The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1679–1699
https://doi.org/10.1214/17-AOAS1126

JOINING THE INCOMPATIBLE: EXPLOITING PURPOSIVE LISTS FOR THE SAMPLE-BASED ESTIMATION OF SPECIES RICHNESS

BY ALESSANDRO CHIARUCCI*, ROSA MARIA DI BIASE†, LORENZO FATTORINI‡, MARZIA MARCHESELLI‡ AND CATERINA PISANI‡

University of Bologna, University of Tuscia† and University of Siena‡*

The lists of species obtained by purposive sampling by field ecologists can be used to improve the sample-based estimation of species richness. A new estimator is here proposed as a modification of the difference estimator in which the species inclusion probabilities are estimated by means of the species frequencies from incidence data. If the species list used to support the estimation is complete the estimator guesses the true richness without error. In the case of incomplete lists, the estimator provides values invariably greater than the number of species detected by the combination of sample-based and purposive surveys. An asymptotically conservative estimator of the mean squared error is also provided. A simulation study based on two artificial communities is carried out in order to check the obvious increase in accuracy and precision with respect to the widely applied estimators based on the sole sample information. Finally, the proposed estimator is adopted to estimate species richness in the Maremma Regional Park, Italy.

1. Introduction. The number of species in a biological assemblage or community, usually referred to as species richness, represents the simplest and most direct indicator of ecological diversity and is largely used as the most convenient proxy for other components of biodiversity [Gaston (1996)]. Since ecologists cannot detect each single plant or animal in a region (especially large regions such as parks, provinces, countries), species richness constitutes an unknown parameter of the community under study, which can be evaluated by means of a purposive survey of the study area, as traditionally performed by ecologists, or estimated through probabilistic sampling.

In purposive sampling (often referred to as "preferential sampling"), plant or animal species are recorded and listed by searching into specific sites or habitats expected to have a larger number of species, high detection rates or high abundances of rare species. On the other hand, in probabilistic sampling, plant or animal species are identified and listed only if present in the selected samples, such as plots, traps, soil samples and hearing points.

In the following we will refer to plants, since they represent excellent model organisms, because of their sessile life, even if, mutatis mutandis, analogous reasoning could be applied to limited mobility animals. Palmer et al. [(2002), page 122]

Received May 2017; revised November 2017.
Key words and phrases. Difference estimator, probabilistic sampling, purposive survey, supporting list, simulation.

emphasize the appeal of purposive surveys carried out by experienced botanists in completing plant species lists, who "*generally have a strong intuition or 'educated guess' about where to direct one's effort*". Moreover, Palmer et al. [(2002), page 122] outline the drawbacks of sample-based strategies which "*are likely to miss the rare or unclassifiable habitats that are likely to contribute most to regional diversity*" and are unlikely to "*outperform the guesses of experienced botanists*". However, when species lists are compiled by subjectively searching for plants, no probabilistic statement can be made about the accuracy and precision of species richness estimators.

On the other hand, estimators of species richness obtained by probabilistic sampling can be objectively evaluated through their sampling distributions, thus allowing for reliable comparisons across areas [e.g., Colwell and Coddington (1994), Hortal, Borges and Gaspar (2006), Cayuela, Gotelli and Colwell (2015)]. Also Palmer et al. [(2002, page 122)] recognize the importance of probabilistic sampling in comparing species richness throughout time and space, even if they point out that "*it would be unwise to dismiss the efficient, yet subjective, contributions of the expert botanist*". Therefore, it is at once apparent that procedures exploiting both the sources of information are highly advisable, although, to our knowledge, no effort has been undertaken in this direction.

Under probabilistic sampling, the estimation of species richness can be performed by using either abundance data or presence-absence data. Procedures based on abundance data include the parametric estimators obtained from fitting the species abundance distribution [e.g., Pielou (1977), and references therein] and some nonparametric estimators, such as those proposed by Chao (1984) and Chao and Lee (1992), usually referred to as Chao1 and ACE, respectively. Procedures based on presence-absence data include the parametric estimators obtained from the model-based extrapolation of species accumulation curve [e.g., Holdridge et al. (1971), Palmer (1990), Colwell and Coddington (1994) and references therein], the nonparametric estimators based on jackknife and bootstrap to compensate for the underestimation associated with the number of observed species, as proposed by Heltshe and Forrester (1983) and Smith and Van Belle (1984), and other estimators such as those proposed by Chao (1987) and Lee and Chao (1994), usually referred to as Chao2 and ICE, respectively [Gotelli and Chao (2013)]. Most of these researches have been recently reviewed in Chao et al. (2014) as well as in Chao and Colwell (2017). Most of these estimators and many case-oriented modifications have been implemented in many easily accessible softwares, such as EstimateS [Colwell (2013)] and vegan R package [Oksanen et al. (2016)] among others.

Estimators based on presence-absence data are certainly the most appropriate for plants and some sessile animals, given the problems in recognizing individual organisms, but can be suitable for animals too. Among these, the nonparametric methods obtained by jackknifing or bootstrapping the number of observed species seem to be more suitable than those based on extrapolating parametric species ac-

cumulation curves [Colwell and Coddington (1994), Colwell et al. (2012), Gotelli and Chao (2013)]. Regarding resampling estimators, it is worth noting that their performance has been checked by means of empirical results [Heltshe and Forrester (1983), Palmer (1990, 1991), Hellmann and Fowler (1999), Chiarucci et al. (2003)], highlighting that the bias reduction does not seem as substantial as should be expected. Theoretical studies on the properties of resampling estimators are due to Cormack (1989), who proved the inadequacy of jackknife to reduce bias when, as in the case of replicated plots or transects, the detection probabilities remain constant in every replication (the so-called M_h model). Subsequently D'Alessandro and Fattorini (2002) proved the design-based inadequacy of the resampling procedures to reduce bias, especially in presence of very rare species. On the other hand, the Chao2 and ICE estimators are more theoretically founded. The Chao2 estimator derives from the Cauchy–Schwarz inequality and Good–Turing formula and it results nearly unbiased when very rare/infrequent species have approximately the same detection probabilities [Chao and Colwell (2017)]. The ICE estimator derives from a rigorous sample coverage theory.

However, in the EstimateS User's Guide [Colwell (2013), Appendix B] it is pointed out that "*nonparametric estimators of species richness are minimum estimators: their computed values should be viewed as lower bounds of total species numbers*". Indeed, for hyper-diverse or severely under-sampled community, sample data do not contain sufficient information to provide accurate point estimates of species richness. Unless some strong parametric assumptions are made, at best one can only evaluate lower bounds because there may be many hard-to-detect species. As pointed out by I. J. Good, "*I don' t believe it is usually possible to estimate the number of unseen species... but only an approximate lower bound to that number. This is because there is nearly always a good chance that there are a very large number of rare species*" [cited by Bunge and Fitzpatrick (1993), page 370]. Therefore, in most ecological and conservation applications, a precise lower bound is preferable to nonaccurate point estimates.

The purpose of this paper is to exploit the use of information derived from species lists compiled by experienced ecologists by means of purposive surveys, henceforth referred to as the supporting list, to improve species richness estimates arising from probabilistic sampling.

The use of auxiliary information to improve estimation has a long standing in sample surveys. The most common way to exploit auxiliary information is the so-called difference (D) estimator and its subsequent modifications, such as the widely used generalized regression and ratio estimators [Särndal, Swensson and Wretman (1992), Chapter 6]. In this paper, a modified version of the D estimator, referred to as the empirical difference (ED) estimator, is proposed to take advantage of the supporting lists together with a presumably asymptotically conservative estimator of its sampling error. In order to check the improvement provided by list exploitation, the ED estimator, together with suitably modified Chao2 and Lincoln–Petersen (LP) estimators, are compared with nonparametric estimators

based on the sole presence-absence data, here considered as benchmark, by means of a simulation study performed on two artificial incidence data sets.

Section 2 contains some notations, while the ED estimator is introduced in Section 3 and a presumably asymptotically conservative estimator of its sampling error is derived in Section 4. Section 5 is devoted to the simulation setting and results. Details of the case study are reported in Section 6 while Section 7 is devoted to discussion and concluding remarks.

2. Notation and setting. It should be pointed out that notations introduced here are directed to statisticians and may sound unusual for ecologists. For this reason, a key to the ecological meaning of the symbols adopted is provided in Section SM1 of the Supplementary Material [Chiarucci et al. (2018)]. Consider a plant community within a delineated study area. From a statistical point of view the community constitutes a without-frame population of N plant individuals spread over the area. Owing to the lack of frame, the most effective probabilistic schemes for sampling plants differ from the traditional schemes and their choice is mainly determined by practical considerations on the nature of the community to be sampled. For example, when dealing with a shrub population, line intercept sampling may be suitable [e.g., Thompson (2002)] while, if the population is formed by the trees in a forest or by the whole plant assemblage containing very different species (e.g., all the vascular plants), plot sampling can be adopted [e.g., Gregoire and Valentine (2008)]. Ideally referring to the plants by their identifying numerical labels, the population can be represented by the set $\mathcal{U} = \{1, \ldots, N\}$ while $\mathcal{S} \subset \mathcal{U}$ denotes the sample of plants selected by means of a suitable scheme ensuring that the first-order inclusion probabilities can be determined directly or by some field measurements [e.g., Fattorini (2007)] for (at least) the selected plants.

