confidence intervals (C95) and their expected relative lengths (ERL), and ratio of the expectations of the standard error estimators to the true values (RAT) for the nonparametric estimators applied to Barro Colorado Island (BCI) and Harvard Forest (HF) communities

		BCI					HF	
ESTIMATOR	***n***	**RB**	**RRMSE**	**C95**	**ERL**	**RAT**	**RB**	**RRMSE**
First-order Jackknife	10	−39	39	0	0.12	71	−36	38
	20	−29	29	0	0.12	81	−25	27
	50	−19	19	0	0.11	87	−17	19
Second-order Jackknife	10	−32	33	0	0.19	85	−28	32
	20	−23	24	4	0.20	92	−19	24
	50	−14	15	34	0.19	95	−14	18
Chao2	10	−39	40	2	0.22	86	−29	39
	20	−30	31	7	0.21	89	−22	31
	50	−21	21	16	0.18	89	−17	23
Chao2-bc	10	−40	41	1	0.21	85	−36	40
	20	−31	31	5	0.20	87	−28	32
	50	−21	22	11	0.17	88	−21	23
iChao2	10	−37	37	1	0.14	51	−26	39
	20	−27	28	4	0.14	52	−20	31
	50	−19	20	9	0.12	53	−15	23
ICE	10	−38	38	0	0.18	84	−30	34
	20	−31	31	0	0.15	86	−25	28
	50	−22	23	0	0.12	85	−21	22
ICE-1	10	−32	33	12	0.28	94	−17	33
	20	−27	28	7	0.22	95	−19	26
	50	−20	21	7	0.16	91	−19	21
Homogeneous model	10	−49	49	0	0.07	49	−46	47
	20	−39	39	0	0.06	53	−35	36
	50	−27	28	0	0.06	57	−25	26

actual coverage of the confidence intervals, in the case of the HF community the software repeatedly crashed when computing the standard error estimates of the ICE-type estimators. Therefore, for the HF community, we report only the RB and RRMSE values. On the other hand, in the case of the BCI community, for each estimator and for each n, we empirically achieved the expectation of the standard error estimators ESE_n, the coverage of the 0.95 confidence intervals $\mathrm{C95}_n$ and their expected relative length ERL, i.e. their expected length divided by K. Then, the ratios $\mathrm{RAT}_n = \mathrm{ESE}_n/\mathrm{SE}_n$ were derived to check the tendency of these model-based estimators to overestimate/underestimate the true standard errors. The simulation results are reported in Table 2.

The results in Table 2 fully confirm the theoretical concerns outlined in Section 5 regarding the nonparametric estimators, as it is apparent from the Monte Carlo expectations of the eight estimators. The expectations of the resampling estimators are very similar to the theoretical expressions, confirming the reliability of the simulation. The expectations of the CHAO2-type estimators follow the trend of the lower bounds they are estimating. The estimator based on HM approximation is the worst and the ICE-type estimators greatly underestimate the L&C approximator due to the unsuitable implementation discussed in Section 5.3.

In general, the most serious drawback is the massive presence of negative bias induced by the loss of rare species. For the sampling effort of 10–20 plots, the negative bias ranges from 20% to 40%,

and it remains over 14% even with 50 plots. In addition, for the BCI community, the model-based estimators of standard errors show underestimations that range from 50% to 5% and that shorten the confidence intervals. The short length of the confidence intervals is apparent when comparing the ERL values in Table 2 with 4 times the relative sampling errors that, in the normal case, approximately represent the ERL necessary to achieve confidence intervals with a coverage of 95%.

The effect of large bias, which incorrectly centres the intervals, combined with the underestimation of standard errors, which reduces the interval lengths, leads to disastrous results for confidence intervals with risible values of coverage. Being funny, we should speak of distrust intervals rather than confidence intervals.

