





Manuel Mendoza¹, Eduardo Mendoza²  and E. Gutiérrez-Peña³ 

Method

Cite this article: Mendoza M, Mendoza E, and Gutiérrez-Peña E (2024). Statistical analysis of species association indices. *Journal of Tropical Ecology*. **40**(e12), 1–8. doi: <https://doi.org/10.1017/S0266467424000105>

Received: 13 July 2022

Revised: 21 November 2023

Accepted: 17 March 2024

Keywords:

Bayesian ecology; camera-trapping data analysis; mammal coexistence; presence–absence data

Corresponding author:

Eduardo Mendoza;

Email: eduardo.mendoza@umich.mx

¹Departamento de Estadística, Instituto Tecnológico Autónomo de México, Ciudad de México, México; ²Instituto de Investigaciones sobre los Recursos Naturales, Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Michoacán, México and ³Departamento de Probabilidad y Estadística, Universidad Nacional Autónoma de México, Ciudad de México, México

Abstract

The study of species association is of great interest in ecology due to its role in understanding key issues such as patterns of habitat use by animals, species coexistence, biotic interactions, and in general factors affecting *community structure and assembly*. There are many indices that ecologists commonly use, all based on the observed frequencies of organism occurrences, to evaluate the association between a pair of species. However, few of these indices correspond to proper statistical measures of association, and the inferential aspects of their analysis are often overlooked. In this paper, we propose a Bayesian approach based on a simple multinomial-Dirichlet structure to provide a comprehensive inferential framework for any set of association indices. Our approach provides a full statistical analysis for any association index of interest, free of special requirements on the sample size. We illustrate our procedure with a camera-trapping real-dataset, but the analysis of any other dataset of the same type can be readily produced using the R package *basa* that accompanies this paper.

Introduction

The question of whether two species tend to occur together, or finding one of them precludes the presence of the other, has occupied the attention of ecologists since the beginning of the past century (Dice 1945). The study of species association patterns constitutes a central issue in ecology due to its role in understanding how species interact and how these interactions are influenced by changes in biotic and abiotic factors, ultimately shaping community characteristics (Lavender *et al.* 2019). Because of this interest, a variety of approaches have been proposed to analyze the relevant information, each usually introducing a different numerical measure of association (Calvert 1922, Dice 1945, Forbes 1907, Michael 1920, Pielou 1967).

Hubálek (1982) reviewed 43 association coefficients and concluded that only 6 of them were admissible according to a set of properties that he determined. Legendre and Legendre (1998) reviewed several measures of association and proposed a classification of them (i.e., similarity, distance, and dependence), distinguishing between two types of input information: abundance and presence–absence of the species. Their list included several well-known statistical measures of dependence such as the Pearson, Kendall, and Spearman correlation coefficients. For each of these cases, they discussed the corresponding inferential procedure.

De Cáceres *et al.* (2008) used two association measures – the Pearson correlation coefficient and the Ochiai index – as fidelity measures to identify diagnostic species of vegetation communities. These authors showed that, under certain conditions, the Ochiai index can be seen as an asymptotic approximation to the Pearson coefficient. Moreover, they noted that only for the Pearson coefficient there exists an associated test of independence (which in this context is known as a non-faithfulness test). They suggested that, instead of a test of hypothesis, a confidence interval could be used to evaluate the strength of the fidelity measured by an index. In their paper, approximate intervals were obtained using bootstrap techniques. In a related study, De Cáceres and Legendre (2009) dealt with the problem of assessing the association between species and groups of sites. They addressed the problem of selecting an appropriate association index and emphasized the importance of inferential analysis in this setting. To determine if some species are associated with site groups, they proposed a significance test for different indices (e.g., the Pearson coefficient and the Indicator Value index). Due to the lack of suitable distributional theory, the authors proposed the use of permutation tests; they recognized, however, that in some instances this approach can lead to poor results. As in the previous case, they relied on bootstrap techniques to produce confidence intervals.