If K species are present in the community, each group of individual plants belonging to the same species may be viewed as a unit, in such a way that the complete species list can be viewed as a population. Ideally, referring to species by their identifying numerical labels, the complete list of species can be represented by the set $\mathrm{C}(\mathcal{U}) = \{1, \ldots, K\}$, henceforth denoted for brevity by C, while K represents the species richness. Usually, the complete list of species is unknown and K must be estimated by means of sample surveys or evaluated through purposive surveys. It should be noticed that species constitute unknown assemblages of individual plants spread over the study area which cannot be sampled directly.

Thus, the most effective way for sampling species is to sample individual plants, or even plant ramets, in such a way that a species is sampled when at least one plant of that species is sampled. Practically speaking, any sample of plants $\mathcal{S} \subset \mathcal{U}$ univocally determines the corresponding sample of species $\mathrm{G}(\mathcal{S}) \subset \mathrm{C}$, henceforth denoted for brevity by G. Hence, the sampling scheme adopted to select plants, univocally determines the first-order inclusion probabilities of species $\theta_1, \ldots, \theta_K$.

Even if the schemes are designed to quantify the inclusion probabilities of (at least) the sampled plants, they do not allow for the quantification of species inclusion probabilities. Indeed, the quantification of the θ_js would entail the knowledge of all the units belonging to each species together with their spatial distribution over the study area [e.g., Fattorini (2007)].

Because a study area cannot be adequately sampled by means of only one plot or transect, n independent replications of the sampling scheme [Barabesi and Fattorini (1998)] are usually performed, giving rise to n samples of plants $\mathcal{S}_1, \ldots, \mathcal{S}_n$ which, in turn, give rise to n samples of species $\mathrm{G}_1, \ldots, \mathrm{G}_n$. The set of the species observed in the whole survey is $\mathrm{G}_{(n)} = \bigcup_{i=1}^{n} \mathrm{G}_i$ and its size SO_n is the number of observed species. Owing to the independence of the replications, the probability that species j enters the pooled sample $\mathrm{G}_{(n)}$ (i.e., it is detected during the whole sample survey) turns out to be $\tau_j = 1 - (1 - \theta_j)^n$.

For each replication i let $z_i = (z_{i1}, \ldots, z_{iK})^T$ be the K-vector in which the jth element z_{ij} is equal to 1 if the species j has been sampled and 0 otherwise. Usually, $z_1, \ldots, z_n$ are organized into a 0–1 matrix of n columns and SO_n rows, commonly referred to as presence-absence or incidence data. The n vectors $z_1, \ldots, z_n$ are independent realizations of the random vector $Z = (Z_1, \ldots, Z_K)^T$ with expectation $\theta = (\theta_1, \ldots, \theta_K)^T$. Now denote by $x = \sum_{i=1}^{n} z_i$ the realization of the random vector $X = (X_1, \ldots, X_K)^T$ in which each marginal variable X_j has a binomial distribution with parameters n and θ_j. Because x_j is the number of replications in which the species j has been sampled, $x_j = 0$ for all the undetected species. Thus, even if theoretically x is a K-vector, it contains an unknown number of zeros [D'Alessandro and Fattorini (2002)].

3. The empirical difference estimator. Denoting by Y a variable such that $y_j = 1$ for each species $j \in \mathrm{C}$, the species richness K can be written as $K = \sum_{j \in \mathrm{C}} y_j$. Let L be the set of species detected by means of a purposive survey of the study area, as traditionally performed by botanists, and let M be their number. Obviously, $\mathrm{L} \subset \mathrm{C}$ in such a way that $M \leq K$.

In accordance with the approach leading to the D estimator [Chiarucci et al. (2018)], the dichotomous variable Y^0 such that $y_j^0 = 1$ if $j \in \mathrm{L}$ and 0 otherwise can be adopted as a proxy for the survey variable Y. The errors in predicting the y_js by means of the y_j^0s, that is, $y_j - y_j^0 = 1 - y_j^0$, are equal to 0 for any species $j \in \mathrm{L}$ and to 1 otherwise. Accordingly, Y^0 is a good proxy for Y when the supporting list is accurate. By using the proxy variable, the species richness can be rewritten as $K = \sum_{j \in \mathrm{C}} y_j^0 + \sum_{j \in \mathrm{C}} (y_j - y_j^0) = M + \sum_{j \in \mathrm{C}} (1 - y_j^0)$, where M is a known constant (i.e., the number of species detected by means of purposive sampling) while the second term is the unknown total of errors to be estimated from the sample. If the probabilities of the species to enter the pooled sample $\mathrm{G}_{(n)}$ were known, the species richness could be estimated by means of the D estimator, which

reduces to

$$\widehat{K}_D = M + \sum_{j \in \mathrm{G}_{(n)}} \frac{1 - y_j^0}{\tau_j} = M + \sum_{j \in \mathrm{G}_{(n)} - \mathrm{L}} \frac{1}{\tau_j} \tag{3.1}$$

because the errors $1 - y_j^0$ vanish for each $j \in \mathrm{L}$.

As stated in the Appendix, the estimator $\widehat{K}_D$ would be unbiased with a closed-form variance which could be unbiasedly estimated from the sample. Actually, the τ_js are unknown depending on the θ_js and the estimator $\widehat{K}_D$ cannot be computed from the available information. As suggested by Fattorini (2006, 2009), the frequencies x_js in which the species enter the n samples can be adopted to estimate the θ_js by means of $\hat{\theta}_j = (x_j + 1)/(n + 1)$. Therefore, the estimate of τ_j is given by $\hat{\tau}_j = 1 - (n - x_j)^n/(n + 1)^n$ and, using the estimated probabilities into (3.1), the ED estimator turns out to be

$$\widehat{K}_E = M + \sum_{j \in \mathrm{G}_{(n)} - \mathrm{L}} \frac{1}{\hat{\tau}_j}. \tag{3.2}$$

An alternative estimator of τ_j has been recently proposed by Chao and Colwell (2017), page 25. As opposite to the D estimator, the ED estimator is biased with expectation and variance which cannot be expressed in closed forms. However, the ED estimator maintains the following appealing properties: first, its realizations are never smaller than the cardinality of the set $\mathrm{G}_{(n)} \cup \mathrm{L}$, that is, the number of species detected by the combination of purposive and sample surveys; furthermore, if the supporting list is perfect, the ED estimator invariably estimates the true species richness without error. Hence, besides the uncertainty due to the estimation of the inclusion probabilities, the uncertainty of $\hat{K}_E$ is completely due to the species in the set $\mathrm{C} - \mathrm{L}$, that is, the species lost in the supporting list which can be partially recovered by the sample survey. It should be noticed that if $\mathrm{G}_{(n)} - \mathrm{L}$ is the empty set, no additional species is detected by the sample survey with respect to the list. In other words, the sample survey has not provided any additional information; in this case, the second term in equation (3.2) is 0 and the ED estimator coincides with M.

4. Sampling error estimation. Because $\widehat{K}_E$ is a biased estimator, there is no sense in estimating its variance. Indeed, as Särndal and Lundström [(2005), page 8] point out, the bias of any estimator should be the main concern. Variance and its estimation are of minor importance since "*if an estimator is greatly biased, it is poor consolation that its variance is low*". Rather, we should estimate the mean squared error $\mathrm{MSE}(\widehat{K}_E) = \mathrm{E}\{(\widehat{K}_E - K)^2\}$ or, more meaningfully, the relative root mean square error

$$\mathrm{RRMSE}(\widehat{K}_E) = \sqrt{\mathrm{MSE}(\widehat{K}_E)}/K.$$

Because neither the mean nor the variance of $\widehat{K}_E$ can be expressed in a closed form, we derived an upper bound for $\mathrm{MSE}(\widehat{K}_E)$ to be subsequently estimated from the available information, in such a way that the resulting estimator should be presumably asymptotically conservative, that is, it should overestimate the actual sampling error. In the Appendix it is proven that,

$$\mathrm{MSE}(\widehat{K}_E) \leq 2K\left(4e^{-1}+1\right) \sum_{j\in \mathrm{C}-\mathrm{L}} \left\{1-\theta_j\left(1-e^{-1}\right)\right\}^n \tag{4.1}$$

in such a way that the right side of (4.1) can be estimated by

$$\widehat{\mathrm{MSE}}(\widehat{K}_E) = 2\widehat{K}_E\left(4e^{-1}+1\right) \sum_{j\in \mathrm{G}_{(n)}-\mathrm{L}} \frac{\{1-\hat{\theta}_j(1-e^{-1})\}^n}{\hat{\tau}_j} \tag{4.2}$$

which, at least asymptotically, should be a conservative estimator. Also in this case, if $\mathrm{G}_{(n)} - \mathrm{L}$ is the empty set, the second term in equation (4.2) is 0 and the MSE estimate turns out to be 0.

