NON PARAMETRIC TESTS AND CONFIDENCE REGIONS TO EVALUATE INTRINSIC DIVERSITY PROFILES OF ECOLOGICAL POPULATIONS

TONIO DI BATTISTA - S. ANTONIO GATTONE

Department of Quantitative Methods and Economic Theory
Universita "G. D'Annunzio" of Chieti
Viale Pindaro, 42
65127 Pescara Italy

dibattis@dmqte.unich.it sgattone@yahoo.com

THE AIM

TO EVALUATE THE DIVERSITY OF BIOLOGICAL COMMUNITIES

↓

HOW DO SPECIES EVOLVE AND INTERACT EACH OTHER?

AVAILABLE APPROACH

- DIVERSITY INDEX: Shannon, Simpson, Brillouin, etc.

Species richness | the degree of evenness

Can be expressed as a function $g(\mathbf{N})$ → $\mathbf{N} = (N_1, N_2, \ldots, N_s)$ → number of units belonging to the *i*-th species

SOME
D
R
A
W
B
A
C
K
S

- IT IS NOT POSSIBLE TO RECOMMEND A SINGLE INDEX AS SUPERIOR TO ALL OTHERS
- DIFFERENT INDEXES MAY INCONSISTENTLY RANK A GIVEN PAIR OF COMMUNITIES
- DESCRIPTION OF WHOLE COMMUNITIES BY ONE STATISTIC RUN THE RISK OF LOOSING MUCH VALUABLE INFORMATION

Using an index
ALTERNATIVE

A DIVERSITY PROFILE APPROACH (Patil and Taillie, 1982)

$$T_j = \frac{1}{N} \sum_{i=j+1}^{s} N_{(i)}, \qquad j = 1, \ldots, s-1$$

where $N_{(1)}, N_{(2)}, \ldots, N_{(s)}$ are the ranked abundance,
$N = \mathbf{1}^T \mathbf{N}$, $\mathbf{T}_s = 0$ and $\mathbf{T}_0 = 1$

The convex curves obtained by plotting the $\left(j, T_j\right)$ pairs

Are termed intrinsic diversity profiles

Let $\mathbf{T}_{C_1}$ and $\mathbf{T}_{C_2}$ be the intrinsic diversity profiles of communities C_1 and C_2

$\mathbf{T}_{C_1} \leq \mathbf{T}_{C_2}$

C_2 is more intrinsically diverse than C_1 (dominance)

$\mathbf{T}_{C_1}$ crosses $\mathbf{T}_{C_2}$

communities are not intrinsically comparable

$\mathbf{T}_{C_1} = \mathbf{T}_{C_2}$

no differences in diversity between C_1 and C_2 (equivalence)

However $\mathbf{T}_{c_1}$ *and* $\mathbf{T}_{c_2}$ are descriptive statistics

THE GOAL OF THE WORK

AN INFERENTIAL APPROACH TO COMPARE TWO OR MORE COMMUNITIES ACCORDING TO THEIR DIVERSITY IS PROPOSED

CONFIDENCE SETS

TESTING HYPOTHESES

LITERATURE

♦ Fattorini and Marcheselli (1999)

HOWEVER, IN THIS SETTING, PROBLEMS ARISE BECAUSE OF

difficulty in assuming any specific distribution both for the abundance vector and for the diversity profile

components of diversity profiles are dependent and heteroschedastic

PROPOSAL

A NON PARAMETRIC METHODOLOGY BASED ON BOOTSTRAP IS PRESENTED

F R A M E W O R K

LET $\mathbf{n} = \left(\hat{\boldsymbol{\pi}}_1, \hat{\boldsymbol{\pi}}_2, \ldots, \hat{\boldsymbol{\pi}}_n\right)$ BE n i.i.d. UNBIASED ESTIMATORS OF THE RELATIVE ABUNDANCE VECTOR $\boldsymbol{\pi}$