In practice, even though association measures play an important role in modern ecological research, there do not seem to be clear guidelines as to which index is more appropriate for a given study. Furthermore, once an index is chosen, little attention is paid to the range of values it could take if computed from similar data, not to mention the plausible values of other indices for

the same data. From a statistical point of view, an assessment of this uncertainty must be provided, and a common suggestion is to compute interval estimates. For traditional frequentist methods, it is often the case that the distributional theory – exact or asymptotic – necessary to obtain confidence intervals is not readily available. When comparing two indices, the use of joint confidence regions is even less common.

Hereafter, we will refer to the above-mentioned coefficients, measures, and indices simply as “indices” and will work only in the setting of presence–absence information. Also, we will use “association” instead of “joint occurrence,” “co-occurrence,” “coincidence,” or any other similar term. Regarding the analysis, we will make use of Bayesian statistical theory and show that a probability distribution can always be obtained that adequately describes the available information concerning a population association index. This so-called posterior distribution combines the observed data with a prior distribution that encapsulates the knowledge available before the study is conducted. The prior distribution could be used, for example, to suggest independence in a statistical sense or lack of association according to a specific index. The Bayesian analysis produces inferences – in particular, interval estimations – that do not require analytical approximations. Moreover, this approach allows us to deal simultaneously with several indices, and thus to gain insights into their joint behavior. Admittedly, in recent years, different options to analyze patterns of association of pairs of species have emerged, most of them associated with the increased suite of packages of the open-source program R. Some of them are mainly focused on providing new tools for assessing species association patterns such as the package *cooccur* (Griffith et al. 2016), whereas others provide miscellaneous functions for analyzing species association and niche overlap, such as *spaa* (Zhang and Ma 2014). However, these packages rely on the use of frequentist statistical methods – with the limitations mentioned above – and do not always offer the possibility to calculate some of the most common indices. The purpose of this paper is to introduce a statistical procedure that can be used to obtain probability intervals for any association index. This procedure allows the researcher to produce other inferences regarding the index of interest, such as pointwise estimates and tests of hypotheses.

The paper is organized as follows. In the next section, we first discuss the structure of the data and describe a Bayesian approach to the analysis of the multinomial model. We also discuss the difference between ecological measures of association and statistical measures of stochastic dependence. We close the section by applying these ideas to produce inferences on association indices. In “Association indices,” we illustrate our method in the context of the analysis of a real dataset concerning camera-trap data of two species inhabiting the *Montes Azules* Biosphere Reserve in Southern Mexico. In “Discussion,” we discuss these results and introduce a complementary analysis. Finally, “Concluding remarks” contains some concluding remarks.

Methodology

Data structure

In a general setting, if a number r of species is considered and presence–absence observations are collected at c different sites, the information can be organized in a $r \times c$ table which is a particular instance of the ecological data matrix as discussed in Legendre and Legendre (1998, Chapter 7). Although some authors have

Table 1. Presence–absence of species E_1 and E_2 .

		E_2		
		1	0	
E_1	1	n_{11}	n_{10}	$n_{1\cdot}$
	0	n_{01}	n_{00}	$n_{0\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 0}$	N

addressed the study of association when more than two species are simultaneously considered (see Pielou 1972, for example), most of the association indices are designed to measure the association between two species. Hence, we will concentrate on the case $r = 2$, where only two species, E_1 and E_2 , are studied. In this case, if a sample of N observations (presence–absence) is available at a selected collection of sites, the data are usually summarized in a 2×2 array as shown in Table 1.