From (4.2), the estimator of $\mathrm{RRMSE}(\widehat{K}_E)$ is given by

$$\begin{aligned} \widehat{\mathrm{RRMSE}}(\widehat{K}_E) &= \frac{\sqrt{\widehat{\mathrm{MSE}}(\widehat{K}_E)}}{\widehat{K}_E} \\ &= \sqrt{\frac{2(4e^{-1}+1)}{\widehat{K}_E} \sum_{j\in \mathrm{G}_{(n)}-\mathrm{L}} \frac{\{1-\hat{\theta}_j(1-e^{-1})\}^n}{\hat{\tau}_j}}. \end{aligned} \tag{4.3}$$

It is worth noting that inequality (4.1) suffices to prove that $\widehat{K}_E$ converges in quadratic mean (and hence also in mean) to K because

$$\lim_{n\to\infty} 2K\left(4e^{-1}+1\right) \sum_{j\in \mathrm{C}} \left\{1-\theta_j\left(1-e^{-1}\right)\right\}^n \left(1-y_j^0\right) = 0.$$

Therefore, $\widehat{K}_E$ is a consistent estimator of K with bias and variance approaching 0 as the number of replications increases.

5. Simulation study.

5.1. *Simulation setting*. In order to check the improvement provided by the exploitation of floristic list in the ED estimator with respect to the nonparametric estimators based on the sole presence-absence data, a simulation study was performed on two artificial plant communities.

Because plants communities in large areas are composed of coexisting species, some of them overlapping and some others avoiding each other, the first artificial community was constituted by $K = 100$ species, partitioned into 4 exhaustive and mutually exclusive groups of 25 species with nested distributions.

Accordingly, without specifying the sampling scheme (which could be plot sampling, line intercept sampling or any other suitable environmental scheme), within each group the θ_js were assumed to decrease geometrically from a maximum of 0.25 with a decreasing factor of 0.8, and if species j was sampled, all the species having first-order inclusion probabilities greater than θ_j were subsequently sampled. The second-order inclusion probabilities, say θ_{jh}, vanished for all the pairs (j, h) belonging to different groups while turned out to be $\theta_{jh} = \min(\theta_j, \theta_h)$ for all the pairs (j, h) belonging to the same group. In order to simulate species sampling, for each replication i $(i = 1, \ldots, n)$ the vector z_i was independently generated by means of the following algorithm: (i) a group of species was randomly selected among the 4 groups; (ii) a random number u was generated from the uniform distribution on $(0, 0.25)$ and all the species of the selected group having first-order inclusion probability greater than u were included in the sample.

The second artificial community was composed by $N = 250{,}000$ forest trees partitioned into $K = 228$ species settled on a rectangular region of size 1×0.5 km^2. As to the apportionment of abundance to species, the number of individuals of the most abundant species was 40,000, two species had 20,000 individuals each, four species had 10,000 individuals each, eight species had 5,000 individuals each, sixteen species had 2,500 individuals each, thirty-two species had 1,250 individuals each, five had 1,000 individuals each, five had 500 individuals each, five had 250 individuals each, five had 100 individuals each, five had 50 individuals each, ten had 20 individuals each, ten had 10 individuals each, twenty had 5 individuals each. Finally, 100 species had only one individual each. Each individual, irrespective of its species, was randomly placed on the rectangle. The resulting community roughly resembled the structure of some tropical forests widely exploited in ecological studies. For this population, the n independent vectors $z_1, \ldots, z_n$ were generated presuming n circular plots of radius $r = 13$ m randomly located on the study region in such a way that $z_{ij} = 1$ if at least one individual of the species j was contained in plot i, while $z_{ij} = 0$ otherwise.

For both the artificial communities, nine types of supporting list were supposed. The first three lists, denoted by L_1, L_2, L_3, were artificially achieved deleting from the communities the 1%, 5% and 10% of the rarest species, respectively. The further three lists, denoted by L_4, L_5, L_6, were achieved deleting from the communities the 1%, 5% and 10% of the most common species, respectively. Finally, the last three lists, denoted by L_7, L_8, L_9, were achieved deleting from the communities the 1%, 5% and 10% of both rarest and most common species, respectively. It should be noticed that the lists missing the most common species are not so unrealistic, because botanists often focused on searching the rarest species, sometimes neglecting the most common ones [Palmer et al. (2002)]. For each artificial community and for $n = 50, 100, 150$, $R = 10{,}000$ presence-absence matrices were generated as previously described.

From each generated matrix, the following nonparametric estimators were computed: species observed (SO_n), first-order jackknife ($\widehat{K}_{\text{jack1}}$), second-order jackknife ($\widehat{K}_{\text{jack2}}$), bootstrap ($\widehat{K}_{\text{boot}}$), Chao2 ($\widehat{K}_{\text{Chao2}}$) and ICE ($\widehat{K}_{\text{ICE}}$). Moreover the

ED estimator $\widehat{K}_E$ was computed for each of the nine supporting lists. Finally, for each list, the estimates of the sampling error were computed by means of (4.3). Then, the relative bias (RB) and the relative root mean squared error (RRMSE) were derived from the Monte Carlo distributions. Moreover, for each of the nine lists, the expectations of the sampling error estimator were also computed.

5.2. *Alternative floristic list exploitations*. Besides the ED estimator, there may be several ways to incorporate the floristic lists in the familiar estimators of species richness. For example, the Chao2 estimator can be adapted treating the floristic list as one sample and the pooled n-sample data as the second sample, and then using equations (3a) or (3b) in Chao and Colwell (2017) to estimate the variance. In a similar way, the list presence can be used in the framework of a simple two-sample capture-recapture analysis, treating the floristic list as the first "capture" sample, and treating the pooled n-plot data as the second "recapture" sample. Then the well-known Lincoln–Petersen estimator can be computed together with its variance estimator [see, e.g., Seber (1982), page 60]. The estimator is approximately unbiased under assumptions (a)–(f) in Seber (1982), page 59. For both the estimators and for the nine lists L_1–L_9, the RB and RRMSE were derived from the Monte Carlo distributions together with the expectation of the estimators of their sampling errors.

5.3. *Simulation results*. Table 1 reports the percentage values of RB and RRMSE for each nonparametric estimator and for the ED, Chao2 and LP estimators corresponding to each supporting list when the first artificial community is considered. Regarding the ED, Chao2 and LP estimators, values in brackets are the expectations of the RRMSE estimators. Table 2 reports the same indicators achieved from the second artificial community.

For $n = 50$, among the nonparametric estimators $\widehat{K}_{\text{Chao2}}$ provides the best RB for both the artificial communities, equal to -15% and -36% respectively, while the ED estimator provides RBs invariably better than -10%. Obviously, the RBs of the ED estimator strictly depend on the accuracy of the supporting lists with better RBs achieved when the lists miss the most common species (lists L_4, L_5, L_6). In these cases, bias disappears in both the artificial communities.

Similar conclusions can be drawn regarding the RRMSEs, because the main part of the sampling errors are due to bias. For $n = 50$, $\widehat{K}_{\text{jack2}}$ provided the smallest RRMSE (about 30%) in the first artificial community and $\widehat{K}_{\text{Chao2}}$ provided the smallest RRMSE (about 39%) in the second artificial community, while the ED estimator provides RRMSEs invariably smaller than 10% for both the artificial communities. Also in this case, the smallest values of RRMSE for the ED estimator are achieved when the lists miss the most common species.

As the number of replications increases to 100 and 150, the RBs and RRMSEs decrease for all the nonparametric estimators. The decrease is less marked in the second artificial community, owing to the massive presence of rare species

TABLE 1

Percentage values of relative bias (RB) and relative root mean squared error (RRMSE) for each nonparametric estimator and for the ED, Chao2 and LP estimators corresponding to each supporting list in the case of the first artificial community. Values in brackets are the expectations of the relative root mean squared error estimators

