

Sampling strategies for diversity indexes estimation in presence of rare species

Strategie di campionamento per la stima degli indici di diversità in presenza di specie rare

C. Pisani and L. Fattorini

Abstract The estimation of diversity indexes is considered when species abundance is estimated by means of plots thrown onto the study area in accordance with probabilistic schemes. In presence of rare species the sample diversity index estimators heavily underestimate their population counterparts. Under the uniform random sampling variance estimation and reduction of finite-sample negative bias are performed using jackknife. Despite its theoretical simplicity, uniform random sampling may lead to uneven coverage of the study area. To avoid the shortcoming, the use of stratified sampling is adopted. Large sample properties of diversity indexes estimators are considered under this alternative scheme as well as the use of jackknife to deal with negative bias.

Abstract *La stima degli indici di diversità è effettuata stimando l'abbondanza delle specie tramite plot collocati nell'area di studio sulla base di disegni campionari. In presenza di specie rare gli indici risultano fortemente sottostimati. Nel caso del campionamento casuale uniforme la stima della varianza e la riduzione della distorsione negativa avvengono attraverso l'uso del jackknife. A fronte della sua semplicità, questo campionamento può causare una copertura non uniforme dell'area di studio e pertanto risulta opportuno utilizzare il campionamento stratificato. Le proprietà per grandi campioni degli stimatori così ottenuti sono analizzate così come l'uso del jackknife per gestire la distorsione negativa.*

Key words: abundance estimation, rare species, uniform random sampling, stratified sampling, jackknife, simulation study

Caterina Pisani

University of Siena, P.zza S. Francesco 8, 53100 Siena e-mail: caterina.pisani@unisi.it

Lorenzo Fattorini

University of Siena, P.zza S. Francesco 8, 53100 Siena e-mail: lorenzo.fattorini@unisi.it

1 Introduction

The concept of ecological diversity relies on the apportionment of abundance into the animal or plant species forming the ecological community under study. Any diversity index is a function of the species abundance vector which is actually unknown and it must be estimated by means of a sample survey.

This paper deals with the estimation of diversity indexes when the species abundance vector is estimated by means of plots placed onto the study region in accordance with probabilistic schemes. The most simple scheme is the so-called Uniform Random Sampling (URS). Under URS, plots are uniformly and independently located giving rise to consistent and large-sample normally distributed diversity indexes estimators. In presence of rare species, diversity indexes estimators are highly negatively biased for a finite number of plots, and bias reduction is achieved by using jackknife (Barabesi and Fattorini, 1998). Despite its theoretical simplicity, URS may lead to uneven coverage of the study region which may cause a massive absence of the less abundant species in the sample. In order to overcome this drawback, a sampling scheme frequently adopted in environmental studies is the so-called one-per-stratum Stratified Sampling (SS). The SS scheme involves partitioning the study region into spatial subsets of equal size and uniformly and independently selecting a plot in each of these subsets.

This paper aims to discuss the results, quoted from Barabesi et al. (2014), on the large sample properties of the diversity indexes estimators under SS, as well as on the sampling variance estimation and the bias occurring for a finite number of plots.