In this table, the cells contain the frequencies of the only four possible cases, where: both species were observed (n_{11}); species E_1 was observed but species E_2 was absent (n_{10}); species E_1 was absent but the species E_2 was observed (n_{01}); and finally, when neither of the species was observed (n_{00}); here, $n_{11} + n_{10} + n_{01} + n_{00} = N$. In the margins of the table are also reported: $n_{1\cdot} = n_{11} + n_{10}$, the total number of cases where E_1 was present; $n_{0\cdot} = n_{01} + n_{00}$, the total number of cases where E_1 was absent, and, in a similar fashion, $n_{\cdot 1}$ and $n_{\cdot 0}$, the total number of cases where E_2 was present and absent, respectively. An equivalent summary is obtained if the observed frequencies are transformed into proportions, as shown in Table 2.

Here, we have $p_{ij} = n_{ij}/N$ for $i, j = 0, 1$; the observed proportions in the sample for the four possible cases. In any case, these sample data are used to describe the population they come from. For this purpose, statistical methods assume that $(n_{11}, n_{10}, n_{01}, n_{00})$ is generated by a probability model. To this end, each observation is recorded as a vector of dimension 4. If the observation is allocated to the cell (i, j) of Table 1, the associated vector X has three entries equal to zero, and one entry equal to one where the only nonzero entry corresponds to the position $4 - 2i + j$. Thus, the observed data are a collection of vectors X_1, \dots, X_N . These vectors are usually assumed to be randomly and independently sampled from the same population, so the statistic $\mathbf{X} = \sum_{k=1}^N X_k$ can also be written as $\mathbf{X} = (n_{11}, n_{10}, n_{01}, n_{00})$ and follows a multinomial distribution with parameters θ and N , where $\theta = (\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00})$ with $\theta_{ij} > 0$, $(i, j = 0, 1)$ and $\theta_{11} + \theta_{10} + \theta_{01} + \theta_{00} = 1$.

Thus:

$$p(\mathbf{x}|\theta) = \frac{N!}{\prod_{i=0}^1 \prod_{j=0}^1 n_{ij}!} \times \prod_{i=0}^1 \prod_{j=0}^1 \theta_{ij}^{n_{ij}}$$

For every nonnegative integer $n_{11}, n_{10}, n_{01}, n_{00}$ such that $n_{11} + n_{10} + n_{01} + n_{00} = N$. In this model, θ_{ij} represents the probability that an observation be allocated to the (i, j) -cell. Within this framework, the analysis aims to provide inferences about the population probabilities $\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00}$ using the observed frequencies in Table 1 or, equivalently, the observed proportions in Table 2.

Table 2. Observed proportions. Presence–absence of species E_1 and E_2 .

		E_2		
		1	0	
E_1	1	p_{11}	p_{10}	$p_{1\cdot}$
	0	p_{01}	p_{00}	$p_{0\cdot}$
		$p_{\cdot 1}$	$p_{\cdot 0}$	I

Bayesian analysis for our multinomial model

Bayesian inference requires a prior distribution for θ , the only unknown parameter of the model. This prior distribution is then updated with the sample information to obtain the posterior distribution, given by $p(\theta | \mathbf{x}) \propto p(\theta)p(\mathbf{x} | \theta)$. In the specific case of our multinomial model, we have

$$p(\theta | \mathbf{x}) \propto p(\theta) \prod_{i=0}^1 \prod_{j=0}^1 \theta_{ij}^{n_{ij}}$$

We can use any prior $p(\theta)$ in this expression if it describes the state of knowledge of the researcher before the sample information is available. A convenient class of models from which this prior can be chosen is the family of Dirichlet distributions,

$$p(\theta) \propto \prod_{i=0}^1 \prod_{j=0}^1 \theta_{ij}^{\alpha_{ij}-1}$$

where $\{\alpha_{ij} > 0; i, j = 0, 1\}$ is a set of fixed constants. In this model, it can be verified that $E(\theta_{ij}) = \alpha_{ij}/\alpha$ and $Var(\theta_{ij}) = [\alpha_{ij}(1 - \alpha_{ij})]/(\alpha + 1)$, where $\alpha = \sum_{i=0}^1 \sum_{j=0}^1 \alpha_{ij}$. Among other convenient features of the Dirichlet family of distributions, it is well known that it is *conjugate* to the multinomial sampling model. This means that, if the prior distribution is Dirichlet and the sample is distributed as multinomial, then the posterior distribution is also Dirichlet. Thus, in our case, we have the Dirichlet posterior:

$$p(\theta | \mathbf{x}) \propto \prod_{i=0}^1 \prod_{j=0}^1 \theta_{ij}^{(\alpha_{ij} + n_{ij}) - 1}$$

