Design-Based Inference on Diversity in Biological Populations

Inferenza basata sul disegno per l'analisi della diversità nelle popolazioni biologiche

Lorenzo Fattorini
Dipartimento di Metodi Quantitativi,
Università di Siena,P.za S. Francesco 8, 53100 Siena (Italy)
fattorini@unisi.it

Riassunto: il lavoro considera la misura della diversità di collettività animali o vegetali tramite opportuni indici e la stima degli stessi per mezzo di indagini campionarie basate sulla ripetizione di schemi di campionamento all'incontro. Sono trattati anche i problemi inerenti l'ordinamento della diversità e la stima del numero delle specie.

Keywords: Diversity indexes, Abundance estimation, Diversity ordering, Species richness.

1. Introduction

Even if a formal definition of diversity is still lacking, in statistical literature the diversity concept traditionally relies on the apportionment of abundance into the animal or plant species forming the ecological community under study. In turn, as pointed out by Pielou (1977, p.269), the term *community* refers to all the organisms in a delineated study area belonging to a taxonomic group of level higher than species. For a long time the primary aim of the statisticians faced with ecological diversity has been its quantification by means of suitable indexes which may take into account some important aspects of diversity such as *evenness*, *dominance* and *rarity* of species. However, depending on the apportionment of abundance among species, these indexes may only be known by means of a complete survey of the community under study, which is not feasible in most situations. Thus, a further statistical problem in analysing diversity lies in estimating species abundance on the basis of sample surveys in order to subsequently make inference on the diversity of the whole community. Unfortunately, on this topic results of practical relevance are lacking. Indeed, most papers devoted to inference on diversity indexes are based on the assumption that individuals are selected from the community by means of simple random sampling with replacement (SRSWOR). However, ecological communities rarely have a list frame, so it is not usually feasible or practical to obtain a sample of community members using SRSWOR. Usually the sampling schemes adopted by ecologists are *encounter schemes* (ES) in which the selected units are those *encountered* from points, lines or plots randomly thrown onto the study area. The purpose of this paper is to focus on those methodologies which make use of ES to estimate ecological diversity in a complete design-based framework. Accordingly, sections 2 and 3 are devoted to the preliminary problems of quantifying ecological diversity by means of suitable indexes and

estimating species abundance by using suitable strategies that take into account the nature of the community under study. Subsequently, in section 4, the design-based estimation of diversity indexes is considered. Finally, sections 5 and 6 are devoted to the peculiar problems of ordering communities according to diversity and estimating species richness, while some future developments for the design-based inference on diversity are considered in section 7.

2. The measurement of ecological diversity

The purpose of this section is not to review the huge literature on ecological diversity indexes (for detailed reviews on this topic see *e.g.* Dennis *et al*, 1979, Magurran, 1988, Frosini, 2003), but rather to point out the most relevant contributions to the traditional approach of measuring diversity together with some new promising proposals. Suppose an ecological community of N individuals partitioned into k species and denote by $\mathbf{p} = [p_1, \ldots, p_k]^{\mathrm{T}}$ the vector of relative abundance $p_l = N_l / N$ $(l = 1, \ldots, k)$ where N_l denotes the abundance of the species l. A very effective and unifying approach for measuring diversity is offered by Patil and Taille (1979a, 1982) on the basis of the intuition that a community is diverse when there is a large number of rare species. Accordingly, the authors propose measuring the rarity of each species and adopting the average community rarity as a diversity index. Thus, if $R_l(\mathbf{p})$ is the rarity of the species *l*, the diversity index turns out to be $\Delta = \sum_{l=1}^{k} p_l R_l(\mathbf{p})$. Patil and Taille (1979a, 1982) discuss the use of (a) *dichotomous-type diversity indexes*, which are obtained when the rarity measure of the species l depends only on its relative abundance, *i.e* $R_l(\mathbf{p}) = R(p_l)$; (b) *rank-type diversity indexes*, which are obtained when the rarity measure of the species *l* is a function of the (descending) rank of p_l on the relative abundance vector **p**. The most familiar diversity indexes belong to these two large families or rarity measures. For example, the Δ_β dichotomous family is obtained when $R(p_l) = (1 - p_l^{\beta}) / \beta$ $(\beta \geq -1)$ which in turn reduces to $k - 1$, or to the Shannon and Simpson indexes for $\beta = -1, 0, 1$, respectively. Alternatively, Patil and Taille (1979a, 1982) focus on a rank-type rarity measure which gives rise to the *right-tail sum family* of diversity indexes T_m $(m = 0, 1, \ldots, k)$, where T_m represents the relative abundance of the $k - m$ rarest species (with $T_0 = 1$ and $T_k = 0$). The plotting of T_m versus m gives provides the *right-tail sum diversity profile* which turns out to be convex and decreasing from 1 to 0. Contemporary to the seminal works by Patil and Taille (1979a, 1982), Rao (1982) proposes an axiomatization of diversity measures on the basis of their capability to split diversity between and within the sub-populations determined by a hierarchical classification of the community under study. Accordingly, the author defines diversity measures to be *perfect* if the diversity splitting can be carried out for multiply-classified communities of any order. Subsequently, Lau (1985) shows that any perfect diversity measure must be of the form $\Delta = \mathbf{p}^{\mathrm{T}} \mathbf{D} \mathbf{p}$ where **D** is a *k*-matrix whose *l*-*h* element d_{lh} denotes the difference (usually in a biological sense) between species l and h $(h \neq l = 1, \ldots, k)$. These indexes are referred to as *Rao's quadratic entropy* and surprisingly they had a limited impact on diversity studies until the work by Champely

