Statistical Methods & Applications (2004) 13: 129–136
DOI: 10.1007/s10260-003-0076-9

SMA

Simultaneous inference on diversity of biological communities

Tonio Di Battista, Stefano Antonio Gattone

Dipartimento di Metodi Quantitativi e Teoria Economica, Universita "G. d'Annunzio", Viale Pindaro, 42, 65127 Pescara, Italy

Abstract. In environmental statistics, surveys on the structure of biological communities are generally carried out by focusing on diversity indexes. A more complete analysis may be performed by means of an appropriate function giving a spectrum of different measures of diversity: diversity profiles. They can be expressed as a function of the unknown abundance vector of the ecological population. In this paper we develop a non parametric approach based on bootstrap in order to make inference on diversity profiles. The proposed procedure is applied on biological data of four parks in Milan, Italy.

Key words: β diversity profiles, bootstrap root estimate, balanced confidence regions, replicated sampling

1. Introduction

In many environmental and ecological surveys the aim is to evaluate the diversity of biological communities. Community diversity, diversity measurement and diversity comparisons have been the focus of ecological studies for more than four decades (for a review, see Gove et al., 1994). Many diversity indices which incorporate both the species richness and the degree of evenness within the species have been developed. Generally, the main indices can be obtained as a vector function of the relative abundance vector $\mathbf{p} = (p_1, p_2, ..., p_s)$ with $\mathbf{p} = N^{-1}\mathbf{N}$ where the i-th element of vector $\mathbf{N}$ represents the number of units belonging to the i-th species and $N = \mathbf{1}\mathbf{N}^T$ is the total number of individuals present in the population.

In particular, Patil and Taillie (1982) developed a measure of diversity known as β-diversity profile given by:

$$\boldsymbol{\Delta}(\mathbf{p}) = \{\Delta_\beta(\mathbf{p}) = \frac{1 - \sum_{i=1}^{s} p_i^{\beta+1}}{\beta} : \beta \in \mathbf{B}\}. \tag{1}$$

For $\beta \geq -1$ we have the one-parameter family of β-diversity indices. We emphasize that (1) is a curve made up of diversity indices. In particular, the Species Count, the Shannon index and the Simpson index are special cases of the β-diversity profile with $\beta = -1, 0$ and 1, respectively.

The parameter β represents the sensibility of the index to rare or common species. Thus, a plot of $\Delta_\beta(\mathbf{p})$ versus β provides simultaneous values of a large collection of indices showing different aspects of diversity in a single picture. As a matter of fact, changes or differences in community diversity can be studied by comparing profiles.

Another approach to evaluate diversity profiles is given by the Good family of diversity indices (Good, 1953), *i.e.*

$$\Delta_{\alpha,\beta} = \sum_{i=1}^{s} p_i^{\alpha}(-\ln p_i)^{\beta} \tag{2}$$

where α and β are taken to be non-negative integers. It is worth noting that, even in this case, $\Delta_{1,1}$ and $\Delta_{2,0}$ are the Shannon and Simpson index, respectively. Starting from (2), Baczkowski and Shamia (1998) further generalize the Good's family determining the range of (α, β) values and their biological meaning.

The use of (1) and (2) is motivated because of it is not possible to recommend a single index as superior to all the others and choosing the appropriate index depends on what sort of question is being asked. Furthermore, studies dealing with the topic of communities comparison by means of diversity indices reached the overwhelming conclusion that no single diversity index adequately summarizes community structure (i.e. see Hurlbert, 1971). Thus, a parametric family of indices whose members have varying sensitivities to the rare and abundant species may be used.

