Book of Short Papers SIS 2018

Editors: Antonino Abbruzzo - Eugenio Brentari

Marcello Chiodi - Davide Piacentino

SIS
2018
PALERMO

49 TH SCIENTIFIC MEETING
OF THE ITALIAN
STATISTICAL SOCIETY
20–22 JUNE

PUBLISHED BY PEARSON

WWW.PEARSON.COM

First printing, novembre 2018 - ISBN-9788891910233

Species richness estimation exploiting purposive lists: A proposal

Stima della ricchezza specifica utilizzando liste mirate. Una proposta

A. Chiarucci[1], R.M. Di Biase[2], L. Fattorini[3], M. Marcheselli[4] and C. Pisani[5]

Abstract The lists of species obtained by purposive sampling can be used to improve the sample-based estimation of species richness. A new estimator is 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. An asymptotically conservative estimator of the mean squared error is also provided.

Abstract *Le indagini mirate producono liste floristiche che possono essere usate per migliorare la stima della ricchezza specifica ottenuta tramite campionamento probabilistico. Viene qui proposto un nuovo stimatore della ricchezza specifica basato sullo stimatore per differenza, stimando le probabilità di inclusione sulla base dei dati di incidenza. Viene anche proposto uno stimatore asintoticamente conservativo dell'errore quadratico medio.*

Key words: Difference estimator, probabilistic sampling, purposive survey, supporting list, species richness.

[1] Alessandro Chiarucci, Department of Biological, Geological, and Environmental Sciences, University of Bologna; email: alessandro.chiarucci@unibo.it

[2] Rosa Maria Di Biase, Department for innovation in Biological, Agro-food and Forest systems, University of Tuscia; email: dibiase.rm@gmail.com

[3] Lorenzo Fattorini, Department of Economics and Statistics, University of Siena; email:lorenzo.fattorini@unisi.it

[4] Marzia Marcheselli, Department of Economics and Statistics, University of Siena; email:marzia.marcheselli@unisi.it

[5] Caterina Pisani, Department of Economics and Statistics, University of Siena; email:caterina.pisani@unisi.it

1 Introduction

Species richness, i.e. the number of species in a biological community, represents the simplest and most direct indicator of ecological diversity, largely used as the most convenient proxy for other components of biodiversity [6]. In the following, we will refer to plants, but analogous reasoning could be applied to limited mobility animals.

Species richness is an unknown parameter of the community under study, especially in large areas, and it can be evaluated by means of a purposive survey, as traditionally performed by ecologists, or estimated through probabilistic sampling.

In purposive sampling, species are recorded and listed by searching into specific sites expected to have a large number of species, high detection rates or high abundance of rare species. However, this is a subjective approach, so it does not allow any probabilistic statement about the accuracy and precision of species richness estimators.

On the other hand, in probabilistic sampling, species are identified and listed only if present in the selected samples. The estimates can be objectively evaluated through their sampling distributions, thus allowing for reliable comparisons across areas [e.g., 5]. Nevertheless, sample-based strategies *"are likely to miss the rare or unclassifiable habits that are likely to contribute most to regional diversity. ... Indeed, it is unlikely that such methods can outperform the guesses of experienced botanists"* [7, page 122]. Also in [7, page 122], the importance of probabilistic sampling in comparing species richness throughout time and space is recognized, even if *"it would be unwise to dismiss the efficient, yet subjective contributions of the expert botanist"*. However, at least to our knowledge, the species lists compiled by means of purposive surveys have never been used to improve species richness estimates arising from probabilistic sampling.

Therefore, we introduce a new species richness estimator, referred to as empirical difference (ED) estimator, exploiting both sources of information [2].

Section 2 contains some notations, while the ED estimator and a presumably asymptotic conservative estimator of its mean squared error are presented in Section 3. Section 4 contains a brief discussion on the improvement provided by list exploitation and concluding remarks.

2 Notation and setting

Consider a plant community within a delineated study area, so that each group of individual plants belonging to the same species may be viewed as a unit and the complete species list can be viewed as a population.

Referring to the species by their identifying numerical labels, the complete list of species can be represented by the set $\mathsf{C} = \left\{1, \ldots, K\right\}$. Since the complete list is usually unknown, the species richness K must be estimated. It should be noticed that species cannot be directly sampled because they constitute unknown assemblages of individual plants spread over the study area. Thus, the most effective way for sampling species is to sample individual plants, so that a species is sampled when at least one plant of that species is sampled.