Estimator	$n = 50$		$n = 100$		$n = 150$	
	RB	**RRMSE**	**RB**	**RRMSE**	**RB**	**RRMSE**
Species observed	−42.9	44.0	−31.9	33.2	−25.4	26.8
First-order jackknife	−26.1	30.9	−16.0	21.7	−10.2	16.9
Second-order jackknife	−18.1	29.7	−8.7	22.8	−3.8	20.1
Bootstrap	−35.1	37.3	−24.4	26.9	−18.2	21.0
Chao2	−14.9	49.5	−6.4	41.5	−2.2	38.5
ICE	−25.7	33.9	−16.5	25.6	−11.2	21.3
ED exploiting L_1	−0.9	1.0 (0.7)	−0.9	0.9 (1.4)	−0.8	0.9 (2.0)
ED exploiting L_2	−4.6	4.7 (3.6)	−4.3	4.4 (6.1)	−4.0	4.2 (7.9)
ED exploiting L_3	−9.2	9.3 (5.9)	−8.4	8.6 (10.4)	−7.7	8.0 (14.4)
ED exploiting L_4	0.0	0.0 (0.4)	0.0	0.0 (0.0)	0.0	0.0 (0.0)
ED exploiting L_5	0.0	0.0 (1.8)	0.0	0.0 (0.1)	0.0	0.0 (0.0)
ED exploiting L_6	0.0	0.0 (3.8)	0.0	0.0 (0.5)	0.0	0.0 (0.0)
ED exploiting L_7	−0.9	1.0 (1.1)	−0.9	0.9 (1.4)	−0.8	0.9 (2.0)
ED exploiting L_8	−4.6	4.7 (5.0)	−4.3	4.4 (6.2)	−4.0	4.2 (7.9)
ED exploiting L_9	−9.2	9.3 (8.9)	−8.4	8.6 (10.6)	−7.7	8.0 (14.4)
Chao2 exploiting L_1	8.1	9.9 (3.8)	3.3	4.4 (2.4)	1.6	2.4 (1.7)
Chao2 exploiting L_2	3.0	5.6 (3.6)	−0.9	2.5 (2.2)	−2.1	2.6 (1.6)
Chao2 exploiting L_3	−3.0	4.8 (3.3)	−5.9	6.2 (2.0)	−6.3	6.6 (1.5)
Chao2 exploiting L_4	10.0	11.8 (4.0)	4.8	5.7 (2.5)	2.8	3.5 (1.9)
Chao2 exploiting L_5	12.9	15.0 (4.6)	6.3	7.4 (3.0)	3.9	4.7 (2.2)
Chao2 exploiting L_6	17.4	20.2 (5.6)	8.7	10.1 (3.7)	5.6	6.5 (2.8)
Chao2 exploiting L_7	8.7	10.5 (3.9)	3.7	4.8 (2.5)	1.9	2.7 (1.8)
Chao2 exploiting L_8	6.0	8.8 (4.4)	0.7	3.2 (2.8)	−1.0	2.2 (2.1)
Chao2 exploiting L_9	3.2	7.9 (5.1)	−2.5	4.1 (3.3)	−4.0	4.6 (2.5)
LP exploiting L_1	−0.9	1.0 (0.1)	−0.9	1.0 (0.1)	−0.8	0.9 (0.1)
LP exploiting L_2	−4.5	4.6 (0.2)	−4.2	4.4 (0.3)	−3.9	4.1 (0.3)
LP exploiting L_3	−8.9	9.1 (0.4)	−8.2	8.5 (0.5)	−7.5	7.8 (0.6)
LP exploiting L_4	0.8	0.9 (1.2)	0.5	0.5 (0.8)	0.4	0.4 (0.7)
LP exploiting L_5	4.4	4.7 (2.8)	2.7	2.9 (2.0)	1.9	2.1 (1.6)
LP exploiting L_6	9.8	10.7 (4.3)	5.8	6.4 (3.0)	4.1	4.5 (2.4)
LP exploiting L_7	−0.1	0.5 (1.2)	−0.4	0.6 (0.9)	−0.4	0.6 (0.7)
LP exploiting L_8	−0.6	1.8 (2.8)	−1.9	2.3 (2.0)	−2.3	2.6 (1.6)
LP exploiting L_9	−1.0	3.6 (4.2)	−3.8	4.4 (2.9)	−4.5	4.9 (2.3)

which renders ineffective the bias reduction induced by the nonparametric estimators [D'Alessandro and Fattorini (2002)].

Regarding the ED estimator, the decreases in RBs and RRMESs are slight because the main source of bias and uncertainty are due to the accuracy of the sup-

TABLE 2

Percentage values of relative bias (RB) and relative root mean squared error (RRMSE) for each nonparametric estimator and for the ED, Chao2 and LP estimators corresponding to each supporting list in the case of the second artificial community. Values in brackets are the expectations of the relative root mean squared error estimators

Estimator	$n = 50$		$n = 100$		$n = 150$	
	RB	**RRMSE**	**RB**	**RRMSE**	**RB**	**RRMSE**
Species observed	−52.7	52.7	−46.9	46.9	−45.9	46.0
First-order jackknife	−45.4	45.5	−37.3	37.5	−36.1	36.5
Second-order jackknife	−40.4	40.7	−30.9	31.3	−29.5	30.5
Bootstrap	−49.7	49.7	−42.8	42.9	−41.7	42.0
Chao2	−35.7	39.2	−28.3	31.9	−27.3	31.2
ICE	−40.1	40.7	−30.3	31.3	−29.3	30.8
ED exploiting L_1	−0.8	0.8 (0.9)	−0.8	0.8 (1.7)	−0.8	0.8 (1.8)
ED exploiting L_2	−4.5	4.6 (4.2)	−4.3	4.3 (7.1)	−4.2	4.2 (8.0)
ED exploiting L_3	−9.5	9.5 (7, 7)	−8.9	9.0 (12.1)	−8.8	8.8 (13.0)
ED exploiting L_4	0.0	0.0 (0.0)	0.0	0.0 (0.0)	0.0	0.0 (0.0)
ED exploiting L_5	0.0	0.0 (0.0)	0.0	0.0 (0.0)	0.0	0.0 (0.0)
ED exploiting L_6	0.0	0.0 (0.0)	0.0	0.0 (0.0)	0.0	0.0 (0.0)
ED exploiting L_7	−0.8	0.8 (0.9)	−0.8	0.8 (1.7)	−0.8	0.8 (1.8)
ED exploiting L_8	−4.5	4.6 (4, 2)	−4.3	4.3 (7.1)	−4.2	4.2 (8.0)
ED exploiting L_9	−9.5	9.5 (7.7)	−8.9	9.0 (12.1)	−8.8	8.8 (13.0)
Chao2 exploiting L_1	13.5	13.6 (3.4)	9.3	9.4 (2.8)	8.9	9.2 (2.7)
Chao2 exploiting L_2	7.9	8.0 (3.3)	4.5	4.6 (2.7)	4.3	4.7 (2.6)
Chao2 exploiting L_3	0.6	1.4 (3.1)	−1.8	2.2 (2.5)	−1.9	2.5 (2.4)
Chao2 exploiting L_4	15.5	15.6 (3.6)	11.0	11.0 (2.9)	10.5	10.8 (2.8)
Chao2 exploiting L_5	19.6	19.7 (4.1)	13.9	14.0 (3.3)	13.4	13.8 (3.2)
Chao2 exploiting L_6	26.6	26.7 (5.0)	18.9	19.1 (4.0)	18.3	18.8 (3.9)
Chao2 exploiting L_7	14.3	14.3 (3.5)	9.9	10.0 (2.9)	9.5	9.8 (2.8)
Chao2 exploiting L_8	12.3	12.4 (3.9)	7.7	7.8 (3.2)	7.4	7.9 (3.1)
Chao2 exploiting L_9	10.2	10.4 (4.6)	5.2	5.5 (3.8)	4.9	5.8 (3.7)
LP exploiting L_1	−0.8	0.8 (0.1)	−0.7	0.8 (0.1)	−0.7	0.8 (0.1)
LP exploiting L_2	−4.3	4.4 (0.3)	−4.0	4.0 (0.5)	−3.8	3.9 (0.5)
LP exploiting L_3	−9.1	9.2 (0.6)	−8.4	8.4 (0.7)	−8.2	8.2 (0.8)
LP exploiting L_4	1.0	1.0 (1.0)	0.8	0.8 (0.8)	0.8	0.8 (0.8)
LP exploiting L_5	5.9	6.0 (2.4)	4.7	4.7 (2.0)	4.5	4.6 (1.9)
LP exploiting L_6	14.2	14.2 (3.8)	10.9	10.9 (3.1)	10.6	10.7 (3.1)
LP exploiting L_7	0.2	0.4 (1.0)	0.1	0.4 (0.8)	0.0	0.4 (0.8)
LP exploiting L_8	1.2	1.4 (2.4)	0.3	1.0 (2.1)	0.4	1.1 (2.0)
LP exploiting L_9	2.8	3.1 (3.8)	0.8	1.7 (3.2)	0.8	1.9 (3.1)

porting lists rather than to the sampling effort. However, the comparison with the ED estimator leads to conclusions similar to those achieved for $n = 50$.

The sole occasion in which the nonparametric estimators outperform some of the nine ED estimators is in the case of the first artificial community and $n = 150$

when $\widehat{K}_{\text{Chao2}}$ reaches a RB of -2.2% and $\widehat{K}_{\text{jack2}}$ reaches a RB of -3.8%, while the ED estimator performed with the lists missing the rarest species shows RBs of -4% (lists L_2, L_8) and -7.7% (lists L_3, L_9).