Since in the literature there doesn't seem to be the concern of inferring communities diversity profiles, in this paper we develop an inferential procedure based on simultaneous confidence intervals in order to compare communities according to their diversity. In particular, we wish to construct a statistical test in order to compare the β diversity profile of two communities. However, when diversity is evaluated by means of $\Delta_\beta(\mathbf{p})$, tests of hypotheses involve unresolved aspects of multivariate analysis and simultaneous inference because of difficulties in assuming any specific distribution both for the relative abundance vector and for each component of $\Delta_\beta(\mathbf{p})$. Furthermore, components of $\Delta_\beta(\mathbf{p})$ are dependent. In this paper, we give a solution to these problems by suitably using a nonparametric technique (Beran, 1988) adopted to build simultaneous confidence sets of a family of parametric functions. This procedure requires the estimator to be a linear function of the components of the mean vector in order to have reasonable asymptotically properties. At this purpose, we suggest the first-order Taylor expansion of (1) in order to achieve the β diversity profile as a linear combination of the entries of the unknown relative abundance vector $\mathbf{p}$ followed by a non parametric methodology based on bootstrap which tests the presence of an intrinsic diversity ordering between two communities under study. Finally, the developed procedure is applied to a real data set.

2. Statistical inference on β diversity profiles

In what follows, we propose a non parametric procedure based on bootstrap in order to derive a hypothesis test for comparing β diversity profiles of two biological communities. The key idea is to build a simultaneous confidence set for the parameter "difference between profiles" by means of the Beran's procedure (Beran, 1988). Hence, the conditions in order to apply this procedure on β diversity profiles have to be provided. Since the method applies on linear combinations of the mean vector, we suggest the use of the first-order Taylor expansion of β-diversity profiles.

In general setting, Beran's procedure aims to construct a simultaneous confidence set for a family of parametric functions of a parameter $\boldsymbol{\lambda}$, say

$$\mathbf{T}(\boldsymbol{\lambda}) = \{T_{\mathbf{u}}(\boldsymbol{\lambda}) : \mathbf{u} \in \mathbf{U}\} \tag{3}$$

with components labelled by an index set $\mathbf{U}$. The parametric function $T_{\mathbf{u}}(\boldsymbol{\lambda})$ is the $\mathbf{u}$-th component of the set $\mathbf{T}(\boldsymbol{\lambda})$. In particular, let $x_n = \{\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_n\}$ be a sample of *i.i.d.* random s-dimensional vectors with vector mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, then, if $T_{\mathbf{u}}(\boldsymbol{\lambda})$ is a linear function of $\boldsymbol{\mu}$, Beran proves (Theorem 4.1 pag. 684) that the simultaneous confidence set has an overall coverage probability $1 - \alpha$ for $\mathbf{T}(\boldsymbol{\lambda})$. Moreover, it is asymptotically balanced, that is, the coverage probability for each component $T_{\mathbf{u}}(\boldsymbol{\lambda})$ remains unchanged as $\mathbf{u}$ varies.

In our framework, let C_1 and C_2 be two communities under study with relative abundance vector $\mathbf{p}_1$ and $\mathbf{p}_2$ and β diversity profile $\Delta(\mathbf{p}_1)$ and $\Delta(\mathbf{p}_2)$ respectively. It is generally accepted that if $\Delta(\mathbf{p}_1) \geq \Delta(\mathbf{p}_2)$, that is, if the profile of community C_1 lies uniformly above that of community C_2, then C_1 is more intrinsically diverse than C_2 (dominance). While, if $\Delta(\mathbf{p}_1) = \Delta(\mathbf{p}_2)$, then there is no difference in diversity between C_1 and C_2 (equivalence). Conversely, if the profiles intersect, then the communities are not intrinsically comparable (intersection).