and Chessel (2002) which in turn was stimulated by the Gore discussion of a paper by Solow and Polasky (1994). The discussant points out the practical importance of the quadratic entropy by emphasizing that "*If in one community a species is replaced by another with similar abundance but very different characteristics, traditional indices of diversity would be unaffected while intuition would suggest an increase of diversity.*" In this framework, Champely and Chessel (2002) propose the use of Euclidean metrics for quantifying distances between species, giving rise to a Euclidean diversity coefficient which involves geometrical interpretations and graphical representations of diversity. However, Izsak and Szeidl (2002) point out some anomalies of quadratic entropy showing that the problem of measuring diversity still remains very open.

3. The estimation of species abundance

It is at once apparent from the previous section that any diversity index is a function, say $\Delta(\mathbf{p})$, of the relative abundance vector $\mathbf{p}$ which, in turn, is a function of the abundance vector $\mathbf{N} = [N_1, \ldots, N_k]^{\mathrm{T}}$, being $\mathbf{p} = \mathbf{N}/(\mathbf{1}^{\mathrm{T}}\mathbf{N})$. However, knowledge of these quantities obviously requires a census of the ecological community under study, which is unfeasible in most cases. Accordingly, abundance is actually unknown and must be estimated by means of a sample survey in order to subsequently estimate Δ. Consider an ecological community on a delineated study area of size *A*, which usually constitutes a without-frame population of *N* organisms spread over the area. Owing to the lack of frame, the most effective schemes for sampling ecological populations differ from the traditional ones and their choice is mainly determined by practical considerations on the nature of the community to be sampled. For example, when dealing with plant populations, *floating plot sampling* is usually adopted while *Bitterlich sampling* and *line intercept sampling* are suitable for sampling tree and shrub communities, respectively. On the other hand, when dealing with animal communities, *line transect sampling* or *point transect sampling* should be performed to handle problems related to the elusive behaviour of animals. All these techniques have empirically developed in field investigations and have long been set apart from the core of the statistical world. More recently, some authors (*e.g.* De Vries, 1986, Thompson, 1992, Schreuder *et al.*, 1993, Overton and Stehman, 1995) have attempted to connect many of these methods with basic sampling theory as well as to focus on the practical advantages to be gained over traditional sampling strategy. Now, denote by S a sample of distinct units selected from the community using a suitable scheme. Referring to units by their labels, S actually represents a subset of the population labels $U = \{1,2,\ldots,N\}$. Obviously, any sampling strategy induces the *sampling design*, *i.e.* the probability distribution assigning $\Pr(S)$ for each $S \in \mathcal{S}$, where $\mathcal{S}$ denotes the family of the possible 2^N samples. When the population frame is available, the inclusion probabilities may be readily determined from the design. On the other hand, when handling without-frame ecological communities, the sampling design is unknown. In this case, the sampling schemes must be strictly ruled in order to determine (directly or by field measurements) the first-order inclusion probabilities at least for the selected units, which in turn allow for the computation of the Horvitz-Thompson estimator. For example, in floating plot sampling, a point is randomly thrown onto the study area and the selected units are those included in a circular or square plot of a pre-fixed size *a* centered at the sample

point. Accordingly, disregarding edge effects which can be removed by suitable modifications of the sampling scheme (see *e.g.* Schreuder *et al.*, 1993), all the inner units have a first-order inclusion probability equal to a/A. Moreover, as to Bitterlich sampling (also referred as to *variable circular plot sampling),* a point is randomly thrown onto the study area and a tree is selected if its bole at breast high subtends an angle greater than a pre-fixed angle α onto the point. In this case the first order inclusion probability of each tree turns out to be proportional to the bole area at breast height, which can be readily determined in the field by measuring the bole circumference at the same height. Alternatively, in line intercept sampling, a transect of fixed length is randomly thrown onto the study area and the selected units are those intercepted by the transect. In this case, Kaiser (1983) shows that the inclusion probability of a unit is the perimeter length of the minimum convex polygon enveloping the unit to the perimeter length of the minimum convex polygon enveloping the whole area. Alternatively, as suggested by Thompson (1992), a point may be thrown onto a baseline (*i.e* the shadow cast by the study area on a straight line lying outside) and a transect of fixed direction and random length is determined by the line starting from the selected point perpendicular to the baseline. In this case, the inclusion probability of a unit is simply the ratio of the length of the shadow cast by the unit onto the baseline to the baseline length. Finally, in line or point transect sampling, a line or a point is randomly thrown onto the area and the selected units are those spotted from it. In this case, the inclusion probabilities are evaluated on the basis of some simplifying assumptions adopted to model the sighting process (see *e.g.* Thompson, 1992, Barabesi and Fattorini, 1998). Thus, inference arising from line and point transect sampling cannot be considered entirely design-based. Without going into each of these sampling schemes here, the problem of estimating abundance by means of ES may be considered from a very general point of view. Let $\mathbf{e}_1,\ldots,\mathbf{e}_k$ be the standard basis of R^k and denote by $\mathbf{y}_1,\ldots,\mathbf{y}_N$ the marks associated with each individual in the community, where $\mathbf{y}_j = \mathbf{e}_l$ if individual *j* belongs to species *l*. As a result of this notation, the abundance vector $\mathbf{N}$ may be expressed as $\mathbf{N} = \sum_{j=1}^{N} \mathbf{y}_j$, in such a way that the estimation of $\mathbf{N}$ reduces to the estimation of a vector of population totals, which is a standard problem in finite population sampling. Thus, denote by $\pi_1,\ldots,\pi_N$ the first-order inclusion probabilities induced by the design. If the sampling scheme allows for the quantification of these probabilities at least for the selected individuals, the Horvitz-Thompson estimator $\hat{\mathbf{N}} = \sum_{j\in\mathsf{S}} \mathbf{y}_j/\pi_j$ constitutes an unbiased estimator for $\mathbf{N}$ with variance-covariance matrix $\mathbf{\Sigma}$ which also depends on the second-order inclusion probabilities $\pi_{jh}\,(h > j = 1,\ldots,N)$ as well as on some characteristics of the ecological community (such as the spatial distribution of the individuals over the study area). As to the estimation of $\mathbf{\Sigma}$, it is well known that an unbiased estimator for $\mathbf{\Sigma}$ exists if and only if $\pi_{jh} > 0$ for any $h > j = 1,\ldots,N$. Thus, problems arise when estimating $\mathbf{\Sigma}$ by means of a unique sample S, since in ES the individuals that are very far apart cannot be encountered jointly by the same plot, point or line. As emphasized by Barabesi and Fattorini (1998), these problems may be readily bypassed by replicating the ES. Indeed, a study area cannot be adequately sampled using one plot or one line or one point only. Accordingly, if the sampling procedure is independently replicated *n* times, in the sense that *n* plots or *n* lines or *n* points are randomly and independently thrown onto the study