→ Replicated sampling - Barabesi and Fattorini (1998)

then, an improved unbiased estimator for $\boldsymbol{\pi}$ is given by

$$\overline{\boldsymbol{\pi}}_n = \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\pi}}_i$$ with variance-covariance $n^{-1}\mathbf{V}$

and an unbiased and consisten estimator for $\mathbf{V}$ is

$$\hat{\mathbf{V}}_n = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\boldsymbol{\pi}}_i - \overline{\boldsymbol{\pi}}_n\right)\left(\hat{\boldsymbol{\pi}}_i - \overline{\boldsymbol{\pi}}_n\right)^T$$

IN THIS SETTING, THE NOW-FAMILIAR NON PARAMETRIC BOOTSTRAP METHOD (EFRON, 1982) MAY BE SUITABLY ADOPTED ON THE *i.i.d.* SAMPLE $\mathbf{n} = \left(\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_n\right)$

PROBLEM

THE INTRINSIC DIVERSITY PROFILE $T(\boldsymbol{\pi})$ IS A FAMILY OF $S-1$ PARAMETRIC FUNCTIONS

ASSERTING CONFIDENCE SET FOR $T(\boldsymbol{\pi})$ INVOLVES DIFFICULT QUESTIONS OF SIMULTANEOUS INFERENCE

A SOLUTION

OF THIS PROBLEM IS GIVEN BY BERAN'S PROCEDURE (JASA, 1988)

BALANCED AND SIMULTANEOUS CONFIDENCE SETS FOR PARAMETERS EXPRESSED AS A LINEAR COMBINATIONS OF MEANS

IN OUR FRAMEWORK, THE PROCEDURE OF BERAN (1988) MAY BE IMPLEMENTED IN ORDER TO OBTAIN SIMULTANEOUS CONFIDENCE SETS FOR THE INTRINSIC DIVERSITY PROFILE $\mathbf{T}(\boldsymbol{\pi})$

IN FACT, WE CAN EXPRESS THE INTRINSIC DIVERSITY PROFILE AS A **LINEAR FUNCTION** OF $\boldsymbol{\pi}$

$$\mathbf{T}(\boldsymbol{\pi}) = \{T_{\mathbf{b}}(\boldsymbol{\pi}) = \mathbf{b}'\boldsymbol{\pi} \ : \mathbf{b} \in B\}$$

WHERE

$$\mathbf{b}_l = \sum_{j=1}^{s+1-l} \mathbf{a}_j \qquad l = 1,\ldots,s-1 \qquad \text{and } B = \{\mathbf{b}_l;\ l = 1,\ldots,s-1\}$$

AND $a_1, a_2, \ldots, a_{s-1}$ ARE THE SET OF STANDARD BASIS OF $\Re^{S-1}$

The statistic proposed for $T_{\mathbf{b}}(\boldsymbol{\pi})$, called by Beran "studentized root", is:

$$R_{n,\mathbf{b}}(\mathbf{n}, T_{\mathbf{b}}(\boldsymbol{\pi})) = n^{1/2} \frac{\mathbf{b}(\hat{\boldsymbol{\pi}}_n - \boldsymbol{\pi})}{[\mathbf{b}'\mathbf{V}_n\mathbf{b}]^{1/2}}$$

- A RESULTING CONFIDENCE SET FOR $T(\boldsymbol{\pi})$

$$C_n = \{t \in \mathbf{T} : L_{n,\mathbf{b}} \leq R_{n,\mathbf{b}}(t_{\mathbf{b}}) \leq U_{n,\mathbf{b}},\ \forall\, \mathbf{b} \in \mathbf{B}\}$$