In this paper, the aim is to test equivalence, dominance or intersection hypotheses of the β diversity profiles. In particular, a suitable null hypothesis is

$$H_0 : \Delta(\mathbf{p}_1) = \Delta(\mathbf{p}_2)$$

that is, diversity is the same for each community (equivalence). H_0 may be viewed as the intersection of simple hypotheses of type $H_{0,\beta} : \Delta_\beta(\mathbf{p}_1) = \Delta_\beta(\mathbf{p}_2)$ for each $\beta \in \mathbf{B}$. Furthermore, let

$$\boldsymbol{\theta} = \Delta(\mathbf{p}_1) - \Delta(\mathbf{p}_2)$$

be the vector of the difference in the diversity profile of the two communities. The equivalence hypothesis of the two profiles may be written as

$$H_0 = \boldsymbol{\theta} = \boldsymbol{\theta}_0 = \mathbf{0}. \tag{4}$$

A few attempts have already been proposed in order to asses simple hypothesis for a single diversity index as $H_{0,\beta} : \Delta_\beta(\mathbf{p}_1) = \Delta_\beta(\mathbf{p}_2)$ (Basharin,1959; Baczkowski,2000). However, there does not seem to be in the literature a suitable statistical test in order to simultaneously compare several indices of diversity.

Therefore, let C be a community with relative abundance vector $\mathbf{p}$ and let $x_n = \{\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2, ..., \hat{\mathbf{p}}_n\}$ be a sample of *i.i.d.* random s-dimensional relative abundance vectors drawn with replacement from C. An estimator for $\mathbf{p}$ is given by

$$\bar{\mathbf{p}} = \frac{1}{n} \sum \hat{\mathbf{p}}. \tag{5}$$

The estimator $\bar{\mathbf{p}}$ obviously represents an unbiased and consistent estimator for $\mathbf{p}$ with covariance matrix $\mathbf{V}$. Accordingly, a consistent estimator for $\mathbf{V}$ is given by

$$\bar{\mathbf{V}} = [\mathbf{I} - diag(\bar{\mathbf{p}})]\mathbf{S}[\mathbf{I} - diag(\bar{\mathbf{p}})]^T / \bar{N}^2 \tag{6}$$

where $\mathbf{S}$ and $\bar{N}$ are straightforward estimators of the covariance matrix of the abundance vector and the population total, respectively.

Then, assuming that for $\beta \in \mathbf{B}$ the first derivatives (see Tong, 1983)

$$\Delta'_\beta(\mathbf{p}) = \frac{\partial}{\partial p_i} \Delta_\beta(\mathbf{p}) \quad i = 1, 2, ..., s,$$

are defined in a neighborhood of $\mathbf{p}$ and are continuous and non-null at $\mathbf{p}$, we can write the β diversity profile estimator of (1) as follows:

$$\Delta(\bar{\mathbf{p}}) = \{\Delta_\beta(\mathbf{p}) + \Delta'_\beta(\mathbf{p})(\bar{\mathbf{p}} - \mathbf{p})^T + o_p\|(\bar{\mathbf{p}} - \mathbf{p})\| : \beta \in \mathbf{B}\}$$

as $\bar{\mathbf{p}} \to \mathbf{p}$.

From the consistency of $\bar{\mathbf{p}}$, we have that for large n

$$\Delta(\bar{\mathbf{p}}) = \mathbf{W} + \Delta'(\mathbf{p})\bar{\mathbf{p}}^T + o_p(n^{1/2}) \tag{7}$$

where $\mathbf{W} = \Delta(\mathbf{p}) - \Delta'(\mathbf{p})\mathbf{p}^T$ is a constant.

Similarly, let x and y be two samples of respective sizes n_1 and n_2 drawn with replacement from communities C_1 and C_2. At each draw the estimates of the relative abundance vectors are given by $\hat{\mathbf{p}}_{h_1}$ and $\hat{\mathbf{p}}_{h_2}$ ($h_1 = 1, ..., n_1$; $h_2 = 1, ..., n_2$). Since $\hat{\mathbf{p}}_{h_1}$ and $\hat{\mathbf{p}}_{h_2}$ are *iid* random variables, then using (5) and (6), unbiased and consistent estimators for $\mathbf{p}_1$ and $\mathbf{p}_2$ are given by $\bar{\mathbf{p}}_1$ and $\bar{\mathbf{p}}_2$, while $\bar{\mathbf{V}}_1$ and $\bar{\mathbf{V}}_2$ are the covariance matrix estimates of $\mathbf{V}_1$ and $\mathbf{V}_2$, respectively.