Similar results were achieved in a simulation study by Chiu (2022), which, to our knowledge, constitutes a unique investigation of design-based nature performed on a real community in which incidence data were generated from the community via the sampling scheme adopted in the simulation. This approach contrasts with the plethora of simulations performed in the literature on richness estimation, most of which generated data from artificial sets of inclusion probabilities in accordance with the BPM. Analogous to our study, the author generated incidence data via quadrat sampling (online supplementary material, Appendix A) performed on the BCI community of $K = 298$ species censused in 1985. For a sampling effort of $n = 40$ quadrats of side 20 m, i.e. for an effort comparable to $n = 50$ plots of 10 m radius, the main concern remains the negative bias, which is invariably greater than 14% for both the nonparametric and pseudo-maximum likelihood estimators and the subsequent risible coverage of the confidence intervals.

Notably, in addition to SPADE, the SPECIES package by Wang (2011) has been widely adopted, as it provides simple R functions to compute species richness estimates from some pseudo-maximum likelihood and nonparametric methods. As the software is designed to work with abundance data, it should not be used with incidence data collected from natural communities. However, because, in the end, all of the estimators considered in SPECIES are functions of $\mathbf{Q}_{n/0}$, they can also be attempted with incidence data. Therefore, we have checked the performance of these estimators by employing the same simulation study that checked the performance of the estimators automated in SPADE. The simulation results are available to the authors but are not reported because they do not show any improvement with respect to the results of Table 2.

7 Data integration proposal

According to the theoretical and empirical investigations described in Sections 5 and 6, it is our impression that the literature on species richness estimators and the related software do not provide naturalists with reliable methods for achieving point and interval estimates from incidence data collected in natural communities via sustainable sampling efforts. To address this issue, an alternative solution is necessary.

7.1 The Lee and Chao estimator

Based on the appealing result from the L&C approximator in Equation (14), we attempted a more effective estimation of this approximator, avoiding the partition between frequent and infrequent species adopted in SPADE. For this purpose, we exploited only the estimator originally proposed by Lee and Chao (1994, Equation 3.18):

$$\widehat{K}_{\mathrm{L\&C},n} = \frac{Q_{\mathrm{obs},n}}{\widehat{C}_n} + \frac{Q_{1,n}}{\widehat{C}_n}\widehat{\gamma}_n^2 \tag{18}$$

henceforth referred to as the L&C estimator, where, quoting again from the same paper (Equation 3.20):

$$\widehat{C}_n = 1 - \frac{Q_{1,n}}{\sum_{x=1}^{n} xQ_{x,n}} \tag{19}$$

Regarding Equation (19), we prove that it is a design-unbiased estimator of $E_\theta(\widehat{C}_n)$ up to the first order of approximation (online supplementary material, Appendix G). Regarding the estimation

Table 3. Monte Carlo percentages of the relative bias (RB) and relative root mean squared errors (RRMSE) for the Lee and Chao estimator applied to Barro Colorado Island (BCI) and Harvard Forest (HF) communities

	BCI		HF	
n	RB	RRMSE	RB	RRMSE
10	−28	29	−16	30
20	−18	19	−7	23
50	−9	11	−6	16

of γ^2 in Equation (18), we slightly differ from Lee and Chao (1994, Equation 3.22), as we adopt the estimator:

$$\widehat{\gamma}_n^2 = \max\left\{\widehat{K}_{0,n}\frac{n\sum_{x=1}^{n}x(x-1)Q_{x,n}}{(n-1)\left(\sum_{x=1}^{n}xQ_{x,n}\right)^2}-1,\ \ 0\right\}, \tag{20}$$

where $\widehat{K}_{0,n}$ is an initial estimate of K. Up to the first-order approximation, we prove that the fraction in Equation (20) is a design-unbiased estimator of the fraction in Equation (15) (online supplementary material, Appendix H). As an initial estimate of K, we used the second-order jackknife estimator, which has proven to be the best performer in our simulation study.