The quantification of the first-order inclusion probabilities of species $\theta_1,\ldots,\theta_K$ entails the knowledge of all the units belonging to each species and their spatial distribution over the study area [5]. Consequently, even if the adopted sampling scheme ensures that the first-order inclusion probabilities can be determined directly or by some field measurements [e.g., 5] for (at least) the selected plants, the species inclusion probabilities cannot be quantified.

A study area cannot be adequately sampled by means of only one plot or transect and for this reason *n* independent replications of the sampling scheme [1] are usually performed, determining *n* samples of plants, which in turn give rise to *n* samples of species $\mathsf{G}_1,\ldots,\mathsf{G}_n$. The set of species observed in the whole survey is $\mathsf{G}_{(n)} = \bigcup_{i=1}^{n} \mathsf{G}_i$ and its size SO_n is the number of observed species.

For each replication *i*, $\mathbf{z}_i = \left[z_{i1},\ldots,z_{iK}\right]^{\mathsf{T}}$ is the *K*-vector in which the *j*th element z_{ij} is equal to 1 if the species *j* has been sampled and 0 otherwise. Usually, $\mathbf{z}_1,\ldots,\mathbf{z}_n$ are organized into a 0-1 matrix of *n* columns and SO_n rows, the *presence-absence* or *incidence data*. These *n* vectors are independent realizations of the random vector $\mathbf{Z} = \left[Z_1,\ldots,Z_K\right]^{\mathsf{T}}$ with expectation $\boldsymbol{\theta} = \left[\theta_1,\ldots,\theta_K\right]^{\mathsf{T}}$. Let $\mathbf{x} = \sum_{i=1}^{n} \mathbf{z}_i$ be the realization of the random vector $\mathbf{X} = \left[X_1,\ldots,X_K\right]^{\mathsf{T}}$, with $X_j \sim B(n,\theta_j)$. 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 virtually $\mathbf{x}$ is a *K*-vector, it contains an unknown number of zeros [3].

3 The empirical difference estimator

The most common way to exploit auxiliary information is the difference (D) estimator and its modifications, such as the widely used generalized regression and ratio estimators [8, chapter 6]. In this paper, the ED estimator is yet another modification of the D estimator which uses as auxiliary information the species list, henceforth referred to as supporting list L.

Consider *Y* and Y^0 dichotomous variables such that $y_j = 1$ for each species $j \in \mathsf{C}$ and $y_j^0 = 1$ if $j \in \mathsf{L}$ and 0 otherwise, so that Y^0 can be adopted as a proxy for the survey variable *Y*. When the supporting list is accurate, Y^0 is a good proxy for *Y*, given that the errors $y_j - y_j^0 = 1 - y_j^0$ are equal to 0 for any species $j \in \mathsf{L}$.

Exploiting the y_j^0 s, $K=\sum_{j\in \mathsf{C}} y_j$ can be rewritten [8, chapter 6] as

$$K=\sum_{j\in \mathsf{C}} y_j^0+\sum_{j\in \mathsf{C}} \left(y_j-y_j^0\right)=M+\sum_{j\in \mathsf{C}} \left(1-y_j^0\right)$$

where M is the number of species in the supporting list and the second member is unknown and needs to be estimated from the sample using the Horvitz-Thompson criterion

$$\hat{K}_D=M+\sum_{j\in G(n)} \frac{1-y_j^0}{\tau_j}=M+\sum_{j\in G(n)-\mathsf{L}} \frac{1}{\tau_j}$$

where $\tau_j=1-\left(1-\theta_j\right)^n$ is the probability that species j is detected during the whole sample survey. The estimator $\hat{K}_D$ is unbiased with a closed-form variance, which could be unbiasedly estimated from the sample. However, it cannot be calculated since the θ_j s (and consequently the τ_j s) are unknown. The frequencies x_j s in which the species enter the n samples can be adopted to estimate the θ_j s as $\hat{\theta}_j=\left(x_j+1\right)/(n+1)$ [4], from which $\hat{\tau}_j=1-\left(n-x_j\right)^n/(n+1)^n$.

The ED estimator turns out to be:

$$\hat{K}_E=M+\sum_{j\in G(n)-\mathsf{L}} \frac{1}{\hat{\tau}_j}$$

$\hat{K}_E$ is biased with expectation and variance which cannot be expressed in closed form. However, as opposite to other species richness estimators, its realizations are never smaller than the cardinality of the set $G_{(n)}\cup \mathsf{L}$. 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 $\mathsf{C}-\mathsf{L}$, i.e. the species lost in the supporting list which can be partially recovered by the sample survey. It should be noticed that if $G_{(n)}-\mathsf{L}$ is the empty set, the ED estimator coincides with M.