Regarding the estimation of the RRMSE of the ED estimator, the estimator (4.3) turns out to be conservative for both the artificial communities for $n = 100, 150$, while for $n = 50$ some underestimations occur when the lists miss the rarest species. Regarding the use of floristic list in the Chao2 and LP estimators, as proposed in Section 5.2, the LP estimators provides performance comparable to that of the ED estimator when the lists miss the rarest species, behaves worse when the lists miss the most common species, but behaves better when the lists miss both common and rare species. As to the Chao2 estimator, the use of floristic lists invariably improves its precision with considerable reductions of the RRSMSE with respect to the case in which only sample data are used. However it performs worse than ED and LP estimators when the lists miss the most common species or both common and rare species; when the lists miss the rarest species the Chao2 estimator provides less than obvious results, with RB and RRMSE that decrease as the fraction of lost species increases. Regarding the estimation of the sampling errors, for both the estimators the simulation results demonstrate a tendency to underestimate their uncertainty, thus involving an overevaluation of their actual precision. Probably, this issue necessitates further investigations.

6. Field study at Maremma Regional Park. A sample survey was planned within the territory of the Maremma Regional Park to estimate the species richness of vascular plants. The Maremma Regional Park is a protected area of 9400 ha located along the Tuscan coastline, Italy, and covered by different types of Mediterranean vegetation, but also by pastures and agricultural areas. The sample survey was performed during the spring-summer period of 2006–2007 by means of $n = 90$ random plots of size 100 m^2. Each plot was 10 m $\times$ 10 m and was centred in the random point once located with a high precision GPS. To facilitate plant recording, each plot was divided into 16 smaller (2.5 m $\times$ 2.5 m) subplots. Plants were identified at species level using floras [e.g., Pignatti (1982)] and monographs. Nomenclature was standardised according to Conti et al. (2005). The number of species observed in the 90 plots was $SO_{90} = 608$.

Besides the sample survey, a floristic list was compiled by Arrigoni (2003) recording plant species in the same area during a period of almost two decades. The floristic list achieved from such a purposive survey contained $M = 846$ species. Among these, 492 species were present in the list and were observed in the plots, while 116 were observed in the plots and missed by the list and 354 were present in the list but missed by the plots. On the whole, a total of 962 species were observed, which constituted the minimum number of species living in the park. Obviously any estimate of species richness should be greater than 962. The incidence data of the 90 plots for the 962 observed species, together with the floristic list, are available in Chiarucci et al. (2018). From the sole incidence data, the nonparametric

estimates of species richness turned out to be $\widehat{K}_{\text{jack1}} = 801.82$, $\widehat{K}_{\text{jack2}} = 883.25$, $\widehat{K}_{\text{boot}} = 698.38$, $\widehat{K}_{\text{Chao2}} = 776.49$ and $\widehat{K}_{\text{ICE}} = 797.83$. All were smaller than the minimum number of species living in the park.

On the other hand, exploiting the information provided by the floristic list, the ED estimate was $\widehat{K}_{\text{ED}} = 973.51$. Moreover, one can treat the floristic list as the first list and the pooled 90-plot data as the second list. In this two-list case, the number of singletons was 470, the number of doubletons was 492, leading to $\widehat{K}_{\text{Chao2}} = 962 + (1/2)(470 \times 470)/(2 \times 492) = 1074.25$. Likewise, one may treat the floristic list as the first capture sample and the pooled 90-plot data as the second recapture sample. Since there were 492 recaptures, the Lincoln–Petersen estimate was $\widehat{K}_{\text{LP}} = (608 + 1) \times (846 + 1)/(492 + 1) - 1 = 1045.29$. These two estimates are very close and higher than the ED estimate. The estimate of the RRMSE of ED based on (4.3) turned out to be 36.3%, which denoted a high level of uncertainty, probably due to an imperfect supporting list as well as to the inadequacy of the sampling effort limited to 90 plots that covered an area of less than 0.01% of the park area. On the other hand, the estimate of RRMSE of Chao2 and LP turned out to be 1.5% and 1.3%. These low values were almost certainly due to the tendency of these estimators to underevaluate the sampling error, as demonstrated in the simulation study.

7. Discussion. Estimating species richness is one of the most relevant and most problematic issues in biodiversity studies [Colwell and Coddington (1994), Gotelli and Colwell (2001), Chiarucci, Bacaro and Scheiner (2011), Colwell et al. (2012), Gotelli and Chao (2013)]. The number of species is often considered one of the most direct and useful measure of biodiversity and it is widely used both for theoretical and applied topics [e.g., see Howard et al. (1998), Gotelli et al. (2009), Wilson et al. (2012)]. Theoretically, the number of species can be easily censused in a small area, as it is the case of the number of plant species growing in a relatively small plot, but estimation methods are mandatory on larger areas, such as parks or regions, and for those organisms that cannot be exhaustively censused, such as the soil fauna species [Nichols and Conroy (1996), Skov and Lawesson (2000), Gotelli and Colwell (2001)]. Therefore, the development of reliable species richness estimators is still a major challenge for present day ecology. This is also demonstrated by the massive use of nonparametric estimators performed by field ecologists and practitioners to get more reliable estimates with respect to the recorded number of species.

Despite the relevance of species richness estimation, and the large use of some nonparametric estimators, there is no theoretical consensus on the reliability of such methods. In particular, a great discrepancy among the performance of the nonparametric estimators based on incidence data has been reported [Heltshe and Forrester (1983), Palmer (1990, 1991), Walther and Morand (1998), Hellmann and Fowler (1999), Skov and Lawesson (2000), Chiarucci et al. (2003), Melo (2004),

Walther and Moore (2005)]. The paper by D'Alessandro and Fattorini (2002), despite being not well recognized in the ecological literature, proved how the limits of the jackknife and bootstrap estimators are due to the inclusion probabilities of rare species that make their performance unsatisfactory. In a recent paper, Xu et al. (2012) used a large data set to compare different estimators and found that non-parametric estimators (Chao1, ACE, Chao2, first- and second-order jackknife and bootstrap estimators) underestimated the real number of species, while area-based estimators (Ugland's method, Shen and He's method, Power-law model, Exponential model, Logistic model, MaxEnt method) overestimated this number. One of the authors of this paper, in commenting on the paper by Xu et al. (2012), titled "*Estimating species richness: still a longway off!*" [Chiarucci (2012)] to remark how difficult the estimation of species richness is on field data. Similar examples are also available in recent literature.

The exploitation of supporting species lists is likely to provide a totally new perspective in the issue of estimating species richness in large areas, by combining the data obtained by a probabilistic sampling with those arising from purposive sampling. The information contained in species lists compiled by field ecologists by means of purposive surveys is typically very difficult to use, since the methods adopted for collecting such data are not well formalised and are highly variable in time and space [Palmer et al. (2002), Diekmann, Kühne and Isermann (2007), Fattorini (2013)]. However, it is well known that probabilistic sampling is likely to miss most of the rare and ecologically most important species, as well as some of the sites with the highest species richness [Palmer et al. (2002), Hédl (2007), Diekmann, Kühne and Isermann (2007)]. Thus, the possibility of using information recorded in purposive sampling, that can be specifically devoted to survey sites giving a high contribution to the species richness of the region, can be seen as a powerful challenge for theoretical and empirical ecologists, as well as for conservation biologists and practitioners.

Now, by using the ED estimator, the lists of species can be used to improve the design-based estimates and if the supporting list is complete the estimator guesses the true richness with no error. Moreover, in the case of incomplete supporting lists, the ED estimator provides estimates that are invariably greater than the number of species detected combining sample-based and purposive surveys. A presumably asymptotically conservative estimator of the mean squared error is also provided. About this aspect, it should be noticed that most papers dealing with species richness estimation propose estimators of the sampling variances rather than of the mean squared errors. However, in the framework of species richness estimation, in which estimators are usually affected by a large amount of negative bias, the estimation of variance is irrelevant, because the major part of the sampling error is due to bias rather than to variance (see Tables 1 and 2). At least to our knowledge, this is the first attempt to estimate mean squared error in species richness estimation.

Simulation results evidence the relevant improvement in bias and precision provided by the proposed estimator with respect to nonparametric estimators, especially when the supporting list is accurate. Also in this case, the rare species are the key problem. If the supporting list misses some rare species, they are unlikely to be recovered in the sample-based survey. Thus the resulting estimator is negatively biased for a quantity approximately equal to the number of rare species missed in the list. Moreover, for the same reason, the summand in (4.2) is likely to be 0, thus involving a possible underestimation of the mean squared error. On the other hand, if the supporting list misses some common species, they are likely to be recovered in the sample-based surveys and bias disappears.

Practically speaking, the proposed estimator is likely to be effective when efforts in compiling lists are mostly directed toward habitats that are likely to host rare species. If most of them are detected, the ED estimator can represent an efficient solution. Thus, at least for a sufficiently large number of replications (plots or transects), the ED estimator is able to merge the information acquired from opportunistic surveys with a statistically sound inference on the true species richness, which is not possible when the nonparametric estimators are adopted.