Using (7) we can express the β diversity profiles estimates of the two communities C_1 and C_2, say $\Delta(\bar{\mathbf{p}}_1)$ and $\Delta(\bar{\mathbf{p}}_2)$, as the first-order Taylor expansions about $\bar{\mathbf{p}}_1$ and $\bar{\mathbf{p}}_2$:

$$\begin{aligned} \Delta(\bar{\mathbf{p}}_1) &= \mathbf{W}_1 + \Delta'(\mathbf{p}_1)\bar{\mathbf{p}}_1^T + o_p(n_1^{1/2}) \\ \Delta(\bar{\mathbf{p}}_2) &= \mathbf{W}_2 + \Delta'(\mathbf{p}_2)\bar{\mathbf{p}}_2^T + o_p(n_2^{1/2}). \end{aligned}$$

Under the null hypothesis H_0 in (4) and for large n_1 and n_2 an estimate of $\boldsymbol{\theta}$ is given by

$$\bar{\boldsymbol{\theta}} = \Delta'(\mathbf{p}_1)(\bar{\mathbf{p}}_1 - \mathbf{p}_1)^T - \Delta'(\mathbf{p}_2)(\bar{\mathbf{p}}_2 - \mathbf{p}_2)^T. \tag{8}$$

Therefore, thorough an approximation, we are able to express $\mathit{\Delta}(\bar{\mathbf{p}})$ as a linear function of the relative abundance vector $\mathbf{p}$. In fact the first derivative of $\mathit{\Delta}_\beta(\mathbf{p})$ turns out to be

$$\mathit{\Delta}'_\beta(\mathbf{p}) = -\frac{\beta+1}{\beta}\mathbf{p}^\beta.$$

In order to express the diversity profile as the parametric function given in (3), we first define

$$\mathbf{u} = -\frac{\beta+1}{\beta}\mathbf{p}^\beta. \tag{9}$$

Hence, the index set is $\mathbf{U} = \{\mathbf{u} : \beta \in \mathbf{B}\}$. By means of (7) and (9) we obtain

$$\mathit{\Delta}(\bar{\mathbf{p}}) = \mathbf{u}\bar{\mathbf{p}}^T + \mathbf{W} + o_p(n^{1/2}) \quad \mathbf{u} \in \mathbf{U}$$

that is the β diversity profile as a linear function of the relative abundance vector with components labelled by the index set $\mathbf{U}$.

Finally, we can write $\bar{\boldsymbol{\theta}}$ in (8) as

$$\bar{\boldsymbol{\theta}} = \mathbf{u}_1(\bar{\mathbf{p}}_1^T - \mathbf{p}_1^T) - \mathbf{u}_2(\bar{\mathbf{p}}_2^T - \mathbf{p}_2^T) \quad \mathbf{u}_1, \mathbf{u}_2 \in \mathbf{U}$$

so that a suitable statistical test to evaluate departure from the null hypothesis is

$$G(\mathbf{u}_1, \mathbf{u}_2) = \frac{\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0}{\left[\frac{\mathbf{u}_1\mathbf{V}_1\mathbf{u}_1^T}{n_1}\right]^{1/2} + \left[\frac{\mathbf{u}_2\mathbf{V}_2\mathbf{u}_2^T}{n_2}\right]^{1/2}} \tag{10}$$

for every value of $\mathbf{u}_1$ and $\mathbf{u}_2 \in \mathbf{U}$.

