Research Article GU J Sci 33(1): 229-236 (2020) DOI: 10.35378/gujs.554644

Gazi University

Journal of Science

http://dergipark.gov.tr/gujs

A New Proposed Estimator for Reducing Bias Due to Undetected Species

Kursad OZKAN*

Isparta University of Applied Science, Faculty of Forestry, Department of Soil Science and Ecology, 32200, Isparta, Turkey

Highlights

- There are undetected species in almost all inventories due to sampling limitation.
- Undetected species causes negative bias in estimation of species diversity.
- A new estimator ($\hat{H}_O$) is proposed to estimate true species diversity.

Article Info

Received: 16/04/2019
Accepted: 26/07/2019

Keywords

Bias-corrected estimator
Diversity
Entropy
Jackknife
Partially density matrix

Abstract

The present paper addresses a new approach to reduce bias when there are undetected species in a plot. Partially density matrix plays essential role in this new proposed estimator. The performance of the new proposed estimator ($\hat{H}_O$) was compared to bias-corrected MLE (MLE_{BC}), Jackknife (JK) and the proposed estimator of Chao and Shen ($\hat{H}_{CS}$) using Principle component analysis (PCA). According to the result of the first PCA applied to the data including the estimators' values of the assemblages, $\hat{H}_O$ is located between JK and $\hat{H}_{CS}$ and its' nearest neighbor becomes JK. The second PCA was applied to the data belonging to the relative estimator values between the pairwise assemblages and, it was found that $\hat{H}_O$ is still located between JK and $\hat{H}_{CS}$ but its' nearest neighbor becomes $\hat{H}_{CS}$ along the first axis at this time. Those results indicated that $\hat{H}_O$ is a better estimator than MLE_{BC}. Thus the new proposed estimator ($\hat{H}_O$) may also be used as an alternative bias-corrected estimator in addition to the other estimators.

1. INTRODUCTION

There is a wide variety of indices to measure biodiversity. In this sense, Shannon entropy, a theory for uncertainly measure introduced first by Claude Shannon [1], is the most well-known entropic measure [2].

Even though Shannon entropy is the most popular theory for biodiversity measure, rare events make Shannon entropy hard to estimate. There are no unbiased estimators of Shannon entropy for samples of fixed sizes. For incomplete samples, the main source of the bias comes from the undetected species.

Many correction methods have been proposed to obtain the reliable entropy estimator with less bias [3-11]. Applications of those correction methods require the number of individuals to estimate sample coverage.

With regard to the sample coverage approach to estimating the species number for an assemblage, assume that there are S species in a reference plot represented to that assemblage and they are labeled by 1,…, S. Let X_i be the number of the individuals of the ith species detected in the plot ($\sum_{i=1}^{S} X_i = n$). The probabilities of relative values of the species (X_i/n) are represented by $\pi_1, \pi_2, \dots, \pi_{S-1}, \pi_S$ $\left(\sum_{i=1}^{S} \pi_i = 1\right)$. Shannon entropy [1], is defined by

$$H = -\sum_{i=1}^{S} \pi_i \log(\pi_i) \tag{1}$$

Suppose that there are 100 species in a plot and 60 of them are singletons. Such plot data tell us that there should be undiscovered species in the plot due to many singletons. In this case, when H is computed using

*e-mail: kursadozkan@isparta.edu.tr

only the detected species, occurrence of sampling bias would be inevitable. That bias should therefore be corrected in order to compare the species diversity of the assemblage to those of the other assemblages.

As mentioned before, many bias corrected methods have been developed. One of the most popular one was built by Chao and Shen [7] for abundance data as an unbiased estimator ($\hat{H}$).

$$\widehat{H} = -\sum_{i=1}^{S} \frac{\tilde{\pi}_i \log \tilde{\pi}_i}{1-(1-\tilde{\pi}_i)^n} I(A_i) = -\sum_{i=1}^{S} \frac{\hat{C}\hat{\pi}_i \log \hat{C}\,\hat{\pi}_i}{1-(1-\hat{C}\hat{\pi}_i)^n} I(A_i) \tag{2}$$

$\hat{C}$ is the estimator of the sample coverage. It is computed by $\hat{C} = 1 - f_1/n$ where f_1 denotes the number of species recorded only once (singletons). $\tilde{\pi}_i = X_i/n$ is the MLE of π_i. $I(A_i)$ is the usual indicator function. If the event A_i is true, then $I(A_i) = 1$, otherwise $I(A_i) = 0$.