area, then the n samples $S_1,\ldots,S_n$ provide n estimates $\hat{\mathbf{N}}_1,\ldots,\hat{\mathbf{N}}_n$ which are *iid* realizations of a random vector with expectation $\mathbf{N}$ and variance-covariance $\boldsymbol{\Sigma}$. Thus, their mean $\overline{\mathbf{N}}$ obviously constitutes an improved estimator for $\mathbf{N}$, with variance-covariance $\boldsymbol{\Sigma}/n$. Moreover, an unbiased and consistent $(n \to \infty)$ estimator for $\boldsymbol{\Sigma}$ may be straightforwardly obtained using the (unbiased) variance-covariance matrix of the n estimates, say $\mathbf{S}$. Finally, the straightforward use of the Central Limit Theorem ensures that $\sqrt{n}\boldsymbol{\Sigma}^{-1/2}(\overline{\mathbf{N}}-\mathbf{N}) \xrightarrow{d} N_k(\mathbf{0},\mathbf{I})$, while for the Delta method $\sqrt{n}\boldsymbol{\Xi}^{-1/2}(\hat{\mathbf{p}}-\mathbf{p}) \xrightarrow{d} N_k(\mathbf{0},\mathbf{I})$, where $\hat{\mathbf{p}} = \overline{\mathbf{N}}/(\mathbf{1}^{\mathrm{T}}\overline{\mathbf{N}})$ is the corresponding estimator of $\mathbf{p}$ and $\boldsymbol{\Xi} = (\mathbf{I}-\mathbf{p}\mathbf{1}^{\mathrm{T}})\boldsymbol{\Sigma}(\mathbf{I}-\mathbf{1}\mathbf{p}^{\mathrm{T}})/(\mathbf{1}^{T}\mathbf{N})^2$ may be consistently estimated by $\hat{\boldsymbol{\Xi}} = (\mathbf{I}-\hat{\mathbf{p}}\mathbf{1}^{\mathrm{T}})\mathbf{S}(\mathbf{I}-\mathbf{1}\hat{\mathbf{p}}^{\mathrm{T}})/(\mathbf{1}^{\mathrm{T}}\overline{\mathbf{N}})^2$. However, it is worth noting that the species richness k is often unknown and some rare species may not be present in the n samples. Hence, if *SO* denotes the number of species observed, the resulting estimates $\overline{\mathbf{N}}$ and $\hat{\mathbf{p}}$ are actually *SO*-vectors containing the estimates of the detected species, while $\mathbf{S}$ and $\hat{\boldsymbol{\Xi}}$ are square matrices both of order *SO*.