We then checked the performance of the L&C estimator by using the same simulation study described in Section 6. Table 3 reports the values of the relative bias and relative root mean squared errors, showing that $\widehat{K}_{\mathrm{L\&C},n}$ provides the best performance with respect to the nonparametric estimators considered in SPADE, even if the bias remains relevant for sampling efforts of 10 plots and, in the BCI community, also 20 plots.

7.2 Data integration

In accordance with a novel drift of survey sampling, usually referred to as data integration (e.g. Kim & Tam, 2021), we attempted to further reduce the bias of the L&C estimator by joining the information acquired in purposive surveys performed to detect the rarest species with the information gathered by the probabilistic surveys, as proposed by Chiarucci et al. (2018). In this article, L_0 denotes the community of the K_0 rare species detected by means of one or several purposive surveys. We then adopted the L&C estimator to estimate the number of species in the residual community $L - L_0$ of species not included in the purposive list. In practice, we propose as estimator of species richness the data integration estimator

$$\widehat{K}_{\mathrm{DI},n} = K_0 + \widehat{K}_{\mathrm{L\&C},n}^{(L-L_0)}, \tag{21}$$

where $\widehat{K}_{\mathrm{L\&C},n}^{(L-L_0)}$ is the L&C estimator in Equation (18) applied to the community $L - L_0$ (see online supplementary material, Appendix I for details on the estimator defined in Equation (21)). In practice, according to Equation (21), we estimate the approximator to K by means of

$$K_{\mathrm{DI},n} = K_0 + K_{\mathrm{L\&C},n}^{(L-L_0)} \tag{22}$$

which is achieved from data integration by the sum of K_0 plus the L&C approximator in Equation (14) applied to the community $L - L_0$ (see online supplementary material, Appendix J for details on the approximator in Equation (22)).

From a theoretical point of view, the improvement entailed by data integration in estimating species richness is easy to see. First, the knowledge that K_0 rare species—hardly detectable by probabilistic sampling—are present in the community obviously reduces the missing of species that constitutes the main reason for the failure of species richness estimators. This fact is clearly apparent from the

species accumulation curves obtained via data integration, which are invariably above those achieved from the whole community, with faster convergence rates to K (online supplementary material, Appendix J, Equation J4). In addition, we prove in online supplementary material, Appendix J, Equation J8 that the approximator defined in Equation (22), which was achieved by data integration, provides more accurate approximations to K than the L&C approximator in Equation (14). Therefore, the data integration estimator in Equation (21) that estimates the approximator in Equation (22) is likely to perform better than the L&C estimator in Equation (18), which instead estimates the approximator in Equation (14).

However, apart from these theoretical considerations, from a practical point of view, it is apparent that the effectiveness of the estimator in Equation (21) strictly depends on the ability to detect the rarest species via opportunistic surveys. To this end, we once again quote from Palmer et al. (2002), which consider purposive surveys to be the unique effective methods for detecting most rare species otherwise missed by probabilistic sampling. In addition, Alessi et al. (2023) emphasize the utility of combining probabilistic and opportunistic sampling, pointing out that 'The combination of probabilistic and preferential sampling approaches may detect different facets of plant community diversity, revealing both common and rare species distributions and abundances'. Based on these considerations, the use of data integration seems appealing from both theoretical and practical points of view.

7.3 Design-based estimation of precision

With respect to the design-based estimation of precision, in contrast to the model-based case, we have no asymptotic result to exploit. We only have available the independence between the columns of the incidence matrix which is not assumed but is ensured by the replications of the sampling scheme. We then estimated the mean squared error of $\widehat{K}_{\mathrm{DI},n}$ via a bootstrap procedure performed by B independent resamplings of $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$. At each bth resampling, we randomly and independently selected the n columns $\boldsymbol{Z}^*_{1,b}, \ldots, \boldsymbol{Z}^*_{n,b}$ that constitute the bth bootstrap incidence matrix $\boldsymbol{Z}^*_{(n),b}$ from which the bth bootstrap estimate $\widehat{K}^*_{\mathrm{DI},n,b}$ was computed via Equation (21). The bootstrap estimator of the mean squared error is given by