For practical purposes, we highlight that for $\beta = 0$, $\mathit{\Delta}_\beta(\mathbf{p})$ is defined as

$$\mathit{\Delta}_{\beta=0}(\mathbf{p}) = \frac{1 - \sum_{i=1}^{s} p_i^{\beta+1}}{\beta} = \sum_{i=1}^{s} p_i \ln p_i$$

so that

$$\mathit{\Delta}'_{\beta=0}(\mathbf{p}) = (\mathbf{1} - \ln \mathbf{p}).$$

At this point, by making use of $G(\mathbf{u}_1, \mathbf{u}_2)$ in (10) we obtain a simultaneous critical set for $\boldsymbol{\theta}$ under H_0. Starting from the two samples x and y, we draw M independent bootstrap samples $(x_1^*, x_2^*, ..., x_M^*)$ and $(y_1^*, y_2^*, ..., y_M^*)$, separately. For every $\mathbf{u}$ in the index set $\mathbf{U}$ we evaluate the bootstrapped roots

$$\{G_j^*(x_j^*, y_j^*, \mathbf{u}_1, \mathbf{u}_2) : 1 \le j \le M\}$$

in order to have a bootstrap approximation of the left-continuous cumulative distribution functions (cdf) of the roots $G(x, y, \mathbf{u}_1, \mathbf{u}_2)$.

More explicitly, let $Q_\beta(\boldsymbol{\theta})$, $Q_{sup}(\boldsymbol{\theta})$ and $Q_{inf}(\boldsymbol{\theta})$ be the cdf's of $G(\mathbf{u}_1, \mathbf{u}_2)$, $sup\{Q_\beta(\boldsymbol{\theta})\}$ and $inf\{Q_\beta(\boldsymbol{\theta})\}$ respectively. Let also $Q_\beta^*(\boldsymbol{\theta})$, $Q_{sup}^*(\boldsymbol{\theta})$ and $Q_{inf}^*(\boldsymbol{\theta})$ denote the corresponding bootstrap estimates of these cdf's obtained by taking bootstrap data from the empirical distributions of the two communities separately. We define a $1-\alpha$ simultaneous confidence set for $\boldsymbol{\theta}$ as

$$D_{\boldsymbol{\theta},1-\alpha} = \{\boldsymbol{\theta} : l_{\beta,\alpha} \leq G(\mathbf{u}_1, \mathbf{u}_2) \leq l_{\beta,1-\alpha}\} \tag{11}$$

where the critical values are obtained by

$$l_{\beta,\alpha} = Q_{\beta}^{*-1}[Q_{inf}^{*-1}(\frac{\alpha}{2})]$$

and

$$l_{\beta,1-\alpha} = Q_{\beta}^{*-1}[Q_{sup}^{*-1}(1-\frac{\alpha}{2})]$$

for every $\beta \in \mathbf{B}$. On the basis of (11) we can derive a decision rule as follows:

- accept $H_0 : \Delta_{\beta}(\mathbf{p}_1) = \Delta_{\beta}(\mathbf{p}_2)$ if $\bar{\boldsymbol{\theta}}$ belongs to $D_{\boldsymbol{\theta},1-\alpha}$
- reject H_0 if $\bar{\boldsymbol{\theta}} \geq D_{\boldsymbol{\theta},1-\alpha}$ or, alternatively, if $\bar{\boldsymbol{\theta}} \leq D_{\boldsymbol{\theta},1-\alpha}$
- finally, if both $\bar{\boldsymbol{\theta}} \geq D_{\boldsymbol{\theta},1-\alpha}$ and $\bar{\boldsymbol{\theta}} \leq D_{\boldsymbol{\theta},1-\alpha}$ hold, then reject H_0, in this case the profiles intersect and the communities C_1 and C_2 are not intrinsically comparable.

