Qual Quant (2017) 51:847–858
DOI 10.1007/s11135-016-0443-x

Bootstrap confidence intervals for biodiversity measures based on Gini index and entropy

Nicola Pesenti[1] · Piero Quatto[1] · Enrico Ripamonti[1]

Published online: 17 October 2016

Abstract Monitoring the richness and the diversity of species living in an ecosystem is an important goal of ecology. To this purpose, measures of biodiversity have been introduced as statistical summaries of the abundance vector. In particular, we take into consideration the Gini–Simpson and the Shannon–Wiener indices, along with the effective number of species calculated through these measures, proposed, respectively, by Laakso and Taagepera (Comp Polit Stud 12:3–25, 1979) and Leti (Statistica descrittiva, Bologna, Il Mulino, 1983). It is an open question how to associate to these indices a measure of uncertainty. In this paper we compare confidence intervals based on these measures, calculated through three different bootstrap methods: percentile, -t and accelerated bias-corrected percentile. We recommend to practitioners to use the percentile procedure, as it is straightforward and computationally feasible, providing results very close to those obtained by more complex techniques.

Keywords Gini–Simpson index · Shannon–Wiener index · Leti index · Laakso–Taagepera index · Bootstrap methods · Confidence intervals

1 Introduction

Biological diversity, also known as biodiversity, is the variation among living organisms within a given habitat or ecosystem. Biodiversity boosts ecosystem productivity and all the species play an important role to this purpose. Thus, it is important to monitor the health and diversity of species over time, such as changes in size and distribution of population of species, habitats, and interactions among communities.

✉ Nicola Pesenti
nicolapesenti@hotmail.it

[1] Department of Economics, Management and Statistics, University of Milan-Bicocca, Milan, Italy

Measures of biodiversity are statistical summaries of the abundance vector, that is the vector of proportions of each species in the community, which helps to understand the condition of biodiversity and the variables affecting it. Two main factors are taken into account when measuring biodiversity: richness and evenness. Richness is a proxy of the number of different kinds of organisms living in the study area. The number of species, s, is a measure of richness and it is the easiest way to estimate biodiversity. Species richness does not take into account the number of individuals for each species. However, biodiversity does not only depend on richness, but also on evenness. Evenness compares the similarity of the population size of each of the species belonging to the same environment; it is a measure of the relative abundance of the different species constituting the richness of an area. Therefore a well-established diversity index must take into consideration both of these factors.

In the current practice, biodiversity measures are generally provided by Environmental Agencies as descriptive indices. However, along with the magnitude, it is also important to evaluate the degree of uncertainty underpinning a certain measure. In the paper we consider this problem. Firstly, we estimate biodiversity using the Gini index, the entropy and their effective number of species. Secondly, we propose to calculate confidence intervals for these measures using percentile, -t and accelerated bias-corrected percentile methods.

In Sect. 2 we describe the data that we will use throughout the paper as a motivating example. In Sect. 3 we provide a brief overview on biodiversity measures, focusing in particular on the Gini–Simpson and the Shannon–Wiener diversity indices. In Sect. 4 we put forward a brief snapshot on the methods to calculate bootstrap confidence intervals that we will employ in the application. In Sect. 5 we present the results of our work and in Sect. 6 we provide a simulation study to support our conclusions, which are presented in Sect. 7.

2 Motivating example

We present biodiversity measures for Northern Ireland seabirds, which are a useful and important indicator to assess the state of the marine environment; in fact, they essentially represent the top of the food chain. They react to a range of factors such as modification in food availability, climate change, predation and pollution. For this reason, they have been studied by the Joint Nature Conservation Committee (JNCC) Seabird Monitoring Program, launched in 1986 in the UK and Ireland. This Program has provided high-quality datasets of population counts and demographic parameters.

The species composition may change over the years depending on the quality and quantity of data for each species. The database includes counts from the 1980s to the present and updates are provided annually. Data are available online at (http://jncc.defra.gov.uk/page-4460).

