

Conservative Confidence Bands for Intrinsic Diversity Profiles

Regioni di Confidenza Conservative per Profili Intrinseci di Diversità

Marzia Marcheselli
Dipartimento di Metodi Quantitativi
Università di Siena
marcheselli@unisi.it

Luca Pratelli
Accademia Navale
Livorno
pratel@mail.dm.unipi.it

Riassunto: Il profilo intrinseco di diversità è forse lo strumento più naturale e anche il più utilizzato per confrontare alcune popolazioni ecologiche rispetto alla loro diversità. Replicando in modo indipendente un opportuno disegno campionario (*e.g. line transect sampling, line intercept sampling, plot smpling*), la cui scelta dipende dalle caratteristiche della popolazione oggetto di studio, si stima il profilo intrinseco e si costruiscono le relative regioni di confidenza conservative.

Keywords: Intrinsic diversity profile, Delta method, replicated designs.

1. Introduction

A number of measures has been proposed to quantify the diversity of ecological populations. However it is well known that a single diversity index is not suitable for comparing communities given that the use of different indices may lead to different communities rankings. In order to compare ecological communities, the diversity of each community may be quantified by means of an *intrinsic diversity profile* which is defined as the plotting of pairs (j, T_j) , where

$$T_j = \sum_{l>j} \pi_{(l)}, \quad j = 0, 1, \dots, k$$

is the right-tail sum diversity index and $\pi_{(1)} \geq \pi_{(2)} \geq \dots \geq \pi_{(k)}$ are the ranked relative abundances of the k species which are present in the community. 'Species' is used as a convenient label for a finite set of distinct categories comprising the population. Obviously $T_0 = 1$ and $T_k = 0$. Thanks to intrinsic diversity profiles it can be stated that a population C is intrinsically more diverse than C' . According to Patil and Taillie (1982), population C is intrinsically more diverse than C' if C' leads to C through a finite sequence of abundance transfer from more to less abundant species without reversing the rank-order of the species, transferring abundance to a new species and re-labelling the species. As a matter of fact, the population C is intrinsically more diverse than C' if and only if the intrinsic diversity profile of C is above that of C' . Moreover, if the two profiles cross one or more times no intrinsic ordering of the compared populations is possible. Since any intrinsic diversity profile is a mapping of the unknown abundance vector, the statistical problem lies in estimating the abundance vector and subsequently, having obtained the diversity profile estimate as a function of the estimated abundances, constructing a confidence band for the profile. Section 2 contains some notations and the estimation of intrinsic diversity profile through the replicated use of a selected

sampling design. In Sections 3 an asymptotically conservative confidence band is constructed.

2. Profile estimation

Let $\mathbf{N} = (N_1, \dots, N_k)$ be the (unknown) abundance vector of a population located in a given study area. The estimation of $\mathbf{N}$ is a well-investigated problem in the framework of the design-based inference. As to the design choice, it should be pointed out that biological populations (*i.e.* plants, animals) rarely have a list frame, precluding the use of the usual sampling schemes. In these cases, the use of replicated encounter designs (*e.g.* *plot sampling*, *line intercept sampling*, *line transect sampling*) seems to be suitable (Thompson (1992), Barabesi and Fattorini (1998)). If n independent replications of an encounter sampling design are performed, n *iid* unbiased estimators $\hat{\mathbf{N}}_1, \dots, \hat{\mathbf{N}}_n$ for $\mathbf{N}$ may be obtained, each with variance-covariance matrix Σ (which may be consistently estimated by means of the sample variance-covariance matrix). Accordingly,

$$\bar{\mathbf{N}}_n = \frac{\hat{\mathbf{N}}_1 + \dots + \hat{\mathbf{N}}_n}{n}$$

is an unbiased, consistent estimator for $\mathbf{N}$, with variance-covariance matrix Σ/n . Moreover, owing to the Central Limit Theorem, $\sqrt{n}(\bar{\mathbf{N}}_n - \mathbf{N})$ converges in distribution to a centered Gaussian vector $\mathbf{V}$, with variance-covariance matrix Σ . Obviously the right-tail sum diversity indices are functions of $\mathbf{N}$; more precisely $T_j = g_j(\mathbf{N})$, where g_j is the function defined by

$$g_j(x_1, \dots, x_k) = \frac{1}{x_1 + \dots + x_k} \sum_{i=j+1}^k x_{(i)},$$

with $x_{(1)} \geq \dots \geq x_{(k)}$. Now, since any diversity profile is uniquely determined by the right-tail sum vector $\mathbf{T} = (g_1(\mathbf{N}), \dots, g_{k-1}(\mathbf{N}))$, a biased estimator for $\mathbf{T}$ is given by

$$\hat{\mathbf{T}}_n = (g_1(\bar{\mathbf{N}}_n), \dots, g_{k-1}(\bar{\mathbf{N}}_n)).$$