$$\widehat{\mathrm{mse}}^*_{n,B} = \frac{1}{B}\sum_{b=1}^{B}(\widehat{K}^*_{\mathrm{DI},n,b} - \widehat{K}_{\mathrm{DI},n})^2 \tag{23}$$

from which we obtained the bootstrap estimate of the relative root mean squared error $\widehat{\mathrm{rrmse}}^*_{n,B} = \sqrt{\widehat{\mathrm{mse}}^*_{n,B}}/\widehat{K}_{\mathrm{DI},n}$. Finally, from the ordered sequence of the B bootstrap estimates, we derived the bootstrap 0.95 confidence interval from the quantiles of order 0.025 and 0.975, $\widehat{K}^*_{0.025}$ and $\widehat{K}^*_{0.975}$. The bootstrap procedure can be easily automated in R (codes are available from the authors). For single case studies, $B = 10{,}000$ resamplings can be performed in approximately 2 min using a standard workstation (Intel(R) Core(TM) i7-8750H with 6 cores, 12 logical processes, and 16 GB of RAM). Obviously, the procedure is more cumbersome in simulation studies when the resampling procedure must be performed for each selected sample. In these cases, $B = 1{,}000$ resamplings were adopted.

7.4 Simulation studies

We checked the performance of the proposed strategy via the same simulation study described in Section 6, supposing an accurate purposive list L_0 containing all of the species with inclusion probabilities smaller than the third decile (Tables B1 and B2 in the online supplementary material, Appendix B). In the case of the BCI community, the list contains $K_0 = 91$ species, whereas in the case of the HF community, the list contains $K_0 = 17$ species. At the rth simulation run, the bootstrap relative root mean squared error estimate and the bootstrap confidence intervals were obtained based on $B = 1{,}000$ resamplings, from which we derived the empirical expectations of the bootstrap relative root mean squared error estimator ($\mathrm{ERRMSEE}_{n,B}$) and the coverage of the 0.95 bootstrap confidence intervals ($C95_{n,B}$).

The simulation results in Table 4 show the effectiveness of the data integration estimator in terms of reducing bias. For the sustainable sampling effort of 10–20 plots, the bias is invariably

Table 4. Monte Carlo percentages of the relative bias (RB) and relative root mean squared errors (RRMSE) of the data integration estimator applied to the Barro Colorado Island (BCI) and Harvard Forest (HF) communities, and percentages of the expectations of the bootstrap relative root mean squared error estimator (EBRRMSEE), coverage of the 0.95 bootstrap confidence intervals (CB95), coverage of the 0.95 bootstrap symmetrized confidence intervals (CB95SYM) and their expected relative lengths (ERLSYM = expected length/K) based on $B = 1{,}000$ resamplings. The simulations were based on 10,000 runs

STAND	n	RB	RRMSE	EBRRMSEE	CB95	CB95SYM	ERLSYM
BCI	10	−2	7	20	19	100	64
	20	4	7	16	52	100	51
	50	5	7	8	98	100	28
HF	10	−8	20	24	59	80	76
	20	−4	16	19	72	86	64
	50	−5	11	11	75	85	40

Table 5. Results of data integration for species richness estimation performed in four protected areas of Tuscany (Central Italy), reporting size, plot type, number of replicated plots (n), number of species observed in the sample (SAMPLED), number of species in the purposive list (LISTED), total number of species sampled and/or listed (POOLED), species richness estimate (ESTIMATE), percentages of the bootstrap relative root mean squared error estimate (BRMSEE) based on $B = 10{,}000$ resamplings and limits of 0.95 bootstrap symmetrized confidence intervals (B95SYM)