3 Biodiversity measures

3.1 Properties of a biodiversity index

In order to study the biodiversity of a system, let's suppose to have collected a sample of n individuals of s different species. Let $(n_1, n_2, \ldots, n_s)$ be the frequencies abundance vector, where n_i is the number of individuals belonging to the species $i\,(i = 1, \ldots, s)$, with

$$\sum_{i=1}^{s} n_i = n$$

and $p = (p_1, p_2, \ldots, p_s)$ be the proportions abundance vector, with p_i representing the probability that an individual selected randomly from a population belongs to the species i, defined as

$$p_i = n_i/n$$

so that

$$\sum_{i=1}^{s} p_i = 1.$$

A biodiversity index I is a statistical measure that summarizes the relative abundance vector p, $I(p) = I(p_1, p_2, \ldots, p_s)$, with the following essential properties (see for instance Frosini 2006):

1. $I(p)$ is non-negative and zero-indifferent function. We recall that a generic function $I_s(p)$ of the vector p of size s, is zero-indifferent if
$$I_s(p_1, \ldots, p_s) = I_{s+t}(p_1, \ldots, p_s, 0, \ldots, 0)$$
for $t > 0$.
2. The index $I(p)$ is minimum when all the individuals belong to a single species i, that is to say
$$\exists i \leq s \quad p_i = 1 \quad \text{and} \quad p_j = 0 \quad \forall j \neq i.$$
3. The index $I(p)$ is maximum when the species are equally common:
$$p_i = \frac{1}{s} \quad \forall i$$
4. The maximum value
$$f(s) = \max[I_s(p)]$$
represents a strictly increasing function $f(s)$ of the number of species s: the higher the value of s, the higher the maximum value, emphasizing the key role played by s in any measure of biodiversity.

3.2 Biodiversity indices

All the indices satisfying the above properties are called biodiversity indices. In the statistical literature different diversity indices have been proposed, satisfying different assumptions. The simplest diversity index is the species richness, namely s, which expresses the number of species present in an environment:

$$s = s(p) = \#(p_i > 0).$$

This is a straightforward and intuitive index, but it lacks of information with respect to the evenness of the distribution. The most widely used measures of biodiversity that take into account both abundance and evenness of species present in the community are the Gini–Simpson and the Shannon–Wiener indices.

3.3 Gini–Simpson diversity index

The Gini–Simpson index was introduced by Gini (1912) as a concentration index and it was proposed as a measure of biodiversity by Simpson (1949). Let's first introduce the probability that two individuals randomly selected from a sample belong to the same species:

$$D(p) = \sum_{i=1}^{s} p_i^2$$

The value of $D(p)$ ranges between 0 and 1: 0 represents maximum diversity and 1 indicates absence of diversity. So, the higher the value of $D(p)$, the lower the diversity. The Gini–Simpson diversity index is obtained as:

$$E(p) = 1 - D(p) = 1 - \sum_{i=1}^{s} p_i^2 \tag{1}$$

and represents the probability that two individuals, randomly selected from a sample, belong to different species. Now, the higher the value of this index, the higher the diversity. The index varies between 0 and 1 and reaches its maximum value when all individuals belong to a single species, i.e.:

$$\exists i \leq s \quad p_i = 1.$$

Moreover, under the constraints:

$$\sum_{i=1}^{s} p_i = 1; \quad 0 \leq p_i \leq 1 \tag{2}$$

its maximum value is:

$$\max[E(p)] = \frac{s-1}{s}$$

corresponding to the case in which all species are equally distributed. In the ecological context, the Gini–Simpson diversity index is viewed as a dominance index because it attributes more weight to common or dominant species. In particular, the weight that the Gini index gives to each species is $1 - p_i$. Hence for a rare species, the value of p_i is very low (close to zero), and the associated weight will be closed to one. So, the presence of rare species causes only small changes in the value of the index.

3.4 Shannon–Wiener diversity index

The Shannon–Wiener diversity index (Shannon 1948; Shannon and Weaver 1949; Wiener 1949), also known as entropy, is a popular diversity index in the ecological literature and it is given by:

$$H(p) = -\sum_{i=1}^{s} p_i \log(p_i) \tag{3}$$

where log denotes the natural logarithm.

The entropy quantifies the uncertainty in predicting the species identity of an individual selected at random from the dataset. This uncertainty increases as the number of species increases and as the individuals are distributed more evenly among the species. A high value of $H(p)$ would be representative of a diverse and equally distributed community and lower values show a less diversity community. A value of 0 would represent a community with just one species (where conventionally $0\log 0 = 0$). Moreover, under the constraint (2), the maximum value of $H(p)$ occurs when each species in the community has the same frequency $p_i = 1/s$:

$$\max(H(p)) = \log(s).$$

In contrast with the Gini–Simpson index, the Shannon–Wiener diversity index is particularly sensitive to the number of rare species in a community, i.e., those species characterized by extremely low relative frequency. In this case, the weight that the entropy gives to each species is $\log(p_i)$.