Relevant improvement in accuracy and precision are also achieved by incorporating purposive list in the Chao2 and LP estimators, even if the estimation of their sampling errors necessitates further investigations. An added value of the use, in the real world, of the purposive list in the estimation of species richness is that the lists of species traditionally compiled by field ecologists with only descriptive aims assume a new value, as a potential tool to improve the performance of a statistical estimation of species richness. Therefore, the work done by field ecologists, or even citizien science group, in preparing species lists will be now considered in a more scientifically sound way and can assume a major relevance. In this new scenario, particular care must be taken in selecting the list to be exploited, especially when the lists are cumulative over many years [Palmer et al. (2002)]. In such cases, the purposive lists may include species that were, but are not any more present in the area (e.g., species extirpated due to habitat conversion) or species that were not considered by the ecologist for the list since they were introduced. In addition, taxonomical problems (e.g., synonyms of species that were split into two or species that were merged) can affect the matching of the purposive with the sample-based lists. All these drawbacks are likely to deteriorate the performance of estimators exploiting floristic lists and need special attention.

As a concluding remark, we would like to point out that our purpose was to encourage the exploitation of the information contained in purposive lists when they are available to improve the sample-based estimation of species richness, rather than to criticize the estimators based on the sole sample information that, when no list is available, provide useful lower bounds. Our proposal completely follows the spirit in which auxiliary information should be used in sample surveys [e.g., Särndal, Swensson and Wretman (1992), page 21]. The ED estimator of species richness exactly works in this direction.

APPENDIX: AN UPPER BOUND FOR THE MEAN SQUARED ERROR OF THE EMPIRICAL DIFFERENCE ESTIMATOR

The absolute difference between $\widehat{K}_E$ and K is bounded by

$$\begin{aligned}
|\widehat{K}_E - K| &= \left| \sum_{j\in \mathrm{C}} y_j^0 + \sum_{j\in \mathrm{C}-\mathrm{L}} \frac{1}{\hat{\tau}_j} I_j - \sum_{j\in \mathrm{C}} 1 \right| \\
&= \left| \sum_{j\in \mathrm{C}} y_j^0 + \sum_{j\in \mathrm{C}} \frac{1-y_j^0}{\hat{\tau}_j} I_j - \sum_{j\in \mathrm{C}} 1 \right| \\
&= \left| -\sum_{j\in \mathrm{C}} (1-y_j^0) + \sum_{j\in \mathrm{C}} \frac{1-y_j^0}{\hat{\tau}_j} I_j \right| = \left| \sum_{j\in \mathrm{C}} \left(\frac{I_j}{\hat{\tau}_j} - 1 \right) (1-y_j^0) \right| \\
&\le \sum_{j\in \mathrm{C}} \left| \frac{I_j}{\hat{\tau}_j} - 1 \right| (1-y_j^0),
\end{aligned}$$

where I_j is the indicator function of the event $j \in \mathrm{G}_{(n)}$. Therefore

$$\left| \frac{I_j}{\hat{\tau}_j} - 1 \right| = I_j \left(\frac{1}{\hat{\tau}_j} - 1 \right) + (1 - I_j)$$

from which

$$|\widehat{K}_E - K| \le \sum_{j\in \mathrm{C}} I_j \left(\frac{1}{\hat{\tau}_j} - 1 \right) (1-y_j^0) + \sum_{j\in \mathrm{C}} (1-I_j)(1-y_j^0).$$

Moreover, because

$$\hat{\tau}_j = 1 - \left(1 - \frac{x_j + 1}{n+1} \right)^n$$

increases with n and because for $n = 1$

$$\hat{\tau}_j = 1 - \left(1 - \frac{x_j + 1}{2} \right) \ge \frac{1}{2},$$

it generally holds that $\hat{\tau}_j \ge \frac{1}{2}$, from which

$$\begin{aligned}
|\widehat{K}_E - K| &\le 2 \sum_{j\in \mathrm{C}} I_j (1-\hat{\tau}_j)(1-y_j^0) + \sum_{j\in \mathrm{C}} (1-I_j)(1-y_j^0) \\
&\le 2 \sum_{j\in \mathrm{C}} (1-\hat{\tau}_j)(1-y_j^0) + \sum_{j\in \mathrm{C}} (1-I_j)(1-y_j^0) \\
&= 2 \sum_{j\in \mathrm{C}} \left(1 - \frac{x_j + 1}{n+1} \right)^n (1-y_j^0) + \sum_{j\in \mathrm{C}} (1-I_j)(1-y_j^0).
\end{aligned}$$

Finally, because $\log(1-x) < -x$ for $x < 1$, it follows that

$$\left(1 - \frac{x_j + 1}{n+1}\right)^n = \exp\left\{n \ln\left(1 - \frac{x_j+1}{n+1}\right)\right\} \le \exp\left\{-\frac{n(x_j+1)}{n+1}\right\}$$

from which

$$\begin{aligned}|\widehat{K}_E - K| &\le 2\sum_{j\in C} \exp\left\{-\frac{n(x_j+1)}{n+1}\right\}(1-y_j^0) + \sum_{j\in C}(1-I_j)(1-y_j^0) \\ &\le 2\sum_{j\in C} \exp\left(-\frac{1+x_j}{2}\right)(1-y_j^0) + \sum_{j\in C}(1-I_j)(1-y_j^0).\end{aligned}$$

Since $(a+b)^2 \le 2a^2 + 2b^2$, it follows that

$$(\widehat{K}_E - K)^2 \le 8\left[\sum_{j\in C} \exp\left(-\frac{1+x_j}{2}\right)(1-y_j^0)\right]^2 + 2\left[\sum_{j\in C}(1-I_j)(1-y_j^0)\right]^2$$

and, as $(\sum_{i=1}^n a_i)^2 \le n\sum_{i=1}^n a_i^2$

$$\begin{aligned}(\widehat{K}_E - K)^2 &\le 2K\left[4\sum_{j\in C}\exp\{-(1+x_j)\}(1-y_j^0)^2 + \sum_{j\in C}(1-I_j)^2(1-y_j^0)^2\right] \\ &= 2K\left[4\sum_{j\in C}\exp\{-(1+x_j)\}(1-y_j^0) + \sum_{j\in C}(1-I_j)(1-y_j^0)\right]\end{aligned}$$

from which

$$\begin{aligned}\mathrm{MSE}(\widehat{K}_E) &= \mathrm{E}\{(\widehat{K}_E - K)^2\} \\ &\le 2K\left[4\sum_{j\in C}\mathrm{E}\left\{\exp\left(-1-\sum_{i=1}^n z_{ji}\right)\right\}(1-y_j^0)\right. \\ &\quad \left. + \sum_{j\in C}\mathrm{E}(1-I_j)(1-y_j^0)\right] \\ &= 2K\left[4e^{-1}\sum_{j\in C}\mathrm{E}\left\{\exp\left(-\sum_{i=1}^n z_{ji}\right)\right\}(1-y_j^0)\right. \\ &\quad \left. + \sum_{j\in C}\{1-\mathrm{E}(I_j)\}(1-y_j^0)\right] \\ &= 2K\left[4e^{-1}\sum_{j\in C}\mathrm{E}\left\{\prod_{i=1}^n \exp(-z_{ji})\right\}(1-y_j^0) + \sum_{j\in C}(1-\tau_j)(1-y_j^0)\right].\end{aligned}$$

Because $z_{1j}, \ldots, z_{nj}$ are independent random variables, it follows that

$$\mathrm{MSE}(\widehat{K}_E) \le 2K\left[4e^{-1}\sum_{j\in\mathrm{C}}\prod_{i=1}^{n}\mathrm{E}\{\exp(-z_{ij})\}(1-y_j^0) + \sum_{j\in\mathrm{C}}(1-\theta_j)^n(1-y_j^0)\right].$$

Moreover, since each z_{ij} is a Bernoulli random variable with parameter θ_j, it holds that

$$\mathrm{E}\{\exp(-z_{ij})\} = e^{-1}\theta_j + 1 - \theta_j = 1 - \theta_j(1-e^{-1})$$

from which it ultimately holds that

$$\begin{aligned}\mathrm{MSE}(\widehat{K}_E) &\le 2K\left[4e^{-1}\sum_{j\in\mathrm{C}}\{1-\theta_j(1-e^{-1})\}^n(1-y_j^0)\right.\\ &\quad\left.+\sum_{j\in\mathrm{C}}(1-\theta_j)^n(1-y_j^0)\right]\\ &\le 2K\left[4e^{-1}\sum_{j\in\mathrm{C}}\{1-\theta_j(1-e^{-1})\}^n(1-y_j^0)\right.\\ &\quad\left.+\sum_{j\in\mathrm{C}}\{1-\theta_j(1-e^{-1})\}^n(1-y_j^0)\right]\\ &\le 2K(4e^{-1}+1)\sum_{j\in\mathrm{C}}\{1-\theta_j(1-e^{-1})\}^n(1-y_j^0).\end{aligned}$$

Acknowledgements. We thank Prof. Luca Pratelli from the Naval Academy of Livorno (Italy) for his suggestions on the whole paper and his support in the derivation of the results of the Appendix. We are also indebted to two anonymous reviewers for their comments that greatly improved the previous draft of this paper and for suggesting to incorporate purposive lists in the Chao2 and Lincoln-Petersen estimators.