Study region	Size (km^2)	Plot type	n	SAMPLED	LISTED	POOLED	ESTIMATE	BRRMSEE (%)	B95SYM
Maremma Regional Park	94	Quadrat plots 10 m side	90	608	846	962	1390	25	966–1814
Poggio all'Olmo Nature Reserve	4	Circular plots of 50 m^2	100	411	224	512	678	19	512–851
Duna Feniglia Nature Reserve	4.74	Quadrat plots of 10 m side	103	202	55	206	365	35	206–536
Foreste Casentinesi National Park	368.46	Circular plots of 13 m radius	87	325	370	454	541	14	454–643

smaller than −10%. With respect to the bootstrap performance, while the precision was conservatively estimated, the coverages of bootstrap intervals are excessively smaller than the nominal level. Therefore, we heuristically enlarged the intervals via a symmetrizing procedure of type:

$$\widehat{K}_{\mathrm{DI},n} \pm \max(\widehat{K}_{\mathrm{DI},n} - \widehat{K}^*_{0.025}, \widehat{K}^*_{0.975} - \widehat{K}_{\mathrm{DI},n}).$$

In addition, to avoid less than obvious results, if the lower bound of the symmetrized interval is smaller than the number of species $Q_{\mathrm{pool},n}$ in the pooled set $D_{(n)} \cup L_0$, i.e. the minimum number of species surely existing in the study region, we set the lower bound equal to $Q_{\mathrm{pool},n}$. As it is apparent in Table 4, the coverages become conservative but at the cost of an increase in the relative interval lengths that however, as n increases, become comparable with those expected in the normal case (approximately 4 times the relative sampling error).

8 Applications to communities of vascular plants in Central Italy

We applied our data integration method to estimate species richness in four communities of vascular plants located in nature reserves and parks in Tuscany (Central Italy). The study regions were the Maremma Regional Park, Poggio all'Olmo Nature Reserve, Duna Feniglia Nature Reserve and Foreste Casentinesi National Park. The necessity of applying an alternative estimation criterion to take advantage of the additional information furnished by the purposive lists available for those study regions is apparent from the awkward results provided by the SPADE software, which are not reported here for brevity but are completely detailed in the online supplementary material, Appendix K. In the case of the Maremma Regional Park and the Foreste Casentinesi National Park, the resulting estimates are invariably smaller than $Q_{\text{pool},n}$. In the case of the Poggio all'Olmo Nature Reserve, only the second-order jackknife provides an estimate greater than $Q_{\text{pool},n}$, while the CHAO2-type estimates, notwithstanding their widespread use in botanical applications (e.g. Cazzolla Gatti et al., 2022), are invariably smaller than $Q_{\text{pool},n}$. The Duna Feniglia Nature Reserve is the sole case in which the achieved estimates provide admissible results. On the other hand, the estimates provided by the data integration estimator in Equation (21) are invariably greater than the minimum number of species present in the study regions with reliable confidence intervals provided by the symmetrized bootstrap procedure based on $B = 10{,}000$ resamplings (Table 5).

9 Concluding remarks

Due to the theoretical drawbacks of pseudo-maximum likelihood estimators, nonparametric species richness estimators based on the sufficient statistic $\boldsymbol{Q}_{n/0}$ (where the sufficiency is with respect to BPM) are widely used, and most of them are automated in the SPADE software. These methods have produced promising results in several simulation studies of a model-based nature, in which abundance data were generated from artificial sets of inclusion probabilities (e.g. Chao & Lee, 1992; Chiu et al., 2014).

On the other hand, when the nonparametric estimators from incidence data are evaluated from a design-based perspective and, as in the present article, their performance is checked in real communities via simulation studies of a design-based nature, these estimators are unsatisfactory, with unacceptable negative bias and risible coverages of confidence intervals. These findings are also confirmed by a simulation study of a similar nature (Chiu, 2022).