3.5 Normalization of diversity indices

It is worth observing that normalization should not be used in the context of biodiversity measures because it removes the effect of the number of species, which is instead of interest in this framework. More precisely, a normalized diversity index does not satisfy property 4 of Sect. 3.1.

3.6 Effective number of species

Diversity indices have a wide variety of range and behaviors. If we apply them to an equally distributed community of s species, each index would return a different value (e.g., s, $1 - 1/s$, $\log(s)$). One might expect that the diversity of an equally distributed population of 10 species is twice the diversity of an equally distributed population of 5 species. However, this does not happen using diversity indices. That is the reason of the introduction of the effective number of species.

Given a system with s species with biodiversity measured by the index I_s, the effective number of species, r, is obtained by solving the equation:

$$\max(I_r) = I_s$$

with solution

$$\mathrm{r} = f^{-1}(I_s)$$

satisfying properties 1–4 of Sect. 3.1, because of the strict monotonicity of the function f.

To derive the effective number of species for the Gini–Simpson index, we need to solve with respect to r the equation:

$$\max(E_r(p)) = 1 - \frac{1}{r} = E_s(p)$$

The solution is the Laakso–Taagepera index (Laakso and Taagepera 1979):

$$\mathrm{F(p)} = \frac{1}{1 - E_s(p)} = \frac{1}{\sum_{i=1}^{s} p_i^2}$$

Similarly, solving the equation:

$$\log(\mathrm{r}) = \max[H_r(\mathrm{p})] = H_s(p)$$

The effective number of species for the Shannon–Wiener index is given by:

$$\mathrm{L}(p) = \exp[H_s(p)] = \prod_{i=1}^{s} p_i^{-p_i}$$

and it is called Leti diversity index (Leti 1983).

4 A snapshot on bootstrap confidence intervals

In this article we use three different methods to calculate bootstrap confidence intervals (CI): the percentile, the -t and the accelerated bias-corrected percentile. For a comprehensive description of these methods, we refer the reader to Carpenter and Bithell (2000) and to Tu and Shao (1995). In particular, we have drawn $B = 1000$ samples with replacement of size n from our vector of data $(x_1, \ldots, x_n)$ with $x_i = k$ if the individual i belongs to the species k (for $i = 1, \ldots, n$ and $k = 1, \ldots, s$), that is to say, we have considered B bootstrap samples from a Multinomial distribution with parameters n and $p = (p_1, p_2, \ldots, p_s)$.

4.1 Bootstrap percentile CI

The bootstrap percentile CI represents the simplest way to calculate accurate bootstrap CI. Let $\hat{\vartheta}_n$ be an estimator for the parameter ϑ and $\hat{\vartheta}_n^*$ the bootstrap estimator based on a bootstrap sample $(x_1^*, \ldots, x_n^*)$, so that $\hat{\vartheta}_n^* = \hat{\vartheta}(x_1^*, \ldots, x_n^*)$. If we define the distribution function of $\hat{\vartheta}_n^*$ as

$$\mathrm{G}^*(\mathrm{t}) = \mathrm{P}(\hat{\vartheta}_n^* \leq \mathrm{t})$$

then the bootstrap percentile method returns the CI

$$[\vartheta_\alpha^*, \vartheta_{1-\alpha}^*]$$

where ϑ_α^* is the $\alpha-$ quantile of the bootstrap version of $\hat{\vartheta}_n^*$:

$$\vartheta_\alpha^* = G^{*-1}(\alpha) = inf\{t \in \mathbb{R} : G^*(t) \geq \alpha\}.$$

Hence, we can approximate the distribution function G^* with the empirical distribution function G_B of the B bootstrap replications.