4. The estimation of diversity indexes

A vast number of results may be observed in literature regarding the estimation of diversity indexes when individuals are selected from the community using SRSWOR (from the early works revised in section III of the book by Grassle *et al* (1979) to the recent work by Chao and Shen, 2003). Pielou (1966) was the first to account for the actual field conditions, proposing the estimation of diversity indexes by means of a pooled quadrat sampling plan. Subsequently, Heyer and Berven (1973) extended Pielou's method to improve efficiency and to estimate the sampling variance, while Zahl (1977) proposed the use of the jackknife in plot sampling in order to estimate the Simpson index. Other jackknifing procedures in plot sampling have been investigated by Heltshe and Bitz (1979), Heltshe and Forrester (1983b) and Gove et al (1994) by means of simulation studies or field work. Barabesi and Fattorini (1998) give some general results on the jackknife estimation of diversity indexes when abundance is estimated by means of independent replications of an ES. The authors focus on the very large class of diversity indexes satisfying $\Delta(p_1,\ldots,p_k) = \Delta(p_1,\ldots,p_k,0,\ldots,0)$, in such a way that their estimation does not require the knowledge of k since the missing species may be ignored when computing $\hat{\Delta} = \Delta(\hat{\mathbf{p}})$. Moreover, the authors denote $\hat{\Delta}$ as $\Delta(\overline{\mathbf{N}})$ to emphasize the fact that they are dealing with a univariate transformation of the mean of the *n iid* estimates $\hat{\mathbf{N}}_1,\ldots,\hat{\mathbf{N}}_n$. Thus, by using very standard results (see e.g. Shao and Tu, 1995), if $g(\mathbf{x}) = \partial\Delta(\mathbf{x})/\partial\mathbf{x}$ exists in a neighbourhood of $\mathbf{N}$, is non-null and continuous at $\mathbf{N}$, then $\sqrt{n}\sigma^{-1}\left\{\hat{\Delta}-\Delta\right\} \xrightarrow{d} N(0,1)$ where $\sigma^2 = \mathbf{g}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{g}$ and $\mathbf{g} = g(\mathbf{N})$. Then the jackknife procedure, performed by deleting one replication at time, provides a variance estimator v_{jack}^2 which is consistent for σ^2/n in such a way that $v_{jack}^{-1}(\hat{\Delta}-\Delta) \xrightarrow{d} N(0,1)$. However, even if $\hat{\Delta}$ is asymptotically unbiased, a bias occurs for

finite samples. Indeed, it is a stated result that most diversity indexes, when evaluated from sample surveys, heavily underestimate the population counterpart. Thus, the jackknife procedure may be used not only to estimate variance but also to reduce bias and generate better-centered confidence intervals. Quoting again from the standard results by Shao and Tu (1995), Fattorini and Barabesi (1998) point out that if $\partial^2 \Delta(\mathbf{x}) / \partial \mathbf{x} \partial \mathbf{x}^{\mathrm{T}}$ exists and is continuous at **N**, then the $\hat{\Delta}$ has a $O(n^{-1})$ asymptotic bias, while the jackknife estimator $\hat{\Delta}_{jack}$ asymptotically achieves a $o(n^{-1})$ bias and $v_{jack}^{-1}(\hat{\Delta}_{jack} - \Delta) \xrightarrow{d} N(0,1)$. Note that these results hold for the most familiar diversity indexes but they do not hold for right-tail sum diversity indexes T_m $(m = 1, \ldots, k-1)$ when $N_l = N_h$ for some $l \neq h$ since T_m is not differentiable at **N**.

5. The ordering of ecological diversity

As pointed out by Hurlbert (1971), a single diversity index is not suitable for comparing ecological communities in that different indexes may lead to different rankings. In order to avoid inconsistent rankings, Patil and Taille (1982) introduce the *concept of intrinsic diversity ordering* defining a community C_2 to be intrinsically more diverse than a community C_1 ($C_1 \prec C_2$) if C_2 is obtained from C_1 through a finite sequence of the following operations: (*a*) transferring abundance from more to less abundant species without reversing the rank-order of the species; (*b*) transferring abundance to a new species; (*c*) re-labelling the species. Subsequently, the authors prove that any intrinsic diversity ordering, if it exists, can be determined by means of the right-tail sum diversity profiles $\{(m, T_m), m = 0, \ldots, k\}$ which are referred to as *intrinsic diversity profiles*. As a matter of fact, if $C_1 \prec C_2$ then the diversity profile of C_2 is everywhere above that of C_1. On the other hand, if the two profiles intersect one or more times, no intrinsic ordering of the two communities is possible. The proposal of Patil and Taille (1982) is further validated by Rousseau *et al.* (1999) who point out that any diversity ordering must take into account both evenness and species richness in such a way that: for equal species richness, evenness determines the order; for equal evenness, species richness determines the order; none of the two components is all-determining (balance property). In this framework, the authors prove that the partial order derived from intrinsic diversity profiles satisfies all these requirements and is the strongest among the partial orders which are proved to share the requirements. Since any intrinsic diversity profile is uniquely determined by the right-tail sum vector $\mathbf{T} = [T_1, \ldots, T_{k-1}]^{\mathrm{T}}$ which in turn is a function, say $t(\mathbf{N})$, of the unknown abundance vector **N**, once again the statistical problem lies in estimating the diversity profiles as functions of the estimated abundance. When the abundance estimates are obtained by means of independent replications of an ES, Fattorini and Marcheselli (1999) derive the asymptotic properties of the sample diversity profiles, say $\hat{\mathbf{T}} = t(\overline{\mathbf{N}}_n)$ under the assumption that $N_l = N_h$ for any $l \neq h$. Indeed, in this case the jacobian matrix $J(\mathbf{N})$ of the function *t* at **N** exists and differs from the null matrix, in such a way that $\sqrt{n}\boldsymbol{\Omega}^{-1/2}(\hat{\mathbf{T}} - \mathbf{T}) \xrightarrow{d} N_{k-1}(\mathbf{0}, \mathbf{I})$ where