The analysis performed in this article shows that the failures of nonparametric estimators from incidence data are due mainly to the massive presence of rare species that are likely to be missed for realistic sampling efforts. In this scenario, the integration of probabilistic samples with accurate lists of rare species seems to be a unique solution for performing reliable design-based inference on species richness. In this setting, the use of the Lee and Chao (1994) criterion for estimating the number of species outside purposive lists seems to be a viable solution. These considerations are supported by the results achieved in the four case studies performed in nature reserves and parks in Central Italy, in which our data integration estimator furnishes estimates that are invariably greater than the minimum number of species surely present in the study region (i.e. those present in the purposive list and/or in the probabilistic sample): these results contrast those of SPADE, whose estimators give, in most cases, illogical estimates lower than these minima. These self-evident results should discourage naturalists from mechanically adopting automated software to estimate species richness.

However, finally, it should be noted that even if the improvements of our proposal with respect to the nonparametric estimators render the proposal appealing, we must stress that the performance of our proposal is strictly linked to the accuracy of the purposive lists. If the purposive lists fail to detect the rarest species, our estimation strategy fails as well.

Acknowledgments

The authors are grateful to Luca Pratelli from Naval Academy of Livorno (Italy) for his help in the theoretical issues of the article and to Alessandro Chiarucci from the 'Alma Mater' University of Bologna for providing the data of the four case studies. The authors acknowledge the funding by PRIN 2020 (cod. 2020E52THS)—Research Projects of National Relevance funded by the Italian

Ministry of University and Research entitled: 'Multi-scale observations to predict Forest response to pollution and climate change' (MULTIFOR, project number 2020E52THS). Funder: Project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4—Call for tender No. 3138 of 16 December 2021, rectified by Decree n.3175 of 18 December 2021 of Italian Ministry of University and Research funded by the European Union—NextGenerationEU; Award Number: Project code CN_00000033, Concession Decree No. 1034 of 17 June 2022 adopted by the Italian Ministry of University and Research, CUP B63C22000650007, Project title 'National Biodiversity Future Center - NBFC'.

Conflicts of interest: No competing interest is declared.

Funding

National Recovery and Resilience Plan (NRRP) European Union—NextGenerationEU, Grant/Award Number: CN_00000033. PRIN 2020 Research Projects MULTIFOR, project number: 2020E52THS.

Data availability

The data underlying this article will be shared on request to the corresponding author.

Author contributions

Conceptualization, L.F.; methodology, L.F.; software, R.M.D.B. and A.M.; formal analysis, L.F.; investigation, L.F., R.M.D.B., and A.M.; data curation, R.M.D.B. and A.M.; writing–original draft preparation, L.F.; writing–review and editing, R.M.D.B. and A.M. All authors have read and agreed to the published version of the manuscript.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series C*.

References

Adde A., Casabona i Amat C., Mazerolle M. J., Darveau M., Cumming S. G., & O'Hara R. B. (2021). Integrated modeling of waterfowl distribution in western Canada using aerial survey and citizen science (eBird) data. *Ecosphere*, *12*(10), e03790. https://doi.org/10.1002/ecs2.3790

Albert C. H., Yoccoz N. G., Edwards T. C., Graham C. H., Zimmermann N. E., & Thuiller W. (2010). Sampling in ecology and evolution – bridging the gap between theory and practice. *Ecography*, *33*(6), 1028–1037. https://doi.org/10.1111/j.1600-0587.2010.06421.x

Alessi N., Bońari G., Zannini P., JimÉnez-Alfaro B., Agrillo E., Attorre F., Canullo R., Casella L., Cervellini M., Chelli S., Di Musciano M., Guarino R., Martellos S., Massimi M., Venanzoni R., Zerbe S., & Chiarucci A. (2023). Probabilistic and preferential sampling approaches offer integrated perspectives of Italian forest diversity. *Journal of Vegetation Science*, *34*(1), e13175. https://doi.org/10.1111/jvs.13175

Bowler D. E., Nilsen E. B., Bischof R., O"Hara R. B., Thin Yu T., Oo T., Aung M., & Linnell J. D. C. (2019). Integrating data from different survey types for population monitoring of an endangered species: The case of the Eld's deer. *Scientific Reports*, *9*(1), 7766. https://doi.org/10.1038/s41598-019-44075-9