2 Statistical properties

By assuming that k species are present in the population, let $\mathbf{N} = (N_1, \dots, N_k)^t$ be the corresponding abundance vector. Any diversity index can be viewed as a function of the abundance vector, say $\Delta(\mathbf{N})$. Denote by $\hat{\mathbf{N}}_n$ the mean of the Horvitz-Thompson estimators based on the n plots under URS. It can be proven that $\hat{\mathbf{N}}_n$ is unbiased with variance-covariance matrix $\text{var}(\hat{\mathbf{N}}_n)$ where $\max[\text{var}(\hat{\mathbf{N}}_n)] = O(n^{-1})$ and that $\text{var}[\hat{\mathbf{N}}_n]^{-1/2}(\hat{\mathbf{N}}_n - \mathbf{N}) \rightarrow N_k(\mathbf{0}, \mathbf{I})$ as $n \rightarrow \infty$. If Δ is differentiable at $\mathbf{N}$ and $\mathbf{d} \neq \mathbf{0}$, where $\mathbf{d}$ represents the gradient of Δ at $\mathbf{N}$, a straightforward application of the Delta method ensures that the properties of $\hat{\mathbf{N}}_n$ are asymptotically transferred to the diversity index estimator $\Delta(\hat{\mathbf{N}}_n)$, i.e. $\Delta(\hat{\mathbf{N}}_n)$ is asymptotically unbiased, consistent and normal with an asymptotical $O(n^{-1})$ variance which can be consistently estimated by $\hat{\sigma}_{\text{URS},n}^2$. However, even if $\Delta(\hat{\mathbf{N}}_n)$ is asymptotically unbiased, a bias occurs for finite samples. Barabesi and Fattorini (1998) consider the use of jackknife to reduce bias and to provide variance estimates. From the standard results on jackknife (e.g. Shao and Tu, 1995), if Δ is twice continuously differentiable at $\mathbf{N}$ and if $\mathbf{d} \neq \mathbf{0}$, denoting by $\hat{\Delta}_{n,jack}$ the jackknife estimator, it follows that $\hat{\sigma}_{n,jack}^{-1}(\hat{\Delta}_{n,jack} - \Delta(\mathbf{N})) \rightarrow N(0, 1)$, where $\hat{\sigma}_{n,jack}^2$ is the jackknife vari-

ance estimator which turns out to be consistent for $\sigma_{\text{URS},n}^2$. Moreover, it holds that $E[\hat{\Delta}_{n,jack}] = \Delta(\mathbf{N}) + o(n^{-1})$. Accordingly, under URS, $\hat{\Delta}_{n,jack}$ shares all the properties of $\Delta(\hat{\mathbf{N}}_n)$, but, while $\Delta(\hat{\mathbf{N}}_n)$ has a $O(n^{-1})$ bias, the bias of $\hat{\Delta}_{n,jack}$ is $o(n^{-1})$. Therefore, Barabesi and Fattorini (1998) recommend the use of $\hat{\Delta}_{n,jack}$ in order to achieve better centered confidence intervals and more reliable assessments of statistical hypotheses on $\Delta(\mathbf{N})$.

Under SS, denoting by $\tilde{\mathbf{N}}_n$ the mean of the Horvitz-Thompson estimators based on the n plots, by extending the results provided by Barabesi, Marcheselli and Franceschi (2012) for the estimation of a scalar parameter to the parameter vector $\mathbf{N}$, Barabesi et. al. (2014) prove that $\tilde{\mathbf{N}}_n$ is unbiased and that, under rather mild conditions on the shape of the spatial subsets partitioning the study region, $\max[\text{var}(\tilde{\mathbf{N}}_n)] = O(n^{-3/2})$ and $\text{var}[\tilde{\mathbf{N}}_n]^{-1/2}(\tilde{\mathbf{N}}_n - \mathbf{N}) \rightarrow N_k(\mathbf{0}, \mathbf{I})$ as $n \rightarrow \infty$. Moreover, the authors suggest the use of a large sample conservative estimator of $\text{var}[\tilde{\mathbf{N}}_n]$ based on “neighbours” plots.

The asymptotical properties of the diversity index estimator $\Delta(\tilde{\mathbf{N}}_n)$ and its bias reduction have been investigated by Barabesi et al. (2014) who show that, if the conditions for the large-sample normality of $\tilde{\mathbf{N}}_n$ are satisfied, the application of the Delta method ensures that $\sigma_{n,SS}^{-1}(\Delta(\tilde{\mathbf{N}}_n) - \Delta(\mathbf{N})) \rightarrow N(0, 1)$ as $n \rightarrow \infty$, where in this case $\sigma_{n,SS}^2 = \mathbf{d}' \text{var}[\tilde{\mathbf{N}}_n] \mathbf{d}$ represents the asymptotic variance of $\Delta(\tilde{\mathbf{N}}_n)$ and $\sigma_{n,SS}^2 = O(n^{-3/2})$. Moreover, an asymptotically conservative estimator $\tilde{\sigma}_{n,SS}^2$ for $\sigma_{n,SS}^2$ based on the “neighbours” plots is proposed. Practically speaking, analogously to the URS case, the properties of $\tilde{\mathbf{N}}_n$ are asymptotically transferred to $\Delta(\tilde{\mathbf{N}}_n)$. Indeed, $\Delta(\tilde{\mathbf{N}}_n)$ is consistent and large-sample normally distributed with a $O(n^{-3/2})$ variance. Thus, SS turns out to be superefficient with respect to URS even in the case of diversity index estimation. As to the bias order, Barabesi et al. (2014) prove the following result.