Apart from the proposed method by Chao and Shen [7], bias-corrected MLE has been also frequently used to reduce bias for abundance data.

With regard to bias-corrected MLE (MLE+$\hat{S}$), $\hat{S}$, reflects the estimator of total number of species (S), and it is computed by the following equation

$$\hat{S} = S_{abun} + \frac{S_{rare}}{\hat{C}_{rare}} + \frac{f_1}{\hat{C}_{rare}} \hat{\gamma}^2 \tag{3}$$

where

$$\hat{\gamma}^2 = \max\left\{\frac{S_{rare}}{\hat{C}_{rare}} \frac{\sum_{i=1}^{\kappa} i(i-1) f_i}{(\sum_{i=1}^{\kappa} i f_i)(\sum_{i=1}^{\kappa} i f_i - 1)} - 1,\ 0\right\}, \tag{4}$$

$\hat{\gamma}^2$ denotes the estimator of squared CV [12,13]. The total number of the abundant and rare species in the plot are $S_{abun} = \sum_{i=\kappa+1}^{n} f_i = \sum_{i=1}^{S} I[X_i > \kappa]$ and $S_{rare} = \sum_{i=1}^{\kappa} f_i = \sum_{i=1}^{S} I[0 < X_i \leq \kappa]$. In the equation, κ =10 is the value of cut-point [14]. $\sum_{k=1}^{n} k f_k = n$ and sample coverage for the assemblage is $\hat{C}_{rare} = 1 - f_1/\sum_{k=1}^{10} k f_k$ where f_1 denotes the singletons. f_k is the number of species that are those represented *i*th. value of the number of the species.

The other estimator is Jackknife methodology for bias correction. Chao and Shen [7] also used this estimator as proposed by Zahl [15] as well as bias-corrected MLE in comparison to the estimator they built.

$$JK = n \ln n - (n-1)\ln(n-1) + \frac{1}{n}\sum_{k=2}^{n} f_k k^2 \ln\frac{(k-1)}{k} - \frac{1}{n}\sum_{k=2}^{n} f_k k \ln(k-1) \tag{5}$$

where n denotes the number of individuals, f_k number of species that are represented exactly k times in the sample ($k = 1,2, \dots m$).

The present paper offers a new estimator based on partially mixed density operator. Using the same data given by Chao and Shen [7] from Janzen, [16,17], Batten [18] and Holst [19], the performance of the new proposed estimator is discussed by comparing to the proposed estimator by Chao and Shen, bias corrected MLE and Jackknife estimator.

2. PROPOSED ESTIMATOR

The density operator for quantum system $\{\ |\psi_i\rangle, p_i\}$ is $\rho \equiv \sum_i p_i \{\psi_i\rangle\langle\psi_i\}$ and von Neuman entropy $S(\rho)$ of a quantum state ρ is defined by the equation

$$S(\rho) \equiv -\mathrm{Tr}\big(\rho log(\rho)\big). \tag{6}$$

In information theory, the Shannon entropy is defined by $H(p) = -\sum_{i=1}^{n} p_i \log p_i$ where $\sum_{i=1}^{n} p_i = 1\ (0 < p_i \leq 1), p = (p_1, p_2, \dots p_n)$ is a probability distribution. If the eigenvalues of ρ are λ_i then the von Neumann entropy can be re-expressed as $S(\rho) = -\sum_{i=1}^{\infty} \lambda_i \log(\lambda_i) = H\big(\lambda(\rho)\big)$ [20].

The von Neumann entropy is zero for any pure state. The main difference between von Neumann entropy and Shannon entropy arises when considering a mixture of quantum states which are not orthogonal.