- THE PROPOSED BOOTSTRAP VERSION OF SIMULTANEOUS CONFIDENCE SET IS OBTAINED BY TAKING THE CRITICAL VALUES

$$L_{n,\mathbf{b}} = H_{n,\mathbf{b}}^{*-1}\left[H_{\inf}^{*-1}\left(\frac{\alpha}{2}\right)\right] \qquad U_{n,\mathbf{b}} = H_{n,\mathbf{b}}^{*-1}\left[H_{\sup}^{*-1}\left(1-\frac{\alpha}{2}\right)\right]$$

corresponding bootstrap estimates

by taking bootstrap samples $\mathbf{n}^* = (\hat{\pi}_1^*, \hat{\pi}_2^*, \ldots, \hat{\pi}_n^*)$ from the $\hat{\boldsymbol{\pi}}_i$'s where, $H^*_{n,\mathbf{b}}$, $H^*_{\inf}$, $H^*_{\sup}$ are the corresponding bootstrap estimates of:

♦ $H_{n,\mathbf{b}}(\boldsymbol{\pi})$ is the cdf of the root $R_{n,\mathbf{b}}$

♦ → $H_{\sup}(\boldsymbol{\pi})$ → $\sup\{H_{\mathbf{n},\mathbf{b}}(R_{\mathbf{n},\mathbf{b}}, \boldsymbol{\pi}) : \in \mathbf{B}\}$

cdf of a suitable transformed root

→ $H_{\inf}(\boldsymbol{\pi})$ → $\inf\{H_{\mathbf{n},\mathbf{b}}(R_{\mathbf{n},\mathbf{b}}, \boldsymbol{\pi}) : \in \mathbf{B}\}$

respectively.

ADVANTAGES OF THE METHOD

♦ C_n is set to have overall coverage probability $1-\alpha$

♦ C_n is balanced, that is

↓

the coverage probability for the confidence interval of each right tail-sum $T_{\mathbf{b}}(\boldsymbol{\pi})$ remains unchanged as $\mathbf{b}$ varies

♦the method is non parametric where other methods such Tukey (1953), Scheffè (1953) and Richmond (1982) require components of $\boldsymbol{\pi}$ independent and normally distributed

HYPOTHESIS TESTING

IN ORDER TO EVALUATE AND MONITOR BIODIVERSITY
AT DIFFERENT SPATIAL (OR TEMPORAL) SCALES

Communities C_1, C_2 with diversity profiles $\mathbf{T}_1$, $\mathbf{T}_2$

$H_0 : \mathbf{T}_1 = \mathbf{T}_2$ against $H_1 : \mathbf{T}_1 \neq \mathbf{T}_2$

- Gove et al (1994) jackknifing approach, but
  ↓
  unresolved questions of simultaneous inference

- Fattorini and Marcheselli (1999) proposed an asymptotically conservative procedure
  ↓
  deriving asymptotic multinormality distribution for $\hat{\mathbf{T}}_1$ and $\hat{\mathbf{T}}_2$

$\boldsymbol{\theta} = \mathbf{T}_1 - \mathbf{T}_2$ is the $s-1$ vector of differences in the right tail-sums of the two communities

→ the equivalence hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 = \mathbf{0}$

In this framework and applying Beran's procedure, a suitable statistical test is

$$G_{n,\mathbf{b}}(\boldsymbol{\theta}_0) = \frac{\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}_0}{\left[\frac{\mathbf{b}'\overline{\mathbf{V}}_1\mathbf{b}}{n_1}\right]^{\frac{1}{2}}\left[\frac{\mathbf{b}'\overline{\mathbf{V}}_2\mathbf{b}}{n_2}\right]^{\frac{1}{2}}}, \qquad \forall \mathbf{b} \in \mathbf{B}$$

where,

$\overline{\boldsymbol{\theta}} = \{\mathbf{b}'(\overline{\boldsymbol{\pi}}_1 - \overline{\boldsymbol{\pi}}_2) : \mathbf{b} \in \mathbf{B}\}$ → $\boldsymbol{\theta}$

$\overline{\boldsymbol{\pi}}_1, \ \overline{\boldsymbol{\pi}}_2$ → $\boldsymbol{\pi}_1, \ \boldsymbol{\pi}_2$

$\overline{\mathbf{V}}_1, \overline{\mathbf{V}}_2$ → $\mathbf{V}_1, \mathbf{V}_2$

are straightforward estimators of the population counterparts.