Theorem 1. *Let Δ be twice differentiable at $\mathbf{N}$ with bounded derivatives. Then,*

$$E[\Delta(\tilde{\mathbf{N}}_n)] = \Delta(\mathbf{N}) + O(n^{-3/2}). \quad (1)$$

Theorem 1 ensures that under SS the bias of $\Delta(\tilde{\mathbf{N}}_n)$ is $o(n^{-1})$ showing the superiority of SS with respect to URS also in term of bias. In order to deal with the finite sample bias of $\Delta(\tilde{\mathbf{N}}_n)$, Barabesi et al. (2014) consider the use of jackknife. To this purpose, denoting by $\tilde{\Delta}_{n,jack}$ the jackknife estimator corresponding to $\Delta(\tilde{\mathbf{N}}_n)$, the authors prove the following result.

Theorem 2. *If Δ is third-order differentiable at $\mathbf{N}$, it holds that*

$$E[\tilde{\Delta}_{n,jack}] = \Delta(\mathbf{N}) + O(n^{-1}). \quad (2)$$

On the basis of (2), it follows that $\tilde{\Delta}_{n,jack}$ has a $O(n^{-1})$ bias even if Δ is a third-order differentiable function at $\mathbf{N}$ and hence under SS the use of jackknife increases the order of the bias instead of reducing it. This drawback obviously depends on the fact that the n Horvitz-Thompson estimators obtained on the basis of the n plots are

independent, but not identically distributed, and they have different expectations. Thus, in contrast to a common tendency of statistical practice, care must be taken in the use of resampling procedures when the conditions under which they are able to work are not met.

3 Simulation study

In order to evaluate the performance of SS with respect to URS, a simulation study was performed (Barabesi et al. (2014)).

A square of size $1ha$ was assumed to be the study region. Within the region a community of $N = 1,000$ trees, partitioned into $k = 36$ species, was considered. Tree locations were generated in such a way that 6 species (3 of them constituted by 300, 200 and 100 trees, respectively, and the remaining 3 constituted by 20 trees each) were uniformly distributed onto the square. In order to simulate species association and species avoidance, 15 species (2 of them constituted by 50 trees each, 3 of them constituted by 20 trees each, and the remaining 10 constituted by 1 tree each) were uniformly distributed within a circle of radius $15m$ settled on the upper part of the quadrat. Similarly, 15 species with the same abundance were uniformly distributed within a circle of radius $15m$ settled in the lower part of the quadrat. It is worth noting that the simulated community was mainly composed by very rare species, since 20 out of the 36 species have abundance equal to 1. The Shannon diversity index $\Delta_{SH} = -\sum_{i=1}^k (N_i/N) \ln(N_i/N)$ and the Simpson diversity index $\Delta_{SI} = 1 - \sum_{i=1}^k (N_i/N)^2$ were considered to quantify diversity and turned out to be 2.35 and 0.85, respectively. Since in a community of 36 species it turns out that $0 \leq \Delta_{SH} \leq 3.58$ and $0 \leq \Delta_{SI} \leq 0.97$, the simulated forest showed a high level of ecological diversity due to the presence of the 20 rare species.