In order to study the asymptotic behaviour of $\mathbf{Y}_n = \sqrt{n}(\hat{\mathbf{T}}_n - \mathbf{T})$, it is necessary to introduce some notations. Let $\{n_1, \dots, n_s\}$ be the set of distinct components of $\mathbf{N}$, where $n_1 > \dots > n_s$. Obviously $s \leq k$ and $s = k$ only if all components of $\mathbf{N}$ are distinct. Moreover, let $J_1, \dots, J_s$ denote the disjoint subsets of $\{1, \dots, k\}$ so that $J_h = \{i: N_i = n_h\}$ and let d_h be the cardinality of J_h . Finally, let $(\Lambda_l)_{l \in D_h}$ be the random vector $(V_{(i)})_{i \in J_h}$, where $D_h = \{d_1 + \dots + d_{h-1} + l, l = 1, \dots, d_h\}$. In particular if $s = k$, J_h is the singleton $\{i\}$ so that $N_i = n_h$ and $D_h = \{h\}$, $\Lambda_h = V_i$. The random vector $\Lambda = (\Lambda_1, \dots, \Lambda_k)$ is Gaussian only if $s = k$. Since g_j is not necessarily (Hadamard) differentiable at $\mathbf{N}$, the delta method (Andersen *et al.*, 1997) cannot be applied to study the asymptotic behaviour of $\mathbf{Y}_n$. However, thanks to the generalized delta method (Marcheselli, 2000), the following result holds.

THEOREM. $\mathbf{Y}_n$ converges in distribution to the random vector $\mathbf{U} = (U_1, \dots, U_{k-1})$ where

$$U_j = \frac{1}{N_1 + \dots + N_k} (-g_j(\mathbf{N}) \sum_{i=1}^j \Lambda_i + (1 - g_j(\mathbf{N})) \sum_{i=j+1}^k \Lambda_i).$$

Since U_j is a linear combination of $\Lambda_1, \dots, \Lambda_k$, the random vector $\mathbf{U}$ is Gaussian if and only if the components of $\mathbf{N}$ are all distinct and, if $\mathbf{N}$ has at least two components that

are equal, it can only be stated that U_j is a Gaussian random variable when $j = d_1 + \dots + d_h$, with $h = 1, \dots, s$. It follows that the knowledge of Σ does not characterize the distribution of $\mathbf{U}$.

It is worth-while to observe that the jackknife variance for $\mathbf{Y}_n$ does not converge in distribution when $s < k$, and then the variance-covariance of $\mathbf{U}$ cannot be the limit of jackknife estimator for variance-covariance of $\mathbf{Y}_n$, when at least two components of $\mathbf{N}$ are equal! Moreover, the jackknife estimator for $\mathbf{Y}_n$ converges in distribution to $\mathbf{U}$ but, in general, its bias is not reduced with respect to the bias of $\mathbf{Y}_n$. Nevertheless, despite these difficulties, an asymptotic conservative confidence band for $\hat{\mathbf{T}}_n - \mathbf{T}$ can be given.

3. Confidence bands for intrinsic diversity profiles

In order to construct a conservative confidence bands, it is convenient to write in a suitable way the random variable U_j . More precisely, set

$$W_h = \sum_{l \in D'_h} \Lambda_l (= \sum_{i \in J'_h} V_i) \quad Z_h = W_s - W_h,$$

with $D'_h = \bigcup_{m=1}^h D_m$ and $J'_h = \bigcup_{m=1}^h J_m$, W_h and Z_h are symmetric Gaussian random variables and, for any $j \in D'_h$, it follows that

$$U_j = a_j W_{h-1} + b_j Z_h + a_j \sum_{l \in E_{h,j}} \Lambda_l + b_j \sum_{l \in D_h \setminus E_{h,j}} \Lambda_l$$

where $E_{h,j} = D_h \cap \{1, \dots, j\}$, $a_j = -g_j(\mathbf{N})/N$ and $b_j = (1 - g_j(\mathbf{N}))/N$, with $N = N_1 + \dots + N_k$. For any $j \in D_h$ and for any one-to-one function $\tau_{h,j}$ of D_h onto J_h , where $\tau_{h,j}$ is an increasing function on $E_{h,j}$, let

$$X_{\tau_{h,j}} = a_j W_{h-1} + b_j Z_h + a_j \sum_{l \in E_{h,j}} V_{\tau_{h,j}(l)} + b_j \sum_{l \in D_h \setminus E_{h,j}} V_{\tau_{h,j}(l)}.$$

It is not difficult to observe that U_j may be obtained suitably "pasting" the random variables of type $X_{\tau_{h,j}}$. Then, if $m_{\alpha,k^*,\infty}$ is the upper α -point of the studentized maximum modulus distribution with parameter $k^* = 2^k - 1$ and ∞ degrees of freedom (i.e. $m_{\alpha,k^*,\infty}$ is the upper α -point of $\max_{1 \leq m \leq 2^k - 1} |Y_m|$, with $Y_1, \dots, Y_{2^k - 1}$ independent $\mathcal{N}(0,1)$ random variables (Stoline and Ury, 1979)), the following result holds.