Boyd R. J., Powney G. D., & Pescott O. L. (2023). We need to talk about nonprobability samples. *Trends in Ecology & Evolution*, *38*(6), 521–531. https://doi.org/10.1016/j.tree.2023.01.001

Bunge J., & Fitzpatrick M. (1993). Estimating the number of species: A review. *Journal of the American Statistical Association*, *88*(421), 364–373. https://doi.org/10.1080/01621459.1993.10594330

Burnham K. P., & Overton W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, *65*(3), 625–633. https://doi.org/10.1093/biomet/65.3.625

Cazzolla Gatti R., Reich P. B., Gamarra J. G. P., Crowther T., Hui C., Morera A., Bastin J.-F., de-Miguel S., Nabuurs G.-J., Svenning J.-C., Serra-Diaz J. M., Merow C., Enquist B., Kamenetsky M., Lee J., Zhu J., Fang J., Jacobs D. F., Pijanowski B., … Liang J. (2022). The number of tree species on Earth. *Proceedings of the National Academy of Sciences*, *119*, e2115329119. https://doi.org/10.1073/pnas.2115329119

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

Chao A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, *43*, 783–791. https://doi.org/10.2307/2531532

Chao A. (2005). Species estimation, applications. In N. Balakrishnan, C.B. Readand & B. Vidakovic (Eds.). *Encyclopedia of statistical sciences* (2nd Ed., vol. *12*, pp. 7907–7916). Wiley.
Chao A., & Chiu C. H (2016a). Nonparametric estimation and comparison of species richness. In *Encyclopedia of life sciences* (pp. 1–11). Wiley.
Chao A., & Chiu C. H (2016b). Species richness: estimation and comparison. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, & J.L. Teugels (Eds.). *Wiley StatsRef: Statistics reference online* (pp. 1–26). Wiley.
Chao A., & Colwell R. K. (2017). Thirty years of progeny from Chao's inequality: Estimating and comparing richness with incidence data and incomplete sampling. *SORT*, *41*, 3–54.
Chao A., Gotelli N. J., Hsieh T. C., Sander E. L., Ma K. H., Colwell R. K., & Ellison A. M. (2014). Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs*, *84*(1), 45–67. https://doi.org/10.1890/13-0133.1
Chao A., & Lee S. M. (1992). Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, *87*(417), 210–217. https://doi.org/10.1080/01621459.1992.10475194
Chao A., Ma C. H., & Yang M. C. K. (1993). Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika*, *80*(1), 193–201. https://doi.org/10.1093/biomet/80.1.193
Chao A., Ma K. H., Hsieh C., & Chiu C. H. (2015). User's guide for online program SpadeR (species-richness prediction and diversity estimation in R). https://sites.google.com/view/chao-lab-website/software/spade
Chao A., & Shen T. J. (2010). User's guide for program SPADE (species prediction and diversity estimation). https://sites.google.com/view/chao-lab-website/software/spade
Chao A., Shen T. J., Ma K. H., & Hsieh T. C. (2003). User's guide for program SPADE (species prediction and diversity estimation). https://sites.google.com/view/chao-lab-website/software/spade
Chiarucci A. (2007). To sample or not to sample? That is the question …for the vegetation scientist. *Folia Geobotanica*, *42*(2), 209–216. https://doi.org/10.1007/BF02893887
Chiarucci A., Di Biase R. M., Fattorini L., Marcheselli M., & Pisani C. (2018). Joining the incompatible: Exploiting purposive lists for the sample-based estimation of species richness. *Annals of Applied Statistics*, *12*(3), 1679–1699. https://doi.org/10.1214/17-AOAS1126
Chiu C. H. (2022). Incidence-data-based species richness estimation via a Beta-Binomial model. *Methods in Ecology and Evolution*, *13*(11), 2546–2558. https://doi.org/10.1111/2041-210X.13979
Chiu C. H., Wang Y. T., Walther B. A., & Chao A. (2014). An improved nonparametric lower bound of species richness via a modified Good-Turing frequency formula. *Biometrics*, *70*(3), 671–682. https://doi.org/10.1111/biom.12200
Colwell R. K., Chao A., Gotelli N. J., Lin S.-Y., Mao C. X., Chazdon R. L., & Longino J. T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, *5*(1), 3–21. https://doi.org/10.1093/jpe/rtr044
Colwell R. K., & Coddington J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences*, *345*(1311), 101–118. https://doi.org/10.1098/rstb.1994.0091
D'Alessandro L., & Fattorini L. (2002). Resampling estimators of species richness from presence-absence data: Why they don't work. *Metron*, *60*, 5–19.
Fuller W. (2009). *Sampling statistics*. Wiley.
Gong G., & Samaniego F. J. (1981). Pseudo maximum likelihood estimation: Theory and applications. *Annals of Statistics*, *9*(4), 861–869. https://doi.org/10.1214/aos/1176345526
Gregoire T. G. (1998). Design-based and model-based inference in survey sampling: Appreciating the difference. *Canadian Journal of Forest Research*, *28*(10), 1429–1447. https://doi.org/10.1139/x98-166
Hedayat A. S., & Sinha B. K. (1991). *Design and inference in finite population sampling*. Wiley.
Heltshe J. F., & Forrester N. E. (1983). Estimating species richness using the jackknife procedure. *Biometrics*, *39*(1), 1–11. https://doi.org/10.2307/2530802
Hurlbert S. H. (1971). The nonconcept of species diversity: A critique and alternative parameters. *Ecology*, *52*(4), 577–586. https://doi.org/10.2307/1934145
Kim J. K., & Tam S. M. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*, *89*(2), 382–401. https://doi.org/10.1111/insr.12434
Lee S. M., & Chao A. (1994). Estimating population size via sample coverage for closed capture-recapture models. *Biometrics*, *50*(1), 88–97. https://doi.org/10.2307/2533199
Mao C. X., & Lindsay B. G. (2007). Estimating the number of classes. *Annals of Statistics*, *35*(2), 917–930. https://doi.org/10.1214/009053606000001280
Norris J. L., & Pollock K. H. (1998). Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species. *Environmental and Ecological Statistics*, *5*(4), 391–402. https://doi.org/10.1023/A:1009659922745
Palmer M. W. (1990). The estimation of species richness by extrapolation. *Ecology*, *71*(3), 1195–1198. https://doi.org/10.2307/1937387