A plot sampling was simulated on the study region by using circular plots of radius $5m$. In order to avoid edge effects and to ensure the same inclusion probability to each tree in the forest, the study region was enlarged by a buffer of width $5m$. Under URS, $n = m^2$ (with $m = 5, \dots, 20$) sample points were uniformly and independently thrown onto the enlarged study region. For each sample point, the plot was constructed and the trees contained in the circle were sampled. Accordingly, the estimate of the Shannon and Simpson diversity indexes and the corresponding asymptotic variance estimate, i.e. the realizations of $\Delta(\hat{N}_n)$ and $\hat{\sigma}_{n,URS}^2$, were achieved along with the corresponding jackknife estimates, i.e. the realizations of $\hat{\Delta}_{n,jack}$ and $\hat{\sigma}_{n,jack}^2$. The estimates of the relative errors were achieved as the realizations of $\hat{\sigma}_{n,URS}/\Delta(\hat{N}_n)$ and $\hat{\sigma}_{n,jack}/\hat{\Delta}_{n,jack}$, respectively. The extremes of confidence intervals at nominal level of 0.95 were implemented as the realizations of $\Delta(\hat{N}_n) \pm z_{0.95} \hat{\sigma}_{n,URS}$ and $\hat{\Delta}_{n,jack} \pm z_{0.95} \hat{\sigma}_{n,jack}$, respectively. For each sampling effort n , the procedure was replicated 10,000 times to achieve the Monte Carlo distribution of the diversity indexes estimators and the corresponding variance estimators.

The whole simulation procedure was repeated under SS. In this case the enlarged study region was partitioned into $n = m^2$ (with $m = 5, \dots, 20$) squares of the same

size and a sample point was uniformly selected in each square. The simulation study was carried out similarly to the URS case.

The simulation results confirm the theoretical findings: under URS the jackknife estimator has to be preferred while the opposite conclusion occurred under SS. It should be remarked that bias reduction should be the main concern when estimating diversity indexes. Indeed, as pointed out by Särndal and Lundström (2005, p. 98), “if an estimator is greatly biased, it is poor consolation that its variance is low”. For such a reason, the performance of the strategy based on URS joined with jackknife estimator and of the strategy based on SS without jackknife correction are compared in terms of relative bias, relative root mean squared error, actual coverage of the confidence intervals at nominal level 0.95 and empirical expectation of the relative error estimators.

Under URS, the bias level of $\hat{\Delta}_{n,jack}$ becomes irrelevant (below 1%) for $n = 49$ when estimating the Shannon index and for $n = 25$ when estimating the Simpson index. On the other hand, by using $\Delta(\tilde{\mathbf{N}}_n)$, $n = 144$ and $n = 64$ plots are respectively necessary for estimating the Shannon and the Simpson indexes at the same level of bias. Even if URS invariably provides smaller bias than SS, the use of $\Delta(\tilde{\mathbf{N}}_n)$ turns out to be invariably more accurate with respect to the use of $\hat{\Delta}_{n,jack}$ - with an efficiency which increases from about 1.5 when $n = 25$ to about 3 when $n = 400$ for both indexes. Under URS and for both indexes, the jackknife variance estimator satisfactorily performs in terms of bias of the relative error estimator and coverage of confidence intervals for n above 100. Under SS and for both indexes, estimator $\tilde{\sigma}_{n,SS}^2$ becomes conservative in terms of bias and coverage for n above 144.

The practical conclusion deriving from both the theoretical findings and the simulation results is that the best strategy to perform design-based diversity index estimation consists in estimating $\mathbf{N}$ by means of the SS scheme and subsequently estimating $\Delta(\mathbf{N})$ by means of $\Delta(\tilde{\mathbf{N}}_n)$ - without further attempt to use jackknife for reducing bias.

References

1. BARABESI, L. AND FATTORINI, L. (1998). The use of replicated plot, line and point sampling for estimating species abundancies and ecological diversity. *Environ. Ecol. Stat.* **5** 353–370.
2. BARABESI, L., FATTORINI, L., MARCHESELLI, M., PISANI, C. AND PRATELLI, L. (2014). The estimation of diversity indexes by using stratified allocations of plots, points or transects. *Submitted*.
3. BARABESI L., FRANCESCHI, S. AND MARCHESELLI M. (2012). Properties of design-based estimation under stratified spatial sampling. *Ann. Appl. Statist.* **6** 210–228.
4. SÄRNDAL, C.E. AND LUNDSTRÖM, S. (2005). *Estimation in Survey with Nonresponse* Wiley, New York.
5. SHAO, J. AND TU, D. (1995). *The jackknife and the bootstrap*. Springer, Berlin.