$\boldsymbol{\Omega} = J(\mathbf{N})\boldsymbol{\Sigma}J(\mathbf{N})^{\mathrm{T}}$. Moreover, by deleting one replication at time, the jackknife variance estimator $\hat{\mathbf{V}}_{jack}$ turns out to be a consistent estimator for $\boldsymbol{\Omega}/n$. On the other hand, the jackknife procedure is not suitable for estimating $\mathbf{T}$ since $\hat{\mathbf{T}}_{jack}$ may fail to be convex. Finally, as to the construction of joint confidence bands, Fattorini and Marcheselli (1999), point out that since any intrinsic diversity profile is obtained by joining the $k+1$ points $\{(m,T_m), m=0,\ldots,k\}$, it actually represents a set of linear combinations of the component of $\hat{\mathbf{T}}$. Thus, since $\hat{\mathbf{T}}$ is asymptotically normal with expectation $\mathbf{T}$ and consistent estimates of its variance-covariance matrix are available, the Richmond (1982) technique may be applied to construct an asymptotically $(1-\alpha)$ conservative confidence band for the true profile. From a practical point of view, the band is bounded from below by joining the $(k+1)$ points $(0,1),(1,\hat{L}_1),\ldots,(k-1,\hat{L}_{k-1}),(k,0)$ and from above by joining $(0,1),(1,\hat{U}_1),\ldots,(k-1,\hat{U}_{k-1}),(k,0)$ where $\hat{L}_m$ and $\hat{U}_m$ equal $\hat{T}_m \pm \psi_{k-1,\alpha}\hat{v}_{mm}$ respectively, $\psi_{g,\alpha}$ is the upper $1-\alpha$ quantile of the studentized maximum modulus distribution with parameter g and ∞ degrees of freedom and $\hat{v}_{mm}^2$ is the (m,m)-element of $\hat{\mathbf{V}}_{jack}$. It is worth noting that Richmond's methods provides confidence bands that are narrower than those obtained using Scheffe's method. However, it must also be noticed that the species richness *k*, which should determine the order of $\hat{\mathbf{T}}$ and $\hat{\mathbf{V}}_{jack}$ as well as the quantile $\psi_{k-1,\alpha}$, usually constitutes an unknown parameter. Actually, $\hat{\mathbf{T}}$ and $\hat{\mathbf{V}}_{jack}$ turn out to be of order SO, while, since *SO* underestimates *k* (see section 6), the sample diversity profiles tend to be shorter than the population counterparts. Moreover, as to the determination of quantiles, a rule of thumb may be to choose *k* as the maximum number of species which might be present in the community. Indeed, since $\psi_{g,\alpha}$ slowly increases with *g*, this rule turns out to be conservative without entailing excessive enlargements of the confidence bands (note that the ratio $\psi_{g+500,\alpha}/\psi_{g,\alpha} < 1.2$ for $g > 15$). In order to rank populations according to their diversity, suitable hypotheses have to be assessed on the basis of the sample diversity profiles. For many years the procedure proposed by Patil and Taille (1979b) has been the unique method for assessing hypotheses on diversity profile. However Gove *et al.* (1994) point out that "*this procedure must be viewed as only an approximate test because it involves difficult and unresolved questions of simultaneous inference*". Alternatively, Fattorini and Marcheselli (1999) propose an asymptotically conservative procedure for comparing couples of diversity profiles by means of some methods previously adopted to make inference on Lorenz curves (Bishop *et al.*,1991). The procedure seems to be suitable in this framework since it is able to distinguish between the three possible outcomes of profile comparison, *i.e* dominance, equivalence and crossing. Denote by $\mathbf{T}_1$ and $\mathbf{T}_2$ the right-tail sum vector of populations C_1 and C_2 respectively. Then consider the equivalence hypothesis $H_0 : \mathbf{T}_1 = \mathbf{T}_2$ as the null one, against the alternative $H_1 : \mathbf{T}_1 \neq \mathbf{T}_2$. Obviously, H_0 implies that $\mathbf{T}_1$ and $\mathbf{T}_2$ have the same dimension $k-1$. By using the well-known union-intersection principle of test construction, H_0 may be decomposed as the intersection of $k-1$ hypotheses $H_{0m}\,(m=1,\ldots,k-1)$ regarding the

equality of the paired components of $\mathbf{T}_1$ and $\mathbf{T}_2$. Thus, denote by $\hat{\mathbf{T}}_1$ and $\hat{\mathbf{T}}_2$ the sample estimate of $\mathbf{T}_1$ and $\mathbf{T}_2$ obtained by means of n_1 and n_2 replications of two selected schemes, by $\boldsymbol{\Omega}_1 / n_1$ and $\boldsymbol{\Omega}_2 / n_2$ the variance-covariance matrices of $\hat{\mathbf{T}}_1$ and $\hat{\mathbf{T}}_2$, by $\hat{\mathbf{V}}_{1\,jack}$ and $\hat{\mathbf{V}}_{2\,jack}$ the respective jackknife estimates of these matrices and by SO_1 and SO_2 the species observed in the two communities. From the previous results by Fattorini and Marcheselli (1999), under H_0 and as $n_1, n_2 \to \infty$, $\boldsymbol{\Psi}^{-1/2}(\hat{\mathbf{T}}_1 - \hat{\mathbf{T}}_2) \xrightarrow{d} N_{k-1}(\mathbf{0}, \mathbf{I})$ where $\boldsymbol{\Psi} = \boldsymbol{\Omega}_1 / n_1 + \boldsymbol{\Omega}_2 / n_2$. Hence, by means of the quantities $Z_m = (\hat{T}_{m1} - \hat{T}_{m2}) / \sqrt{\hat{v}^2_{1mm} + \hat{v}^2_{2mm}}$ where $\hat{T}_{m1}$ and $\hat{T}_{m2}$ are the *m* components of $\hat{\mathbf{T}}_1$ and $\hat{\mathbf{T}}_2$ while $\hat{v}^2_{1mm}$ and $\hat{v}^2_{2mm}$ are the (m,m) elements of $\hat{\mathbf{V}}_{1\,jack}$ and $\hat{\mathbf{V}}_{2\,jack}$, Bishop *et al.* (1991) suggest the following conservative rule: accept H_0 if $|Z_m| \le \psi_{k-1,\alpha}$ for any $m = 1, \ldots, k-1$; reject H_0 and accept the dominance of $\mathbf{T}_1(\mathbf{T}_2)$ if there is at least one significant positive (negative) difference and no significant negative (positive) differences; reject H_0 and accept the crossing of the two profiles if there is at least one significant positive difference and one significant negative difference. Also in this case a problem arises since the value of k is unknown. Fattorini and Marcheselli (1999) suggest considering $\hat{\mathbf{T}}_1$ and $\hat{\mathbf{T}}_2$ of order $\max(S0_1, S0_2)$, thus setting at 0 the last component of the vector with the lower number of observed species. Moreover, as to the determination of the critical value $\psi_{k-1,\alpha}$ a suitable conservative solution is to choose *k* as the maximum number of species which might be present in the two communities. Fattorini and Marcheselli (1999) apply the proposed procedure for the mutual comparison of diversity for the avian populations settled in 11 parks in Milano and Pavia (Italy). Subsequently, Marcheselli (2003) considers the case in which $N_l = N_h$ for some $l \ne h$, which may occur, for example, in the presence of some rare species with unit abundance. The author proves that in this case the jackknife does not provide consistent estimators of $\boldsymbol{\Omega}$ and proposes an alternative conservative estimator based on a generalization of the Delta method.