3. An application

The proposed procedure is experimentally applied to obtain a diversity ordering of avian populations settled in four parks in Milan (Italy): Groane, Lambro, Forlanini and Trenno Parks. A detailed description of the sampling design is given in Fattorini and Marcheselli (1999). We remark that the abundance vector of each population was estimated using independent replications of line transect sampling so that independent estimates $\hat{\mathbf{p}}_{h_i}$ were available for $h_i = 1, ..., n_i$ and $i = 1, 2, 3, 4$. Accordingly, the bootstrap procedures described in section (2) can be applied. Ten evenly spaced β values in the range $-1 \leq \beta \leq 1$ were generated to construct the diversity profile for each park. The restriction that $\beta \geq -1$ ensures that $\Delta_{\beta}(\mathbf{p})$ has certain desirable properties (Patil and Taillie, 1982). On the other hand, for $\beta > 1$ the profiles tend to converge quickly beyond this point and the null hypothesis shall never be rejected.

We emphasize that this approach allows the researcher to perform a more complete analysis on the diversity of each park. In fact, profiles take into account all the aspects of diversity as the number of species, the presence of rare or abundant species, etc. Our procedure allows the researcher to make inference on this global aspect of diversity. Since the aim of the research is to obtain a diversity ordering of the biological communities for each park (Fattorini and Marcheselli, 1999), the use of standard approaches based on comparing simple hypothesis for each single diversity index is not suitable as may lead to different rankings. Thus, the simultaneous inference procedure proposed in section 2 is needed. At this purpose, an estimate of $\boldsymbol{\theta}$, the difference in the diversity profile for each of the possible couples of parks (dotted lines in Fig. 1) is evaluated. Under the null hypothesis of no difference in diversity, we derive a 0.95 simultaneous confidence sets for $\boldsymbol{\theta}$ (solid lines in