Because the estimator is biased, there is no sense in estimating its variance. Rather, we should estimate the relative root mean square error

$$\mathsf{RRMSE}=\frac{\sqrt{\mathsf{MSE}\left(\hat{K}_E\right)}}{K}=\frac{\sqrt{\mathsf{E}\left\{\left(\hat{K}_E-K\right)^2\right\}}}{K}$$

Because neither of the mean nor the variance of $\hat{K}_E$ can be expressed in a closed form, we derived an upper bound of $\mathsf{MSE}\left(\hat{K}_E\right)$ to be subsequently estimated from the available information, in such a way that the resulting estimator should be presumably asymptotically conservative. In [2, appendix] it is proved that

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

whose right side can be estimated by

$$\mathrm{M\hat{S}E}(\hat{K}_E) = 2\hat{K}_E(4e^{-1}+1)\sum_{j\in G_{(n)}-L}\frac{\left\{1-\hat{\theta}_j(1-e^{-1})\right\}^n}{\hat{\tau}_j}$$

Also in this case, if $G_{(n)}-L$ is the empty set, the $\mathrm{M\hat{S}E}$ turns out to be 0. Consequently,

$$\mathrm{RRM\hat{S}E}(\hat{K}_E) = \frac{\sqrt{\mathrm{M\hat{S}E}(\hat{K}_E)}}{\hat{K}_E} = \sqrt{\frac{2(4e^{-1}+1)}{\hat{K}_E}\sum_{j\in G_{(n)}-L}\frac{\left\{1-\hat{\theta}_j(1-e^{-1})\right\}^n}{\hat{\tau}_j}}$$

It should be noticed that the ED estimator is a consistent estimator of K with bias and variance approaching 0 as the number of replications increases. Indeed, from the $\mathrm{MSE}(\hat{K}_E)$ inequality follows that $\hat{K}_E$ converges in quadratic mean (and hence also in mean) to K, since $\lim_{n\to\infty} 2K(4e^{-1}+1)\sum_{j\in C}\left\{1-\theta_j(1-e^{-1})\right\}^n(1-y_j^0)=0$.

4 Discussion

In order to check the improvement provided by the exploitation of floristic lists, a simulation study was performed. $\hat{K}_E$ was compared to other species richness estimators, using nonparametric estimators of species richness.

The results of the simulation (for the full explanation of the simulation process and its results see [2]) show noticeable improvement in bias and precision provided by the proposed estimator, especially when the supporting list is accurate. 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. It should be noticed that lists missing the most common species are not unrealistic, because botanists are often focused on searching the rarest species, sometimes neglecting the most common ones [7].

If most of the rare species are included in the supporting list, the ED estimator can represent an efficient solution and the RRMSE estimator is presumably conservative, since some underestimations only occur when the supporting lists miss rare species. At least to our knowledge, this is the first attempt to estimate the MSE in species richness estimation, as most papers dealing with this problem propose estimators of the sampling variance. However, the estimation of variance in species richness estimation is irrelevant, because the major part of the sampling error is due to bias.

It is worth noting that it is important to work with updated floristic lists, so as to avoid the inclusion of species no longer in the area and other taxonomical problems (e.g. synonyms of species that were split into two or species that were merged). Thus, these drawbacks could deteriorate the performances of estimators exploiting floristic lists.

References

1. Barabesi, L., Fattorini, L.: The use of replicated plot, line and point sampling for estimating species abundances and ecological diversity. Environ. Ecol. Stat. (1998) doi: 10.1023/A:1009655821836
2. Chiarucci, A., Di Biase, R.M., Fattorini, L., Marcheselli, M., Pisani, C.: Joining the incompatible: Exploiting purposive lists for the sample-based estimation of species richness. Ann. Appl. Stat. Forthcoming
3. D'Alessandro, L., Fattorini, L.: Resampling estimators of species richness from presence-absence data: Why they don't work. Metron, **60**, 5-19 (2002)
4. Fattorini, L.: Applying the Horvitz–Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. Biometrika (2006) doi: 10.1093/biomet/93.2.269
5. Fattorini, L.: Statistical inference on accumulation curves for inventorying forest diversity: A design-based critical look. Plant Biosyst. (2007) doi: 10.1080/11263500701401786
6. Gaston, K.J.: Species richness: measure and measurement. In: Gaston, K.J. (ed.) Biodiversity. A Biology of Numbers and Difference, pp. 77–113. Blackwell Science, Oxford (1996)
7. Palmer, M.W., Earls, P.G., Hoagland, B.W., White, P.S., Wohlgemuth, T.: Quantitative tools for perfecting species lists. Environmetrics. (2002) doi: 10.1002/env.516
8. Särndal, C.E., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer, New York (1992)