Partially mixed state is selected in order to estimate unbiased values in the present study. Consider a quantum system of two spin one half particles. One particle is in the z-up spin state and the other in the x-up state. From the classical Shannon point of view, we have a mixture with $p_1 = p_2 = 0.5$, giving and entropy of $-(0.5 \log 0.5) - (0.5 \log 0.5) = 1$. The partially density matrix in the z-presentation is given by:

$$\rho = \frac{1}{2}\ |\uparrow\rangle\langle\uparrow|\ + \frac{1}{2}\frac{(\ |\uparrow\rangle + |\downarrow\rangle)}{\sqrt{2}}\frac{(\langle\uparrow| + \langle\downarrow|\ )}{\sqrt{2}}. \tag{7}$$

The density matrix should be diagonalised to calculate von Neuman entropy (S_{vN}), i.e. to calculate its eigenvalues. That can be generated by solving secular equation $\begin{pmatrix} 0.75-\lambda & 0.25 \\ 0.25 & 0.25-\lambda \end{pmatrix} = 0$. This yields λ= 0.1464 or 0.8536, (i.e. $\lambda_{min} = 0.1464$ and $\lambda_{max} = 0.8536$) and, $S_{vN} = 0.6007$.

With regard to the new proposed estimator λ_{max} is the main parameter. Let f_k (abundance frequency counts), $k = 1,2, \dots, m,$ be the number of species that are detected exactly k times in the plot. In this case, $\hat{k}$ denotes the value to be added to the number of individuals of a species that is detected exactly k times in the plot. In other worlds, $\hat{k}$ is the estimated number of individuals of the species detected k times that referring to undetected species and it is computed by:

$$\hat{k} = \frac{\lambda_{max} k}{k^k}. \tag{8}$$

λ_{max} is calculated by the following equation

$$\rho_{f_k} = \frac{f_k}{S_{obs}}\ |\uparrow\rangle\langle\uparrow|\ + \frac{(S_{obs} - f_k)}{S_{obs}}\frac{(\ |\uparrow\rangle + |\downarrow\rangle)}{\sqrt{2}}\frac{(\langle\uparrow| + \langle\downarrow|\ )}{\sqrt{2}} \tag{9}$$

where S_{obs} denotes the total number of detected species in a plot.

Total value ($\hat{k}$) of the number of individuals of *i*th detected species and the estimated number of individuals of *i*th detected species obtained by referring to an undetected species is defined by the following equation.

$$\tilde{k} = \hat{k} + k\ . \tag{10}$$

When $\tilde{k}$ is found, $\tilde{p}_k$ is computed by:

$$\tilde{p}_k = \frac{\tilde{k}}{\sum_{i=1}^{S} f_k k + \sum_{i=1}^{S} f_k \hat{k}} = \frac{\tilde{k}}{\sum_{i=1}^{S} f_k \tilde{k}} \tag{11}$$

where $\tilde{p}_k$ is the relative value obtained from total value of the number of individuals, *i*th detected species and the estimated number of individuals of *i*th detected species that is referred to an undetected species.

Finally bias-corrected entropy ($\hat{H}_O$) is defined as:

$$\hat{H}_O = -\sum_{i=1}^{S} \frac{\tilde{p}_k \ln \tilde{p}_k}{1-(1-\tilde{p})^{\left(\sum_{i=1}^{S} f_k k + \sum_{i=1}^{S} f_k \hat{k}\right)}} I(A_i) = -\sum_{i=1}^{S} \frac{\tilde{p}_k \ln \tilde{p}_k}{1-(1-\tilde{p})^{\sum_{i=1}^{S} f_k \tilde{k}}} I(A_i) \tag{12}$$

where $\bar{p} = 1/S$ and $I(A_i)$ is the usual indicator function. If the event A_i is true, then $I(A_i) = 1$, $I(A_i) = 0$, otherwise. In this approach, the species corresponding to each of the k values, in particular doubletons play roles to reduce bias even though singletons have the highest contribution value.