6. The estimation of species richness

Species richness represents the simplest and most direct index of ecological diversity and it is often used as a convenient *proxy* for several aspects of biodiversity. The problem of estimating species richness has been studied for many years. Bunge and Fitzpatrick (1993) list more than 125 references on the topic. This section only focuses on the procedures based on *presence-absence* data, which bypass the estimation of species abundance and may be suitably applied under independent replications of an ES. Generally speaking, presence-absence data may be represented by an $n \times k$ matrix $[\mathbf{z}_1, \ldots, \mathbf{z}_n]$ where $\mathbf{z}_i = [z_{i1}, \ldots, z_{ik}]^{\mathrm{T}}$ denotes the vector in which $z_{il} = 1$ if at least one individual of species $l\,(l = 1, \ldots, k)$ is detected at occasion $i\,(i = 1, \ldots, n)$ and $z_{il} = 0$ otherwise. It is worth noting that the same kind of data also arise in capture-recapture experiments. Hence, the myriad of mark-recapture methodologies adopted to estimate

the size of a closed population may also be adopted in estimating species richness. However, most of these procedures are based on the assumption that the z_{il} are independent over all *i* and *l*, but while the independence between occasions may be ensured by field work, under realistic sampling schemes, the independence between species detections may sound like an oxymoron for any ecologist who is familiar with the concept of *inter-specific association.* Thus, more realistically, consider an ES and denote by θ_l the probability of detecting species *l* (which is given by the probability of sampling at least one individual of such species) and by θ_{lh} the probability of detecting species *l* and *h* jointly. Obviously, even if the ES is strictly ruled to quantify the inclusion probabilities of selected individuals, the inclusion probabilities of species are unknown, depending on their abundance as well as on their spatial distributions. However, under *n* independent replications of an ES, $\mathbf{z}_1,\ldots,\mathbf{z}_n$ constitutes *n iid* variables with expectation $\boldsymbol{\theta}=[\theta_1,\ldots,\theta_k]^{\mathrm{T}}$ and variance-covariance matrix $\boldsymbol{\Theta}=\boldsymbol{\Phi}-\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}$ where $\boldsymbol{\Phi}$ is a *k* square matrix having θ_{lh} as its *lh* element, with $\theta_{ll}=\theta_l$. Thus, their mean $\bar{\mathbf{z}}$ constitutes an unbiased estimator for $\boldsymbol{\theta}$ with variance-covariance $\boldsymbol{\Theta}/n$. Moreover $\sqrt{n}\boldsymbol{\Theta}^{-1/2}(\bar{\mathbf{z}}-\boldsymbol{\theta})\xrightarrow{d}N_k(\mathbf{0},\mathbf{I})$. As to the estimation of species richness, it is worth noting that $k=k(\boldsymbol{\theta})=\sum_{l=1}^{k}I(\theta_l>0)$, in such a way that the species observed may be rewritten as $SO=k(\bar{\mathbf{z}})=\sum_{l=1}^{k}I(\bar{z}_l>0)$. Obviously *SO* underestimates *k*. Thus, Heltshe and Forrester (1983a) propose the use of first-order jackknife to reduce the bias of SO while subsequently Smith and van Belle (1984) consider the use of second-order jackknife and bootstrap. Theoretical studies on the properties of these estimators have long been neglected. Recently, D'Alessandro and Fattorini (2002) have outlined the asymptotical inadequacy of these procedures to reduce bias since $k(\boldsymbol{\theta})$ has null derivatives at $\boldsymbol{\theta}$. Moreover, as to the finite-sample properties, the authors prove the inability of re-sampling procedures to reduce bias in the presence of species with very small inclusion probabilities. An alternative way to estimate *k* may be the fitting of the species accumulation curve $\boldsymbol{\gamma}=[\gamma_1,\ldots,\gamma_n]^{\mathrm{T}}$, where $\gamma_i=k-\sum_{l=1}^{k}(1-\theta_l)^i$ denotes the expected number of species observed in *i* occasions. Under the assumption that the z_{il} are independent, Colwell et al. (2004) prove that $\mathbf{c}=[c_1,\ldots,c_n]^{\mathrm{T}}$, where $c_i=SO-\binom{n_l}{i}^{-1}\sum_{n_l>0}\binom{n-n_l}{i}$, $\binom{a}{b}=0$ if $a<b$ and n_l denotes the number of occasions in which species *l* is detected, constitutes the minimum variance unbiased estimator for $\boldsymbol{\gamma}$ (note that the same result was previously proved by Smith and Grassle, 1977, when individuals are selected using SRSWOR). Thus a maximum likelihood estimation of *k* as well as of the parameters contained in $\boldsymbol{\gamma}$ may be performed together with an extrapolation of the curve with the objective of estimating the number of additional species that would be found by additional occasions. In order to avoid over-parametrizations, Colwell et al. (2004) suggest working with the more parsimonious curve $\gamma_i=k\left\{1-\sum_{h=1}^{H}\lambda_h(1-\theta_h)^i\right\}$ where *H* is the number of species having different inclusion probabilities and λ_h is the proportion of species having inclusion probability θ_h, with $\sum_{h=1}^{H}\lambda_h=1$. Confidence intervals for *k* and for extrapolated values of the