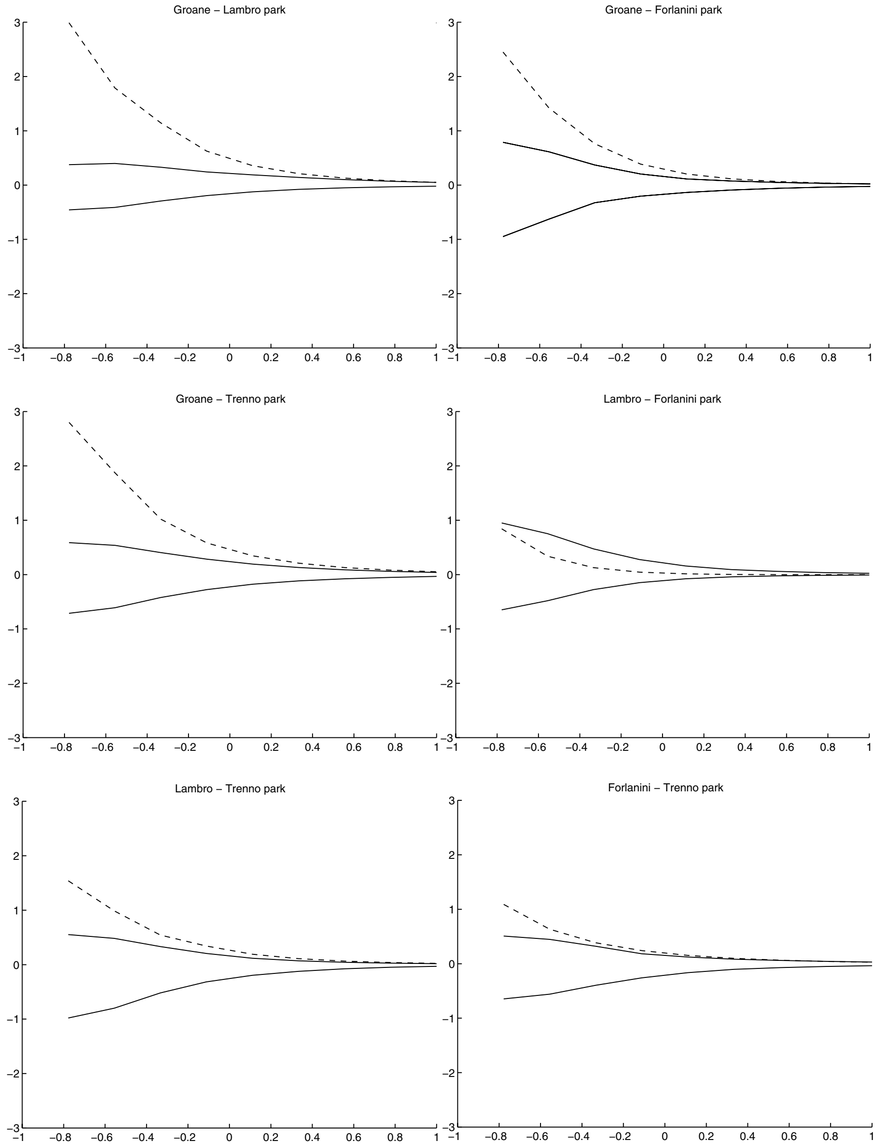

Fig. 1. β diversity profile difference estimates (dotted line) and 0.95 simultaneous confidence sets (solid line)

Fig. 1). In Fig. 1, we show the results of the test of hypothesis for each couple of β profiles. If the dotted line is bounded by the solid line, we accept the null hypothesis of equivalence in diversity between the two communities. Conversely, if the dotted line goes beyond the upper or the lower limit of the solid line we reject H_0 and accept dominance. Finally, it might happen that the dotted line is beyond both the upper and the lower limit of the solid line. In this former case we reject H_0 and accept the crossing alternative, that is, the two communities are not comparable in terms of diversity. Results show a good performance of the method suggested in order to describe the diversity ordering of the communities. In particular, we can conclude that Groane Park and Trenno Park turn out to be the most intrinsically diverse and the least intrinsically diverse, respectively. Moreover, Lambro and Forlanini Parks are at an intermediate diversity level being these parks equivalent in terms of β diversity profiles.

4. Discussion

In this paper we proposed a method for constructing a critical area for testing hypotheses of equivalence between two biological communities by means of a suitable non parametric procedure. A balanced simultaneous critical set for the difference of β diversity profiles has been obtained. In order to evaluate the performance of the method, a data set previously analyzed in Fattorini and Marcheselli (1999) regarding the comparison of the biological diversity of four parks in Milan, Italy, has been used. In particular, we remark that the diversity ordering obtained turns out to be the same as that obtained using the concept of intrinsic diversity profiles. This confirms the findings in Gove et al. (1994) that the $\Delta_\beta(\mathbf{p})$ profiles are isotonic to intrinsic diversity ordering so that, if an intrinsic diversity ordering exists they will preserve it. These results are obtained without any model assumption neither of the species abundance nor for the β profile estimator. In this work we concentrated our attention on the β-family of diversity index in (1). Further work on the Good's family index is certainly worthwhile.

References

Baczkowski AJ, Joanes DN, Shamia GM (1998). Range of validity of α and β for a generalized diversity index $H(\alpha, \beta)$ due to Good. Mathematical Biosciences 148: 115–128

Baczkowski AJ, Shamia, G (2000) The distribution of a generalized diversity index due to Good. Environmental and Ecological Statistics 7: 329–342

Basharin, GP (1959) On a statistical estimate for the entropy of a sequence of independent random variables. Theory of Probability and its Applications 4: 333–336

Beran R (1988) Balanced Simultaneous Confidence Sets. Journal of the American Statistical Association 83: 679–686

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

Good, IJ (1953) The population frequencies of species and the estimation of population parameters. Biometrika 49: 237–264

Gove HJ, Patil JP, Swindel, BF (1994) Ecological diversity and forest management. Handbook of Statistics 12: 409–462

Hurlbert SH (1971) The non concept of species diversity: a critique and alternative parameters. Ecology 52: 577–586

Patil GP, Taillie C (1982) Diversity as a Concept and its Measurement. Journal of the American Statistical Association 77: 548–567

Tong, YL (1983) Some Distribution Properties of the Sample Species-Diversity Indices and their application. Biometrics 39: 999–1008