Suppose that there are 59 species, with only one induvial ($f_1 = 59$) and, the number of the total species (S_{obs}) is 78 in a sample plot. The partially density matrix for f_1 is:

$$\rho_{f_1} = \frac{59}{78}\ |\uparrow\rangle\langle\uparrow|\ + \frac{(78-59)}{78}\frac{(\ |\uparrow\rangle + |\downarrow\rangle)}{\sqrt{2}}\frac{(\langle\uparrow| + \langle\downarrow|\ )}{\sqrt{2}}$$

$$= \frac{59}{78}\ |\uparrow\rangle\langle\uparrow|\ + \frac{19}{78}\frac{(\ |\uparrow\rangle + |\downarrow\rangle)}{\sqrt{2}}\frac{(\langle\uparrow| + \langle\downarrow|\ )}{\sqrt{2}}$$

$$= \begin{pmatrix} 0.878205 - \lambda & 0.121795 \\ 0.121795 & 0.121795 - \lambda \end{pmatrix}.$$

This yields $\lambda_{min} = 0{,}1026$ and $\lambda_{max} = 0{,}8974$. In this case, $\hat{k} = 0.8974 \times 1/1^1 = 0.8974$ and $\tilde{k} = 0.8974 + 1 = 1.8974$ for f_1. If $\sum_{i=1}^{S} \tilde{k}$ =185, then $\tilde{p}_1 = 0.0102$. This $\tilde{p}_1$ value is valid for all the singeletons. After computing $\tilde{p}_k$ for all k, bias corrected entropy ($\widehat{H}_O$) is found by using equation 12.

3. DATA EXAMPLES

The data including the frequency counts for beetles of DT (day-time) and NT (night-time) (Table 1) found in the paper of Chao and Shen [7] was obtained from "Osa primary-hill, dry season, 1967" in Janzen's papers [16,17]. As explained by Chao and Shen [7], for DT and NT plots, most of the species has only one, two or three individuals, and there are only a few abundant species (Table 1). It is clear that a relatively large fraction of species has been missed in the plots due to occurrences of most singletons in both of the plots.

As reported by Chao and Shen [7], MLE_{BC} has the lowest precision for DT and NT Since the bias-corrected MLE yields the highest estimate (Table 2). The new proposed estimator ($\widehat{H}_O$) has the closest value to the value of the proposed estimator of Chao and Shen ($\widehat{H}_{CS}$) for both of the plots (Table 2). In this case, it may be said that $\widehat{H}_O$ shows a much better precision than MLE_{BC} for both of those plots.

Table 1. *Frequency counts for beetles data (DT (Day-Time), NT (Night-Time)), bird data (EW (Endemic woodland), (CP) Conifer plantation) and coin data (OS (Observed site), RS (Reserve site)) from Chao and Shen [7]*

	DT		**NT**		**EW**		**CP**		**OS**		**RS**	
	k	f_k	k	f_k	k	f_k	k	f_k	k	f_k	k	f_k
	1	59	1	56	1	4	1	2	1	102	1	156
	2	9	2	9	2	3	2	1	2	26	2	19
	3	3	3	7	3	5	3	2	3	8	3	2
	4	2	5	2	5	1	4	1	4	2	4	1
	5	2	7	1	6	1	5	1	5	1	5	0
	6	2	10	1	11	1	9	1	6	1	6	0
	11	1	14	1	16	1	11	1	7	1	7	0
			16	1	21	1	14	1				
			18	1	25	1	20	1				
					26	1	30	2				
					35	1	65	1				
S_{obs}	78		79		20		14		141		178	
n	127		170		170		198		204		204	

With regard to the data taken from Batten [18] by Chao and Shen [7], abundant species is dominant in both of EW and CP plots (Table 1). As a result of this, Chao and Shen [7] concluded that there were two missing

species for the EW and only one species was undiscovered for the CP. It means that almost all species were discovered. As a result of this, all estimators and $\widehat{H}_O$ exhibit very close values to the value of MLE.

Coin data given by Holst [19] has been discussed previously by Chao and Lee [12]. The data for the two sides are given in Table 1. For two sides, as concluded by Chao and Lee [12], MLE is likely to be negatively biased. As can be seen in Table 2, $\widehat{H}_O$ yields the closest estimate to $\widehat{H}_{CS}$ for OS while it has almost the same value with the value of JK for RS data.