4.2 Bootstrap-t CI

With this method we need to calculate a bootstrap version of the estimator $\hat{\vartheta}_n$, i.e., $\hat{\vartheta}_n^* = \hat{\vartheta}(x_1^*, \ldots, x_n^*)$, as well as a bootstrap version of the standard deviation of the estimator $\hat{\vartheta}_n$

$$\hat{\sigma}_\alpha^* = \hat{\sigma}(x_1^*, \ldots, x_n^*).$$

So, if the "studentized" statistic

$$T_n^* = \frac{\hat{\vartheta}_n^* - \hat{\vartheta}_n}{\hat{\sigma}_n^*}$$

is pivotal, then the bootstrap-t CI for ϑ with nominal level $(1 - 2\alpha)$ is given by

$$[\hat{\vartheta}_n - t_{1-\alpha}^* \hat{\sigma}_n, \hat{\vartheta}_n - t_\alpha^* \hat{\sigma}_n]$$

where t_α^* is the $\alpha-$ quantile of the bootstrap version of the "studentized" statistic.

Differently from the percentile method, the bootstrap-t involves a deviance estimate, $\hat{\sigma}_n$. However, it is possible to obtain such an estimate, using a second-level bootstrap: for each bootstrap sample $(x_1^*, \ldots, x_n^*)$, we generate $M = 100$ additional bootstrap samples to calculate the usual sample variance. Consequently, such a double bootstrap method is computationally more intensive than the percentile bootstrap.

4.3 Bootstrap accelerated bias-corrected percentile CI

The bias-corrected accelerated (BCa) bootstrap is a generalization of the percentile method introducing the two constants z_0 (bias correction) and a (acceleration) in order to adjust for bias and skewness of the bootstrap distribution. Consider an increasing function h, and the transformation $\hat{\eta}_n = h(\hat{\vartheta}_n)$ with standard deviation $\sigma = 1 + a\eta$, such that

$$P\left(\frac{\hat{\eta}_n - \eta}{\sigma} + z_0 \leq z\right) = H(z)$$

where H is a continuous, strictly increasing and symmetrical transformation (e.g. the standard normal distribution function). Thus, a CI for the parameter ϑ with nominal level $(1 - 2\alpha)$ is given by

$$[\vartheta_\alpha, \vartheta_{1-\alpha}]$$

where

$$\vartheta_\alpha = h^{-1}\left(\hat{\eta}_n + \frac{z_0 + z_\alpha}{1 - a(z_0 + z_\alpha)} \hat{\sigma}_n\right)$$

$$z_\alpha = H^{-1}(\alpha)$$

$$\hat{\sigma}_n = 1 + a\hat{\eta}_n$$

Under these assumptions it is possible to derive the equalities:

$$G^*(\vartheta_\alpha) = \left(z_0 + \frac{z_0 + z_\alpha}{1 - a(z_0 + z_\alpha)}\right)$$

$$z_0 = H^{-1}(G^*(\hat{\vartheta}_n))$$

where

$$\mathrm{G}^*(t) = P(\hat{\vartheta}_n^* \le t)$$

is the distribution function of the bootstrap version of $\hat{\vartheta}_n$. So, the quantiles:

$$\begin{aligned}\vartheta_\alpha^* &= G^{*-1}\left(H\left(z_0 + \frac{z_0 + z_\alpha}{1 - a(z_0 + z_\alpha)}\right)\right)\\ &= inf\left\{t \in \mathbb{R} : G^*(t) \ge H\left(z_0 + \frac{z_0 + z_\alpha}{1 - a(z_0 + z_\alpha)}\right)\right\}\end{aligned}$$

lead to the $(1 - 2\alpha)$ BCa CI for the parameter ϑ:

$$\left[\vartheta_\alpha^*, \vartheta_{1-\alpha}^*\right].$$

According to Efron (1987), we have estimated the bias correction and the acceleration by using the suitable function implemented in R.

5 Results

5.1 Descriptive analysis

Data from the JNCC are available for five countries (England, Scotland, Wales, Ireland, Northern Ireland). To the purposes of this application, we will focus only on Northern Ireland data. The practice of considering separately ecological data from different GB countries is particularly recommendable, being the territory very large and characterized by geographic areas being climatically and morphologically very different.

We remark that JNCC data are described through different labels, which indicate, in particular, the adjustment and the accuracy of the counting method. For these details, we refer the reader to the explanation available online. In this paper, we consider seven species of seabird (excluding three species with very low and not reliable counts) and we only take into account entries labeled as "accurate".

We show results for the period 2006–2009, calculating the Gini–Simpson index ($E(p)$), the Shannon–Wiener index ($H(p)$) and the corresponding effective number of species, given by the Laakso–Taagepera index ($F(p)$) and the Leti diversity index ($L(p)$). Table 1 shows the total observations per year, and Table 2 shows the diversity indexes per year.