SUPPLEMENTARY MATERIAL

Supplement to "Joining the incompatible: Exploiting purposive lists for the sample-based estimation of species richness" (DOI: 10.1214/17-AOAS1126SUPP; .zip). The Supplementary Material contains a table explaining the ecological meaning of some symbols adopted in the papers (Section SM1), details about the D estimator (Section SM2), the R code for computing the ED estimator and to estimate its RRMSE (Section SM3). Moreover, the txt file of the incidence data and the floristic list adopted to estimate the species richness of vascular plants in the Maremma Regional Park, Italy, is also available.

REFERENCES

ARRIGONI, P. V. (2003). The flora of the Maremma Natural Park (Tuscany, central Italy). *Webbia Journal of Plant Taxonomy and Geography* **58** 151–240.

BARABESI, L. and FATTORINI, L. (1998). The use of replicated plot, line and point sampling for estimating species abundances and ecological diversity. *Environ. Ecol. Stat.* **5** 353–370.

BUNGE, J. and FITZPATRICK, M. (1993). Estimating the number of species: A review. *J. Am. Stat. Assoc.* **88** 364–373.

CAYUELA, L., GOTELLI, N. J. and COLWELL, R. K. (2015). Ecological and biogeographic null hypotheses for comparing rarefaction curves. *Ecol. Monogr.* **85** 437–455.

CHAO, A. (1984). Nonparametric estimation of the number of classes in a population. *Scand. J. Stat.* **11** 265–270. MR0793175

CHAO, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43** 783–791.

CHAO, A. and COLWELL, R. H. (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. and LEE, M. (1992). Estimating the number of classes via sample coverage. *J. Am. Stat. Assoc.* **87** 210–217.

CHAO, A., GOTELLI, N. J., HSIEH, T. C., SANDER, E. L., MA, K. H., COLWELL, R. K. and ELLISON, A. M. (2014). Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies. *Ecol. Monogr.* **84** 45–67.

CHIARUCCI, A. (2012). Estimating species richness: Still a long way off! *J. Veg. Sci.* **23** 1003–1005.

CHIARUCCI, A., BACARO, G. and SCHEINER, S. M. (2011). Old and new challenges in using species diversity for assessing biodiversity. *Philos. T. Roy. Soc. B* **366** 2426–2437.

CHIARUCCI, A., ENRIGHT, N. J., PERRY, G. L. W., MILLER, B. P. and LAMONT, B. B. (2003). Performance of nonparametric species richness estimators in a high diversity plant community. *Divers. Distrib.* **9** 283–295.

CHIARUCCI, A., DI BIASE, R. M., FATTORINI, L., MARCHESELLI, M. and PISANI, C. (2018). Supplement to "Joining the incompatible: Exploiting purposive lists for the sample-based estimation of species richness." DOI:10.1214/17-AOAS1126SUPP.

COLWELL, R. K. (2013). EstimateS: Statistical estimation of species richness and shared species from samples. Version 9. User's Guide and application. Published at http://purl.oclc.org/estimates.

COLWELL, R. K. and CODDINGTON, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philos. T. Roy. Soc. B* **345** 101–118.

COLWELL, R. K., CHAO, A., GOTELLI, N. J., LIN, S. Y., MAO, C. X., CHAZDON, R. L. and LONGINO, J. T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblage. *J. Plant Ecol.* **5** 3–21.

CONTI, F., ABBATE, G., ALESSANDRINI, A. and BLASI, C., eds. (2005). *An Annotated Checklist of the Italian Vascular Flora*. Palombi, Roma.

CORMACK, R. M. (1989). Log-linear models for capture-recapture. *Biometrics* **45** 395–413.

D'ALESSANDRO, L. and FATTORINI, L. (2002). Resampling estimators of species richness from presence-absence data: Why they don't work. *Metron* **60** 5–19. MR1973845

DIEKMANN, M., KÜHNE, A. and ISERMANN, M. (2007). Random vs non-random sampling: Effects on patterns of species abundance, species richness and vegetation-environment relationships. *Folia Geobot.* **42** 179–190.

FATTORINI, L. (2006). Applying the Horvitz–Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika* **93** 269–278. MR2278082

FATTORINI, L. (2007). Statistical inference on accumulation curves for inventorying forest diversity: A design-based critical look. *Plant Biosyst.* **141** 231–242.

FATTORINI, L. (2009). An adaptive algorithm for estimating inclusion probabilities and performing Horvitz–Thompson criterion in complex designs. *Comput. Statist.* **24** 623–639.

FATTORINI, S. (2013). Regional insect inventories require long time, extensive spatial sampling and good will. *PLoS ONE* **8** e62118.

GASTON, K. J. (1996). Species richness: Measure and measurement. In *Biodiversity. A Biology of Numbers and Difference* (K. J. Gaston, ed.) 77–113. Blackwell Science, Oxford.

GOTELLI, N. J. and CHAO, A. (2013). Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. In *Encyclopedia of Biodiversity*, 2nd ed. (S. A. Levin, ed.) **5** 195–211. Elsevier Ltd, Waltham.

GOTELLI, N. J. and COLWELL, R. K. (2001). Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* **4** 379–391.

GOTELLI, N. J., ANDERSON, M. J., ARITA, H. T., CHAO, A., COLWELL, R. K., CONNOLLY, S. R., CURRIE, D. J., DUNN, R. R., GRAVES, G. R., GREEN, J. L., GRYTNES, J. A., JIANG, Y. H., JETZ, W., KATHLEEN LYONS, S., MCCAIN, C. M., MAGURRAN, A. E., RAHBEK, C., RANGEL, T. F., SOBERÓN, J., WEBB, C. O. and WILLIG, M. R. (2009). Patterns and causes of species richness: A general simulation model for macroecology. *Ecol. Lett.* **12** 873–886.

GREGOIRE, T. G. and VALENTINE, H. T. (2008). *Sampling Strategies for Natural Resources and the Environment*. Chapman & Hall, Boca Raton, FL.

HÉDL, R. (2007). Is sampling subjectivity a distorting factor in surveys for vegetation diversity? *Folia Geobot.* **42** 191–198.

HELLMANN, J. J. and FOWLER, G. W. (1999). Bias, precision, and accuracy of four measures of species richness. *Ecol. Appl.* **9** 824–834.

HELTSHE, J. F. and FORRESTER, N. E. (1983). Estimating species richness using the jackknife procedure. *Biometrics* **39** 1–11.

HOLDRIDGE, L. R., GRENKE, W. C., HATHEWAY, W. H., LIANG, T. and TOSI, J. A. (1971). *Forest Environments in Tropical Life Zones*. Pergamon Press, Oxford.

HORTAL, J., BORGES, P. A. V. and GASPAR, C. (2006). Evaluating the performance of species richness estimators: Sensitivity to sample grain size. *J. Anim. Ecol.* **75** 274–287.

HOWARD, P. C., VISKANIC, P., DAVENPORT, T. R. B., KIGENYI, F. W., BALTZER, M., DICKINSON, C. J., LWANGA, J. S., MATTHEWS, R. A. and BALMFORD, A. (1998). Complementarity and the use of indicator groups for reserve selection in Uganda. *Nature* **394** 472–475.

LEE, S. M. and CHAO, A. (1994). Estimating population size via sample coverage for closed capture-recapture models. *Biometrics* **50** 88–97.

MELO, A. S. (2004). A critique of the use of jackknife and related non-parametric techniques to estimate species richness. *Community Ecol.* **5** 149–157.

NICHOLS, J. D. and CONROY, M. J. (1996). Estimation of species richness. In *Measuring and Monitoring Biological Diversity. Standard Methods for Mammals* (D. E. Wilson, F. R. Cole, J. D. Nichols, R. Rudran and M. Forster, eds.) 226–234. Smithsonian Institution Press, Washington, DC.

OKSANEN, J., GUILLAUME BLANCHET, F., FRIENDLY, M., KINDT, R., LEGENDRE, P., MCGLINN, D., MINCHIN, P. R., O'HARA, R. B., SIMPSON, G. L., SOLYMOS, P., STEVENS, M. H. H., SZOECS, E. and WAGNER, H. (2016). vegan: Community ecology package. R package version 2.4-1. https://CRAN.R-project.org/package=vegan.

PALMER, M. W. (1990). The estimation of species richness by extrapolation. *Ecology* **71** 1195–1198.

PALMER, M. W. (1991). Estimating species richness: The second-order jackknife reconsidered. *Ecology* **72** 1512–1513.

PALMER, M. W., EARLS, P. G., HOAGLAND, B. W., WHITE, P. S. and WOHLGEMUTH, T. (2002). Quantitative tools for perfecting species lists. *Environmetrics* **13** 121–138.

PIELOU, E. C. (1977). *Mathematical Ecology*. Wiley, New York.

PIGNATTI, S. (1982). *Flora d'Italia*, *Vol.* 3, Edagricole edizioni.