Table 2. *The results of MLE, bias-corrected MLE (MLEBC), Jackknife (JK), the proposed estimator of Chao and Shen ($\widehat{H}_{CS}$) and the new proposed estimator ($\widehat{H}_O$) for beetles data, bird data and coin data (note that the values of MLE, MLEBC, JK and ($\widehat{H}_{CS}$) were taken from Chao and Shen [7])*

Assemblages	***MLE**	***MLE$_{BC}$**	***JK**	***$\widehat{H}_{CS}$**	$\widehat{H}_O$
DT	4.08	5.11	4.62	4.7	**4.69**
NT	3.83	4.62	4.24	4.3	**4.33**
EW	2.41	2.47	2.48	2.49	**2.48**
CP	2.06	2.09	2.10	2.09	**2.09**
OS	4.80	5.72	5.41	5.56	**5.53**
RS	5.13	7.20	5.96	6.63	**5.96**

Principle component analysis (PCA) was applied to observe the locations of the estimators using the data found in Table 2. As a result of the PCA, eigenvalues of the axes were found 0.918376 (96.264%), 0.0301362 (3.1589%), 0.0053526 (0.56106%) and 0.000156992 (0.016456%) respectively.

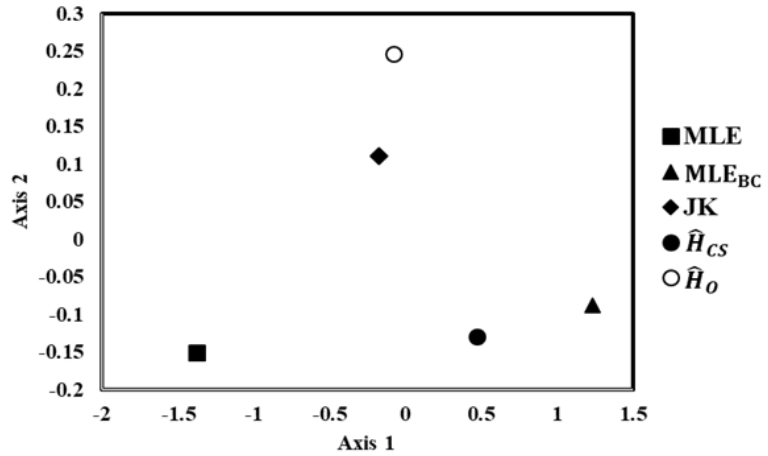

Figure 1. *PCA results of the estimators*

It is clear that most of the percentage of the total variance (96.264%) is explained by the first axis. According to the first axis, MLE is located in the left site opposite to the location of MLE$_{BC}$. $\widehat{H}_O$ is located between JK and $\widehat{H}_{CS}$ and its' nearest neighbor estimator is JK along the first axis.

The relative estimator values between the pairwise assemblages were also computed (Table 3). By using this data, the applied PCA results were given in Figure 1. The first axis explains 95.69% of the total variance. According to the first axis, $\widehat{H}_O$ is still located between JK and $\widehat{H}_{CS}$, but, instead of JK, its' nearest neighbor is $\widehat{H}_{CS}$ along the first axis in this time (Figure 1).

Table 3. *The relative values of the estimators in accordance with the pairs of the assemblages*

	MLE	**MLE$_{BC}$**	**JK**	$\widehat{H}_{CS}$	$\widehat{H}_O$
DT/NT	1.065274	1.106061	1.089623	1.093023	**1.081408**
DT/EW	1.692946	2.068826	1.862903	1.88755	**1.892368**
DT/CP	1.980583	2.444976	2.2	2.248804	**2.242596**
DT/OS	0.85	0.893357	0.853974	0.845324	**0.847976**

DT/RS	0.795322	0.709722	0.775168	0.708899	**0.787081**
NT/EW	1.859223	2.210526	2.019048	2.057416	**2.073773**
NT/CP	1.859223	2.210526	2.019048	2.057416	**2.073773**
NT/OS	0.797917	0.807692	0.783734	0.773381	**0.784140**
NT/RS	0.746589	0.641667	0.711409	0.648567	**0.727829**
EW/CP	1.169903	1.181818	1.180952	1.191388	**1.185074**
EW/OS	0.502083	0.431818	0.45841	0.447842	**0.448103**
EW/RS	0.469786	0.343056	0.416107	0.375566	**0.415924**
CP/OS	0.429167	0.365385	0.38817	0.375899	**0.378122**
CP/RS	0.401559	0.290278	0.352349	0.315234	**0.350969**
OS/RS	0.935673	0.794444	0.907718	0.838612	**0.928188**