From Table 2 it does emerge a very similar trend between the two indices $E(p)$ and $H(p)$, as well as between the effective number of species $F(p)$ and $L(p)$, with a slightly decline over 2006–2008 and a recovery in 2009. We recall that the indices $F(p)$ and $L(p)$ represent the number of equally common species which would produce the value of the index that it has been really observed. For the Gini–Simpson index this value is between three and six species, while for entropy is between six and nine species. This happens as the entropy attributes larger weight to rare species than the Gini–Simpson index.

Table 1 Total counts of seabirds in Northern Ireland, period 2006–2009

Year	Total count
2006	16.928 units
2007	14.970 units
2008	15.124 units
2009	10.696 units

Table 2 Diversity indices values, period 2006–2009

Index	2006	2007	2008	2009
E(p)	0.790	0.764	0.709	0.836
H(p)	1.931	1.909	1.761	2.105
F(p)	4.754	4.230	3.436	6.087
L(p)	6.898	6.746	5.819	8.206

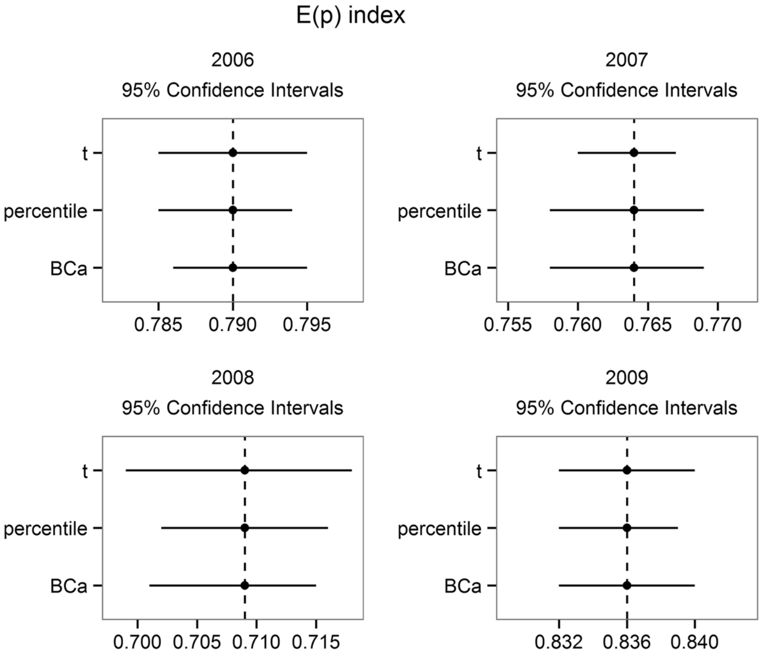

Fig. 1 Confidence intervals in the period 2006–2009 for the three bootstrap methods -t, percentile and BCa associated with the E(p) index

5.2 Comparison of confidence intervals

In Fig. 1 we show CI for the E(p) index across years, as obtained by means of the three bootstrap methods. The same analysis has been done with index H(p) (Fig. 2), index F(p) (Fig. 3) and index L(p) (Fig. 4).

Three main results pop out from our analysis. First, the width of the bootstrap CI obtained for the four measures of diversity considered is very small, and the confidence bounds calculated adopting different bootstrap methods is very similar. Second, we can notice that results obtained with the percentile method and with BCa are very close. This might indicate that, in this context, the BCa procedure does not provide a compelling advantage. Third, occasional differences do emerge by comparing the bootstrap-t method with the other two methods. This is reasonable, as the -*t* method is indeed conceptually different as compared to the percentile procedures. We may observe that, in terms of length of the CI, the bootstrap-*t* provides results sometimes better (e.g., for the index $E(p)$—year 2007, or for the index $H(p)$ for year 2009), and sometimes worse (e.g., $E(p)$ index, for the year 2008, or $H(p)$ for the year 2006) than those obtained with the other methods.

6 A simulation study

In order to support our conclusions, a simulation study was performed. A new population $(n_1, n_2, \ldots, n_s)$, of $s = 25$ groups, was generated from a uniform distribution in the interval [0, 1000], where n_i is the number of elements belonging to the group $i(i = 1, \ldots, s)$. The relative abundance vector $p = (p_1, p_2, \ldots, p_s)$ was obtained from the population, with $p_i = n_i/n$ and the population size $n = \sum_{i=1}^{s} n_i$.