curve are obtained by a re-sampling procedure. Surprisingly, the same results on the efficiency of **c** as an unbiased estimator of $\boldsymbol{\gamma}$ and the same re-parametrization of the curve are independently considered by D'Alessandro (2003) under the more realistic situation in which *n* independent replications of an ES are adopted to sample the community. In this case the asymptotic normality of some selected component of **c**, say $c_{i1},\ldots,c_{im}$, is proven for fixed $i_1 < \ldots < i_m$ and $n \to \infty$, the consistency of the jackknife estimator of their variance-covariance matrix is also proven in such a way that k and the remaining $\boldsymbol{\gamma}$ parameters are estimated by non-linear generalized least-squares. Also in this case a re-sampling procedure is adopted to obtain confidence intervals for *k* and for the extrapolated values of the curve.

7. Future developments

As to the estimation of population totals by mean of replicated ES, Barabesi (2003) points out the equivalence of the procedure to a Monte Carlo integration. In this framework the random selection of points *n* over a baseline or over the whole study area is equivalent to the crude Monte Carlo integration and provides estimators with $O(n^{-1})$ variances. Thus a more attractive procedure consists of partitioning the baseline or the study region into N intervals or quadrats of equal size and generating a random point in each of them. In environmental sampling this scheme is referred to as *unaligned systematic sampling* while in the framework of integration techniques it is referred to as the modified Monte Carlo method. Barabesi and Pisani (2004) and Barabesi and Marcheselli (2005) show that unaligned systematic sampling provides improved estimators of totals with $o(n^{-1})$ variances. Accordingly, the unaligned systematic placement of points to estimate abundance and diversity indexes seems to be a promising strategy which requires further theoretical and empirical investigation. Moreover, as to the techniques for reducing the bias of diversity index estimators, Chao and Shen (2003) propose a method (which combines Horvitz-Thompson adjustment for missing species with the concept of sample coverage) which seems to outperform the jackknife procedures. Unfortunately, these conclusions are obtained under SRSWOR of individuals. Thus they should be generalized to more realistic sampling schemes. Finally, Dardanoni and Forcina (1999) extend the Bishop *et al.* (1991) procedure for the comparison of two Lorenz curves to the case of more than two curves. Accordingly, their procedure could be adapted, *mutatis mutandi*, to the comparison of more than two intrinsic diversity profiles.

References

Barabesi L. (2003) A Monte Carlo integration approach to Horvitz-Thompson estimation in replicated environmental designs, *Metron*, 61, 355-374.

Barabesi L and Fattorini L (1998) The use of replicated plot, line and point sampling for estimating species abundances and ecological diversity, *Environmental and Ecological Statistics*, 5, 353-370.

Barabesi L. and Marcheselli M. (2005) Some large-sample results on a modified Monte Carlo integration method, *Journal of Statistical Planning and Inference*, in press.

Barabesi L. and Pisani C. (2004) Steady-state ranked set sampling for replicated environmental sampling designs, *Environmetrics*, 15, 45-56.

Bishop J.A., Formby J.P. and Smith W.J. (1991) Lorenz dominance and welfare: changes in the U.S. distribution of income, 1967-1986, *The Review of Economic and Statistics*, 73, 134-139.

Bunge J. and Fitzpatrick M. (1993) Estimating the number of species: a review, Journal of the American Statistical Association, 88, 364-373.

Champely S and Chessel D. (2002) Measuring biological diversity using Euclidean metrics, *Environmental and Ecological Statistics*, 9, 167-177.

Chao A. and Shen T.J. (2003) Nonparametric estimation of Shannon's index of diversity when there are unseen species in the sample, *Environmental and Ecological Statistics*, 10, 429-443.

Colwell R.K., Mao C.X. and Chang J. (2004) Interpolating, extrapolating, and comparing incidence-based species accumulation curves, *Ecology*, 85, 2717-2727.

D'Alessandro L. (2003) *Inference on species accumulation curve*, Ph. D. Thesis, Università di Firenze, Florence (in Italian).

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.