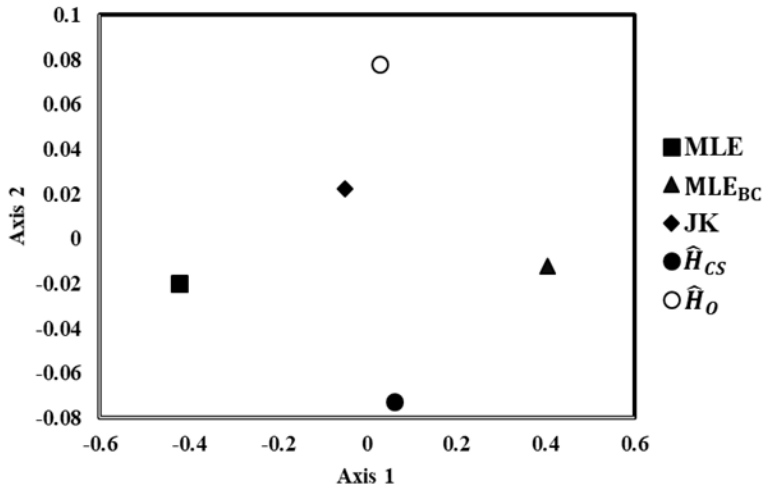

Figure 2. *PCA results of the proportional values between the estimators of the assemblages*

4. CONCLUDING REMARKS AND DISCUSSION

A new approach is proposed to reduce bias when there are undetected species in a plot. This new proposed estimator is based on partially mixed density matrix. If the species inventory is complete and all species are found in the plot, then the maximum likelihood approach (MLE) gives accurate results. Like the other estimators, (i.e. bias-corrected MLE (MLE$_{BC}$), Jackknife (JK) and the proposed estimator of Chao and Shen [7] ($\hat{H}_{CS}$)) the new proposed estimator ($\hat{H}_{O}$) computes a value, almost the identical value of MLE since a complete inventory includes many species and, very few of them are rare, meaning of slightly bias. However, such an inventory is not cost-effective and causes considerable time consuming. Bias-corrected estimators therefore play essential role to accurately measure biodiversity.

As reported by Chao and Shen [7], Chiu *et al*. [21] and Chao and Chiu [22], due to sampling limitation, there are undetected species in almost every species inventory. Thus the simple count of species in a plot underestimates the true species richness. To estimate accurate or true species richness, rare species are important because they yield the most information about the undetected species. Computations of the estimators (i.e., MLE$_{BC}$, JK and $\hat{H}_{CS}$) for obtaining the bias corrected values are based on this notion. The new proposed estimator was built considering the same notion. In this sense, the species with only one individual (singletons) in a reference plot give the most information about occurrences of another species that were not able to be detected during the inventory. In this approach, the other counts ($k=2, 3,\ldots,m$) in particular doubletons play also role to reduce bias.

The main difference of the new proposed estimator from the other estimators is that relative values ($\tilde{p}_k$) are computed by summing up the number of individuals of the species detected k times and the estimated number of individuals of the species detected k times that refers to undetected species.

The results belonging to DT, NT, EW, CP and OS of the new proposed estimator ($\widehat{H}_O$) seem to be very reasonable in comparison with the results of $\widehat{H}_{CS}$. The most important differentiation appears between the results of $\widehat{H}_O$ and $\widehat{H}_{CS}$ when OS data is used. However it is not unexpected result because $\widehat{H}_O$ is a different approach from $\widehat{H}_{CS}$ and it computes the values using its' self-algorithm.

According the PCA results, as can be seen in Figure 1 and Figure 2, $\widehat{H}_O$ is located between JK and $\widehat{H}_{CS}$ along the first axes that explains most of the total variance. However there is an important difference between the results of the two PCA. Namely, by using the data including the values of the estimators, the applied first PCA results shows that the nearest neighbor of $\widehat{H}_O$ is JK. That result may be evaluated that the accuracy of $\widehat{H}_O$ is insufficient. However, the relativeness among the estimated values is usually essential in particular biodiversity modelling and mapping attempts from ecological point of view. The second PCA was therefore applied to the data including the relative values of the estimators between the pairwise assemblages and it was found that instead of JK, the nearest neighbor of $\widehat{H}_O$ became $\widehat{H}_{CS}$.