PROPOSITION. For any $\alpha \in]0, 1[$

$$P\left(\bigcup_h \bigcup_{j \in D_h} \{|U_j| > \sigma_{h,j} m_{\alpha,k^*,\infty}\}\right) \leq \alpha,$$

where $\sigma_{h,j}^2 = \sup_{\tau_{h,j} \in M_{h,j}} \text{Var}[X_{\tau_{h,j}}]$ and $M_{h,j}$ is the set of the one-to-one functions of D_h onto J_h , which are increasing on $E_{h,j}$.

For sake of simplicity, a conservative confidence band for the intrinsic diversity profile, which does not depend on the partition $\{J_h\}_h$, is now given when Σ is a diagonal matrix. Analogous results obviously hold for a generic matrix Σ . Let $\sigma_{(1)}^2 \geq \dots \geq \sigma_{(k)}^2$ be the ranked variances of $V_1, \dots, V_k$ (i.e. the ranked diagonal elements of Σ). For any j such that $g_j(\mathbf{N}) \geq 1/2$, let

$$s_j^2 = \frac{1}{N^2} [g_j^2(\mathbf{N}) \sum_{m=1}^j \sigma_{(m)}^2 + (1 - g_j(\mathbf{N}))^2 \sum_{m=j+1}^k \sigma_{(m)}^2]$$

and, for any j such that $g_j(\mathbf{N}) < 1/2$, let

$$s_j^2 = \frac{1}{N^2} [(1 - g_j(\mathbf{N}))^2 \sum_{m=1}^j \sigma_{(m)}^2 + g_j^2(\mathbf{N}) \sum_{m=j+1}^k \sigma_{(m)}^2].$$

It is easy to prove that $\sigma_{h,j}^2 \leq s_j^2$ for each $j \in D_h$ and for every $h=1 \dots s$. Hence, from the proposition, it follows that $P(\bigcup_j \{|U_j| > s_j m_{a,k^*,\infty}\}) \leq \alpha$.

The previous relation allows to obtain an asymptotic conservative confidence bands for the intrinsic diversity profile when Σ is a diagonal matrix. As a matter of fact, an asymptotically $(1-\alpha)$ conservative confidence band is given by

$$\{(j, t_j) : (\hat{\mathbf{T}}_n)_j - \frac{\hat{s}_j}{\sqrt{n}} m_{a,k^*,\infty} \leq t_j \leq (\hat{\mathbf{T}}_n)_j + \frac{\hat{s}_j}{\sqrt{n}} m_{a,k^*,\infty}, j = 1, \dots, k-1\},$$

where $\hat{s}_j^2$ is the random variable

$$\frac{1}{N_n^2} [\sup((\hat{\mathbf{T}}_n)_j^2, (1 - (\hat{\mathbf{T}}_n)_j)^2) \sum_{m=1}^j \hat{\sigma}_{(m)}^2 + \inf((\hat{\mathbf{T}}_n)_j^2, (1 - (\hat{\mathbf{T}}_n)_j)^2) \sum_{m=j+1}^k \hat{\sigma}_{(m)}^2],$$

$\hat{\sigma}_{(1)}^2 \geq \dots \geq \hat{\sigma}_{(k)}^2$ are the ranked consistent estimators of $\sigma_1^2, \dots, \sigma_k^2$, which are obtained by means of the sample variance-covariance matrix, k is the number of the observed species and $N_n = \sum_j (\bar{\mathbf{N}}_n)_j$. Note that, if the variances σ_j^2 are similar, the previous band is comparable with the analogous one which is obtained when $\mathbf{N}$ has all distinct components.

References

- Andersen P.K., Borgan O., Gill R.D., Keiding N. (1997) *Statistical models based on counting processes*, Springer.
- Barabesi L., 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.
- Marcheselli M. (2000) A generalized delta method with applications to intrinsic diversity profiles, *J. Appl. Prob.*, 37, 504-510.
- Patil G.P., Taillie C. (1982) Diversity as a concept and its measurement, *J. Amer. Statist. Assoc.*, 77, 548-567.
- Stoline M.R., Ury H.K. (1979) Tables of the studentized maximum modulus distribution and an application to multiple comparisons among means, *Technometrics*, 21, 87-93.
- Thompson S.K. (1992) *Sampling*, Wiley.