Dardanoni V. and Forcina A. (1999) Inference for Lorenz curve orderings, *Econometric Journal*, 2, 49-75.

De Vries P.G. (1986) *Sampling Theory for Forest Inventories*, Springer-Verlag, Berlin.

Dennis B, Patil G.P., Rossi O., Stheman S and Taille C. (1979) A bibliography of literature on ecological diversity and related methodology. In *Ecological Diversity in Theory and Practice*, J.F. Grassle, G.P. Patil, W. Smith and C. Taille (Eds.), International Co-operative Publishing House, Fairland (MD), 319-354.

Fattorini L. and Marcheselli M. (1999) Inference on intrinsic diversity profiles of biological populations, *Environmetrics*, 10, 589-599.

Frosini B.V. (2003) Descriptive measures of ecological diversity. In *Environmetrics*, A.H. El-Shaarawi and J. Jureckova (Eds.), Encyclopedia of Life Support Systems (EOLSS), EOLSS Publishers, Oxford (UK) (http://www.eolss.net).

Gove J.H., Patil G.P., Swindel B.F. and Taille C. (1994) Ecological diversity and forest management, In *Handbook of Statistics*, Vol 12 (Environmental Statistics), G.P. Patil and C.R. Rao (Eds.), Elsevier, Amsterdam, pp. 409-462.

Grassle J.K., Patil G.P., Smith W. and Taille C. (1979) *Ecological Diversity in Theory and Practice*, International Co-operative Publishing House, Fairland (MD).

Heltshe J.F. and Bitz D.V. (1979) Comparing diversity measures in sampled communities, In *Ecological Diversity in Theory and Practice*, J.F. Grassle, G.P. Patil, W. Smith and C. Taille (Eds.), International Co-operative Publishing House, Fairland (MD), 133-144.

Heltshe J.F. and Forrester N.E. (1983a) Estimating species richness using the jackknife procedure, *Biometrics*, 39, 1-11.

Heltshe J.F. and Forrester N.E. (1983b) Estimating diversity using quadrat sampling, *Biometrics*, 39, 1073-1076.

Heyer R.V. and Berven K.A. (1973) Species diversity of herpetofaunal samples from similar microhabitats at two tropical sites, *Ecology*, 54, 642-645.

Hurlbert S.H. (1971) The nonconcept of species diversity: a critique and alternative parameters, *Ecology*, 52, 577-586.
Izsak J. and Szeidl L.(2002) Quadratic diversity: its maximization can reduce the richness of species, *Environmental and Ecological Statistics*, 9,423-430.
Kaiser L.(1983) Unbiased estimation in line-intercept sampling, *Biometrics*, 39, 965-976.
Lau K.S. (1985) Characterization of Rao's quadratic entropies, *Sankhya*, A 47, 295-309.
Magurran A.E. (1988) *Ecological Diversity and its Measurement*, Princeton University Press, Princeton (NJ).
Marcheselli M. (2003) Asymptotic results in jackknifing nonsmooth functions of the sample mean vector, *The Annals of Statistics*, 31, 1885-1904.
Overton W.S. and Stehman S.V. (1995) The Horvitz-Thompson theorem as a unifying perspective for probability sampling with examples from natural resource sampling, *The American Statistician*, 49, 261-268.
Patil G.P. and Taille C. (1979a) An overview of diversity. In *Ecological Diversity in Theory and Practice*, J.F. Grassle, G.P. Patil, W. Smith and C. Taille (Eds), International Co-operative Publishing House, Fairland (MD), 3-27.
Patil G.P. and Taille C. (1979b) A study of diversity profiles and ordering for a bird community in the vicinity of Colstrip, Montana, in "Contemporary Quantitative Ecology and Related Econometrics", G.P. Patil and M. Rosenzweig (Eds.), International Co-operative Publishing House, Fairland (MD), 23-48.
Patil G.P. and Taille C. (1982) Diversity as a concept and its measurement, *Journal of the American Statistical Association*, 77, 548-567.
Pielou E.C. (1966) The measurement of diversity in different types of biological collections, *Journal of Theoretical Biology*, 13, 131-144.
Pielou E.C. (1977) *Mathematical Ecology*, Wiley, New York.
Rao C.R. (1982) Diversity and dissimilarity coefficients: a unified approach, *Theoretical Population Biology*, 21, 24-43.
Richmond J. (1982) A general method for constructing simultaneous confidence intervals, *Journal of the American Statistical Association*, 77, 455-460.
Rousseau R., Van Hecke P. Nijssen D. and Bogaert J. (1999) The relationship between diversity profiles, evenness and species richness based on partial ordering, *Environmental and Ecological Statistics*, 6, 211-223.
Schreuder H.T., Gregoire T.G. and Wood G. (1993) *Sampling Methods for Multiresources Forest Inventories*, Wiley, New York.
Shao J. and Tu D. (1995) *The Jackknife and the Bootstrap*, Springer-Verlag, Berlin.
Smith W. and Grassle J.F. (1977) Sampling properties of a family of diversity measures, *Biometrics*, 33, 283-292.
Smith E.P. and van Belle G. (1984) Nonparametric estimation of species richness, *Biometrics*, 40, 119-129.
Solow A.R. and Polasky S. (1994) Measuring ecological diversity, *Environmental and Ecological Statistics*, 1, 95-107.
Thompson S.K. (1992) *Sampling*, Wiley, New York.
Zahl S. (1977) Jackknifing an index of diversity, Ecology, 58, 907-913.