As a conclusion, all the results indicate that $\widehat{H}_O$ is no fewer than MLE_{BC}. Thus the new proposed estimator ($\widehat{H}_O$) can also be used as an alternative bias-corrected estimator as well as the other estimators. This proposed estimator may be also used to estimate unbiased diversity using transformed data obtained from species abundance data since it takes account of all the species data found in an assemblage.

CONFLICTS OF INTEREST

No conflict of interest was declared by the author.

REFERENCES

[1] Shannon, C.E., "A mathematical theory of communication", Bell System Technical Journal, 27(3): 379-423, (1948).

[2] Gorelick, R., "Combining richness and abundance into a single diversity index using matrix analogues of Shannon's and Simpson's indices", Ecography, 29: 525-530, (2006).

[3] Miller, G., "Note on the bias of information estimates", Information Theory in Psychology: Problems and Methods, 95-100, (1955).

[4] Grassberger, P., "Finite sample corrections to entropy and dimension estimates", Physics Letters A, 128(6-7): 369-373, (1988).

[5] Holste, D., Grosse, I., Herzel, H., "Bayes' estimators of generalized entropies", Journal of Physics A: Mathematical and General, 31(11): 2551-2566, (1998).

[6] Grassberger, P., "Entropy Estimates from Insufficient Samplings", ArXiv Physics e-prints, 0307138, (2003).

[7] Chao, A., Shen, T.J., "Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample", Environmental and Ecological Statistics, 10(4): 429-443, (2003).

[8] Schurmann, T., "Bias analysis in entropy estimation", Journal of Physics A: Mathematical and Theoretical, 37(27): L295-L301, (2004).

[9] Bonachela, J.A., Hinrichsen, H., Munoz, M.A., "Entropy estimates of small data sets", Journal of Physics A: Mathematical and Theoretical, 41(20): 1-9, (2008).

[10] Zhang, Z., "Entropy Estimation in Turing's Perspective", Neural Computation, 24(5): 1368-1389, (2012).

[11] Chao, A., Wang, Y.T., Jost, L., "Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species", Methods in Ecology and Evolution , 4(11): 1091-1100, (2013).

[12] Chao, A., Lee, S.M., "Estimating the number of classes via sample coverage", Journal of the American Statistical Association, 87(417): 210-217, (1992).

[13] Chao, A., Hwang, W.H., Chen, Y.C., Kuo, C.Y., "Estimating the number of shared species in two communities", Statistica Sinica, 10: 227-246, (2000).

[14] Chao, A., Ma, M.C., Yang, M.C., "Stopping rules and estimation for recapture debugging with unequal failure rates", Biometrika, 80(1): 193-201 (1993).

[15] Zahl, S., "Jackknifing an index of diversity", Ecology, 58(4): 907-913, (1977).

[16] Janzen, D.H., "Sweep samples of tropical foliage insects: description of study sites, with data on species abundances and size distributions", Ecology, 54(3): 659-686, (1973a).

[17] Janzen, D.H., "Sweep samples of tropical foliage insects: effects of seasons, vegetation types, elevation, time of day, and insularity", Ecology, 54(3): 687-708, (1973b).

[18] Batten, L.A., "Bird communities of some Killarney woodlands", Proceedings of the Royal Irish Academy. Section B: Biological, geological, and chemical science. Royal Irish Academy, 76: 285-313, (1976).

[19] Holst, L., "Some asymptotic results for incomplete multinomial or Poisson samples", Scandinavian Journal of Statistics, 8: 243-246, (1981).

[20] Li, Y., Busch, P., "Von Neumann entropy and majorization", Journal of Mathematical Analysis and Applications 408(1): 384–393, (2013).

[21] Chiu, C.H., Wang, Y.T., Walther, B.A., Chao, A., "An improved nonparametric lower bound of species richness via a modified Good–Turing frequency formula", Biometrics, 70: 671-682, (2014).

[22] Chao, A., Chiu, C.H., "Species richness: estimation and comparison", Wiley StatsRef: Statistical Reference Online, 1-26, (2016).