THE KEY IDEA:

to construct a simultaneous confidence set for parameter $\boldsymbol{\theta}$ in order to define a critical area of the test under $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 = \mathbf{0}$

$$\left.\begin{array}{l} Q_{\mathbf{b}}(\boldsymbol{\pi}) \\ Q_{\sup}(\boldsymbol{\pi}) \\ Q_{\inf}(\boldsymbol{\pi}) \end{array}\right\} \quad \text{cdf of} \quad \left\{\begin{array}{l} G_{\mathbf{b}} \\ \sup\{Q_{\mathbf{b}}(\boldsymbol{\theta}) : \mathbf{b} \in \mathbf{B}\} \\ \inf\{Q_{\mathbf{b}}(\boldsymbol{\theta}) : \mathbf{b} \in \mathbf{B}\} \end{array}\right.$$

$$\left\{\begin{array}{l} Q^*_{\mathbf{b}}(\boldsymbol{\pi}) \\ Q^*_{\sup}(\boldsymbol{\pi}) \\ Q^*_{\inf}(\boldsymbol{\pi}) \end{array}\right.$$

cdf bootstrap estimates by taking bootstrap data from the empirical distributions of the two communities separately finally, the $1-\alpha$ simultaneous confidence set for $\boldsymbol{\theta}$ is

$$D_{\theta,1-\alpha} = \left\{ \boldsymbol{\theta} : l_{\mathbf{b}} \leq G_{\mathbf{b}} \leq u_{\mathbf{b}}, b \in \mathbf{B} \right\}$$

where

$$l_{\mathbf{b}} = Q_{\mathbf{b}}^{*-1}\left[Q_{\inf}^{*-1}\left(\frac{\alpha}{2}\right) \right] \qquad u_{\mathbf{b}} = Q_{\mathbf{b}}^{*-1}\left[Q_{\sup}^{*-1}\left(1-\frac{\alpha}{2}\right) \right]$$

are the critical values obtained from the empirical bootstrap distribution

DECISION RULE

- if both $\overline{\boldsymbol{\theta}} \geq D_{\boldsymbol{\theta},1-\alpha}$ and $\overline{\boldsymbol{\theta}} \leq D_{\boldsymbol{\theta},1-\alpha}$ hold
  ⇓
  reject H_0
  communities not intrinsically comparable

- if $\overline{\boldsymbol{\theta}} \subseteq D_{\boldsymbol{\theta},1-\alpha}$
  ⇓
  accept H_0
  equivalence

- if $\overline{\boldsymbol{\theta}} \geq D_{\boldsymbol{\theta},1-\alpha}$ or $\overline{\boldsymbol{\theta}} \leq D_{\boldsymbol{\theta},1-\alpha}$
  ⇓
  reject H_0
  dominance

AN APPLICATION

AVIAN COMMUNITIES OF 4 PARKS IN MILAN (ITALY)

↓

for a detailed description of the sample design
Fattorini and Marcheselli (1999)

♦ for each park we compute the diversity profile as $\mathbf{T}(\bar{\pi}_n)$ and derive simultaneous confidence sets

♦ in order to rank the park according to their diversity we evaluate an estimate of $\boldsymbol{\theta}$ for each of the possible couples of parks

♦ under the null hypothesis of no difference in diversity we derive a 0.95 simultaneous confidence set for $\boldsymbol{\theta}$

RESULTS

♦Groane and Trenno parks turn out to be the most intrinsically and the least intrinsically diverse, respectively

♦Lambro and Forlanini parks can be located at an intermediate diversity level being these parks equivalent in terms of intrinsic diversity profiles

♦Confidence regions for Groane park clearly show departures from symmetry

Therefore, we have the following ranking of the four parks:

1. Groane (the most diverse)
2. Lambro and Forlanini (intermediate level of diversity)
3. Trenno (the least diverse)