We applied the three bootstrap techniques, percentile, -t and BCa to the new population, resampling from a multinomial distribution with parameters n, the population size and p, the relative abundance vector. For each method, Gini and Entropy indices were calculated and $B = 1000$ bootstrap replications were generated, and for the deviance estimation in the bootstrap-t, $M = 100$ second level bootstrap samples were used. Table 3 shows the results,

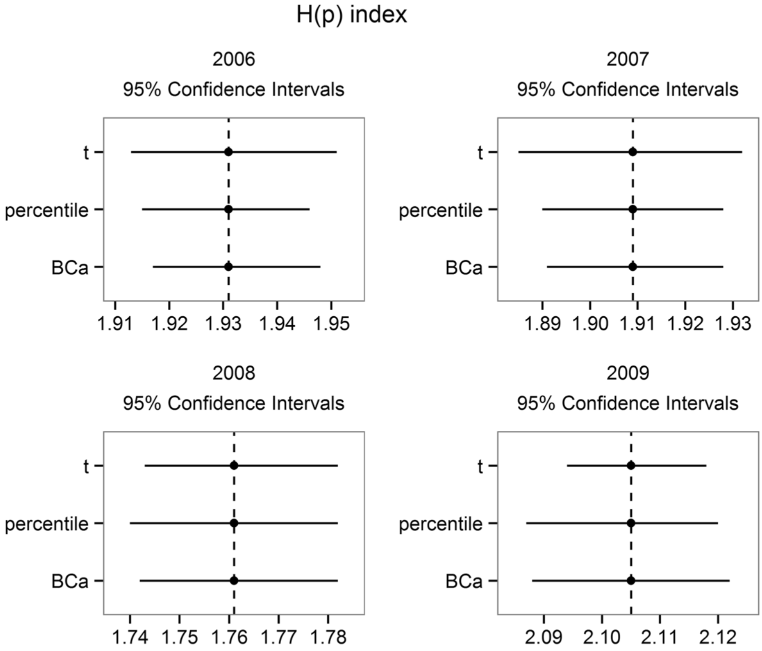

Fig. 2 Confidence intervals in the period 2006–2009 for the three bootstrap methods -t, percentile and BCa associated with the H(p) index

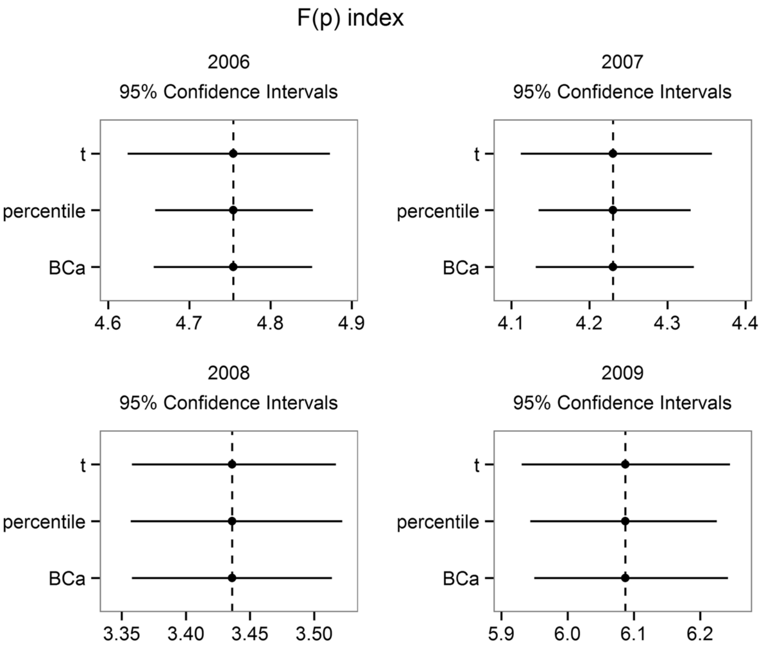

Fig. 3 Confidence intervals in the period 2006–2009 for the three bootstrap methods -t, percentile and BCa associated with the F(p) index