SÄRNDAL, C. E. and LUNDSTRÖM, S. (2005). *Estimation in Survey with Nonresponse*. Wiley, New York.

SÄRNDAL, C. E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer, New York.

SEBER, G. A. F. (1982). *The Estimation of Animal Abundance*. Griffin, London.

SKOV, F. and LAWESSON, J. E. (2000). Estimation of plant species richness from systematically placed plots in a managed forest ecosystem. *Nord. J. Bot.* **20** 477–483.

SMITH, E. P. and VAN BELLE, G. (1984). Nonparametric estimation of species richness. *Biometrics* **40** 119–129.

THOMPSON, S. K. (2002). *Sampling*, 2nd ed. Wiley, New York.

WALTHER, B. A. and MOORE, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography* **28** 815–829.

WALTHER, B. A. and MORAND, S. (1998). Comparative performance of species richness estimation methods. *Parasitology* **116** 395–405.

WILSON, J. B., PEET, R. K., DENGLER, J. and PÄRTEL, M. (2012). Plant species richness: The world records. *J. Veg. Sci.* **23** 796–802.

XU, H., LIU, S., LI, Y., ZANG, R. and HE, F. (2012). Assessing non-parametric and area-based methods for estimating regional species richness. *J. Veg. Sci.* **23** 1006–1012.

A. CHIARUCCI
DEPARTMENT OF BIOLOGICAL, GEOLOGICAL, AND ENVIRONMENTAL SCIENCES
UNIVERSITY OF BOLOGNA
VIA IRNERIO 42
40126, BOLOGNA
ITALY
E-MAIL: alessandro.chiarucci@unibo.it

R. M. DI BIASE
DEPARTMENT FOR INNOVATION IN BIOLOGICAL, AGRO-FOOD AND FOREST SYSTEM
UNIVERSITY OF TUSCIA
VIA ANTONIO PACINOTTI 3
01100 VITERBO
ITALY
E-MAIL: dibiase.rm@gmail.com

L. FATTORINI
M. MARCHESELLI
C. PISANI
DEPARTMENT OF ECONOMICS AND STATISTICS
UNIVERSITY OF SIENA
PIAZZA SAN FRANCESCO 17
53100, SIENA
ITALY
E-MAIL: lorenzo.fattorini@unisi.it
marzia.marcheselli@unisi.it
caterina.pisani@unisi.it

Supplement to

JOINING THE INCOMPATIBLE: EXPLOITING PURPOSIVE LISTS FOR THE SAMPLE-BASED ESTIMATION OF SPECIES RICHNESS

By Alessandro Chiarucci, Rosa Maria Di Biase, Lorenzo Fattorini, Marzia Marcheselli and Caterina Pisani

SM1. Ecological meaning of some notations adopted in the paper

Symbol	Ecological meaning
$\mathcal{U}$	the set of individual plants in the assemblage under study
N	the number of individual plants in the assemblage, i.e. the size of $\mathcal{U}$
C	the set of species in the assemblage
K	the number of species in the assemblage, i.e. the size of C. This quantity is usually referred to as the species richness and denoted by S in ecological studies
$\mathcal{S}$	a sample of individual plants selected from $\mathcal{U}$ by one run of the sampling scheme adopted (e.g. one plot or transect)
G	the set of species detected by the sample $\mathcal{S}$
θ_j	the unknown probability of detecting the j-th species by one run of the sampling scheme, usually referred to as the detection or incidence probability in ecological studies
n	number of replications of the sampling scheme (e.g. number of plots or transects)
$\mathrm{G}_{(n)}$	the set of species detected by the n replications of the sampling scheme
SO_n	the number of species detected by the n replications of the sampling scheme, i.e. the size of $\mathrm{G}_{(n)}$ usually denoted by S_{obs} in ecological studies
τ_j	the unknown probability of detecting the j-th species by the n replications of the sampling scheme
z_{ij}	the indicator equal to 1 if the j-th species is detected in the i-th replication and equal to 0 otherwise, usually organized into a matrix of SO_n rows and columns and referred to as presence-absence or incidence data in ecological studies
z_j	number of replications in which the j-th species is detected, usually referred to as species incidence frequency in ecological studies

SM2. Preliminaries on the Difference Estimator. The basic idea of the paper is to adopt the difference (D) estimator in the framework of the estimation of species richness. To this purpose, it seems suitable to introduce the D estimator from a general point of view. Quoting from Särndal et al. (1992), let $\mathcal{U}$ be a finite population of N units and let y_j for $j \in \mathcal{U}$ be the population values of a survey variable Y. Suppose to be interested in the estimation of the population total

$$T = \sum_{j \in \mathcal{U}} y_j.$$

Suppose also to know the values y_j^0 for each $j \in \mathcal{U}$ of an auxiliary variable Y^0 which is presumed to be a proxy for the survey variable Y. In this case it is possible to exploit the auxiliary information fournished by Y^0 to estimate T. Exploiting the y_j^0s , T can be rewritten as

$$T = \sum_{j \in \mathcal{U}} y_j^0 + \sum_{j \in \mathcal{U}} e_j = T^0 + \sum_{j \in \mathcal{U}} e_j$$

where T^0 is the total of the auxiliary variables and $e_j = y_j - y_j^0$ is the error made when predicting y_j by means of y_j^0. Obviously, while T^0 is known, the total of errors is unknown and must be estimated from the sample using the Horvitz-Thompson (HT) criterion.

To this purpose, let $\mathcal{S}$ be a sample selected from $\mathcal{U}$ by means of a probabilistic scheme inducing first- and second-order inclusion probabilities π_j and π_{jh} $(h > j \in \mathcal{U})$. Then, the D estimator is given by

$$\hat{T}_D = T^0 + \sum_{j \in \mathcal{S}} \frac{y_j - y_j^0}{\pi_j}.$$

In practice the D estimator makes use of the HT criterion to estimate the error made when the total of the proxy variable is adopted to predict the total of the interest variable. Särndal et al (1992) note that $\hat{T}_D$ can be rewritten as

$$\hat{T}_D = T^0 + \sum_{j \in \mathcal{S}} \frac{y_j}{\pi_j} - \sum_{j \in \mathcal{S}} \frac{y_j^0}{\pi_j} = \hat{T}_{HT} + (T^0 - \hat{T}_{HT}^0)$$

where $\hat{T}_{HT}$ and $\hat{T}_{HT}^0$ are the HT estimators of T and T^0, respectively. From this expression it is apparent that the D estimator can be viewed as a correction of the HT estimator, performed on the basis of the additional information provided by the auxiliary variable. If the total of the auxiliary variable

is underestimated from the sample, the same sample is likely to provide also an underestimation of the total of the survey variable, hence the estimate $\hat{T}_{HT}$ is increased by the quantity $T^0 - \hat{T}^0_{HT}$. On the contrary, if the total of the auxiliary variable is overestimated from the sample, the same sample is likely to provide also an overestimation of the total of the survey variable, hence the estimate $\hat{T}_{HT}$ is decreased by the quantity $T^0 - \hat{T}^0_{HT}$.

The D estimator has some appealing properties. It is unbiased with variance

$$\mathrm{Var}(\hat{T}_D) = \sum_{j \in \mathcal{U}} \frac{1-\pi_j}{\pi_j} e_j^2 + \sum_{h>j \in \mathcal{U}} \frac{\pi_{jh} - \pi_j \pi_h}{\pi_j \pi_h} e_j e_h.$$

Moreover, if the proxy predicts the interest variable without error, i.e. if $y_j^0 = y_j$ for each $j \in \mathcal{U}$, the D estimator guesses the true total without error. Therefore, the D estimator is likely to be efficient when Y^0 is a good proxy for Y.

SM3. The R function `ederich( )`. The R function `ederich( )` returns the empirical difference estimate and the estimate of its relative standard error. The script of the function is reported in the following.

```
ederich=function(M,L) {
# M presence-absence matrix, dimension kxn
# k number of species
# n number of plots
# L floristic list, length n
n=ncol(M)
l=sum(L)
s=apply(M,1,sum)
# s: number of the occurrences in M for each species
data=data.frame(s)
names(data)[1]="occurrences"
data$ob_sp=ifelse(data$occurrences==0,0,1)
# ob_sp=1 if the i-th species is sampled and ob_sp=0 otherwise
data$list=L
# the L list is added to data
d=data[data$ob_sp-data$list==1,]
# new dataframe with the species in M but not in L
K_ED=l+sum(1/(1-(1-(1+d$occurrences)/(n+1))^n))
# species richness estimate
```

```
RRMSE=sqrt((2 * (4 * exp(-1)+1))/K_ED * sum(((1-(1-exp(-1))*
(d$occurrences+1)\(n+1))^n)\(1-(1-(d$occurrences+1)\(n+1))^n)))
#RRMSE estimate
est=c(K_ED,RRMSE)
names(est)=c("species richness estimate","RRMSE estimate")
return(est)}
```

References

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Särndal, C. E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.