Palmer M. W., Earls P. G., Hoagland B. W., White P. S., & Wohlgermuth T. (2002). Quantitative tools for perfecting species lists. *Environmetrics*, *13*(2), 121–137. https://doi.org/10.1002/env.516
Särndal C. E., Swensson B., & Wretman J. (1992). *Model assisted survey sampling*. Springer.
Shao J., & Tu D. (1995). *The Jackknife and the bootstrap*. Springer.
Smith E. P., & Van Belle G. (1984). Nonparametric estimation of species richness. *Biometrics*, *40*(1), 119–129. https://doi.org/10.2307/2530750
Stevens D. L. (1994). Implementation of a national monitoring program. *Journal of Environmental Management*, *42*(1), 1–29. https://doi.org/10.1006/jema.1994.1057
Thompson S. K. (2012). *Sampling* (3rd Ed.). Wiley.
Wang J. P. Z. (2010). Estimating species richness by a Poisson-compound gamma model. *Biometrika*, *97*(3), 727–740. https://doi.org/10.1093/biomet/asq026
Wang J. P. Z. (2011). SPECIES: An R package for species richness estimation. *Journal of Statistical Software*, *40*(9), 1–15. https://doi.org/10.18637/jss.v040.i09
Wang J. P. Z., & Lindsay B. G. (2005). A penalized nonparametric maximum likelihood approach to species richness estimation. *Journal of the American Statistical Association*, *100*(471), 942–959. https://doi.org/10.1198/016214504000002005