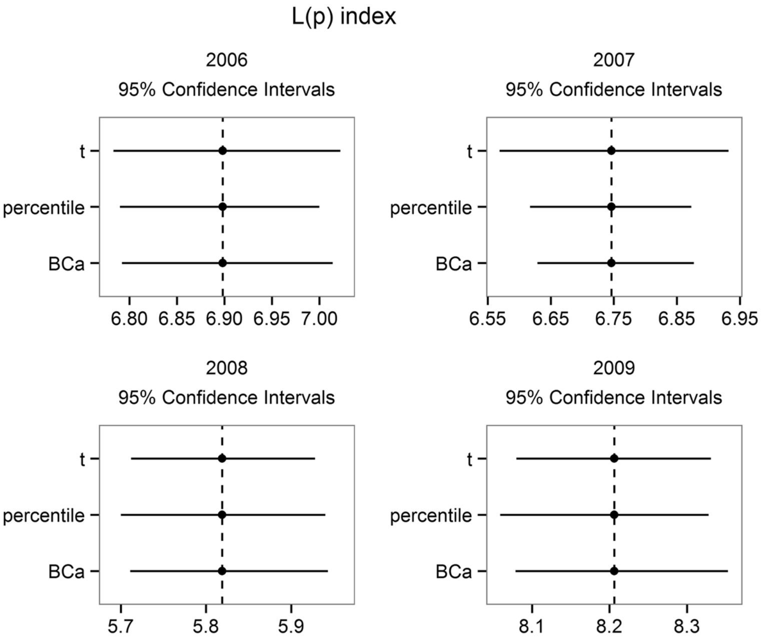

Fig. 4 Confidence intervals in the period 2006–2009 for the three bootstrap methods -t, percentile and BCa associated with the L(p) index

Table 3 Simulation study: Gini and Entropy index values and CI widths for each bootstrap method

	Value	Gini 0.9446	Entropy 3.0092
CI width	Percentile	0.0018	0.0218
	-t	0.0017	0.0208
	BCa	0.0017	0.0102

which are very similar to those obtained with JNCC data. In particular, the widths of the bootstrap CI are very small for the three methods and very close together for both of the indices.

7 Conclusions

In this paper, we have considered two indices of biodiversity commonly used in the context of ecological applications, such as the Gini–Simpson index and the Shannon–Wiener index. In addition, we have considered the effective number of species corresponding to these indices and given by the Laasko–Taagepera and the Leti diversity indices, respectively. Even though these indices are frequently adopted in ecological reports, it is not still in use the practice of showing the indices along with a measure of uncertainty, such as that provided by a CI.

Bootstrap methods represent a useful tool to estimate CI in case of complex indices, such as those previously described. In this paper we have considered three different techniques: percentile, bootstrap-*t* and BCa. Our simulations have shown that all three methods lead to quite convergent and consistent results, with the only exception, in some cases, of the -*t* method. Following these results we might recommend practitioners to adopt, in this context, the percentile method, as it is not computationally demanding and it is as well less demanding in terms of theoretical assumptions. More generally, these findings are in agreement with some simulation studies provided by Carpenter and Bithell (2000) and Tu and Shao (1995).

References

Carpenter, J., Bithell, J.: Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat. Med. **19**, 1141–1164 (2000)

Efron, B.: Better bootstrap confidence intervals. J. Am. Stat. Assoc. **82**, 171–185 (1987)

Frosini, B.V.: Descriptive measures of ecological diversity. In: Jureckova, J., El-Shaarawi, A.H. (eds.) Environmetrics. Encyclopedia of Life Support systems. Eolss Publishers, Oxford (2006)

Gini, C.: Variabilità e mutabilità. Studi economico-giuridici della Facoltà di Giurisprudenza dell'Università di Cagliari. Anno Terzo, Parte Seconda (1912)

Laakso, M., Taagepera, R.: The "Effective" number of parties: a measure with application to West Europe. Comp. Polit. Stud. **12**, 3–25 (1979)

Leti, G.: Statistica descrittiva. Bologna, Il Mulino (1983)

Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–432 (1948)

Shannon, C.E., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press, Urbana (1949)

Simpson, E.H.: Measurement of diversity. Nature **163**, 688 (1949)

Tu, D., Shao, J.: The Jackknife and the Bootstrap. Springer, New York (1995)

Wiener, N.: Cybernetics. Wiley, New York (1949)