Of particular interest is the situation where the researcher has little prior information or is not willing or allowed to make inferences based on his/her prior beliefs. This can be addressed using a reference prior within the Dirichlet family. A common choice is to take $\alpha_{ij} = 1/2$ (see Box and Tiao 1973, Section 1.3, and Berger and Bernardo 1992). The corresponding posterior distribution is then:

$$p(\theta | \mathbf{x}) \propto \prod_{i=0}^1 \prod_{j=0}^1 \theta_{ij}^{(n_{ij} - 1/2)}$$

and thus $E(\theta_{ij} | \mathbf{x}) = (n_{ij} + 1/2)/(N + 2)$. This conjugate analysis in the general setting of contingency tables has been extensively discussed in the Bayesian literature. A recent review can be found in Gutiérrez-Peña and Mendoza (2017). Once the posterior distribution $p(\theta | \mathbf{x})$ is obtained, inferences about θ describing some of its characteristics (location measures as point estimates or probability intervals, for example) can be produced. These features can be obtained analytically, as in the case of the mean and variance, or computed via numerical or simulation-based methods. For the interested reader, a brief account of the general Bayesian

approach to inference is included in Section S.1 of the Supplementary Material.

Association indices

For the sake of simplicity, in this section, we will focus on the Ochiai and Pearson indices to illustrate our ideas, although it must be clear that the proposed analysis can be applied to any index. First, we will recall how the Ochiai and Pearson indices are calculated. For details, see Hubálek (1982).

- (i) Ochiai: $O = n_{11} / \sqrt{(n_{10} + n_{11})(n_{01} + n_{11})}$; $O \in [0, 1]$.
- (ii) Pearson: $R = (n_{00}n_{11} - n_{10}n_{01}) / \sqrt{(n_{0\cdot})(n_{1\cdot})(n_{\cdot 0})(n_{\cdot 1})}$; $R \in [-1, 1]$.

Each one of these indices provides a measure of the association between two species which can be used to assess the strength of such association *in the observed sample*. In the case of the Ochiai index, the closer the value of O is to one, the stronger the association. In fact, O reaches this maximum if and only if $n_{10} = n_{01} = 0$ (no discordant sites), whereas the minimum (zero) is attained when $n_{11} = 0$ if there is a positive number of discordant sites (otherwise it is not defined). The Pearson index also reaches its respective maximum value when there are no discordant sites. On the other hand, when the Ochiai index takes the value zero, and there are enough discordant sites, the Pearson coefficient can take a value close to -1 , suggesting a negative relationship. Conversely, if the Pearson coefficient is zero (suggesting no association), then the Ochiai index equals $\sqrt{n_{11}/N}$ which might be close to one. These results suggest that the assessment of the strength of the association of the species in the population, based on an observed value of a particular index in a specific sample, must be made with caution. The Supplementary Material (Section S.2.1) includes some numerical examples that illustrate these differences among indices.

In any case, it is easy to see that the indices considered in this section, as well as all others we reviewed in the literature, can be calculated in terms of the observed proportions instead of using the absolute frequencies. Take, as an example, the Ochiai index. According to the definition, $O = n_{11} / \sqrt{(n_{10} + n_{11})(n_{01} + n_{11})}$. However, if we divide by N both the numerator and denominator of this expression, we get $O = p_{11} / \sqrt{(p_{10} + p_{11})(p_{01} + p_{11})}$.

The most noticeable consequence of this result stems from the fact that p_{ij} is the observed proportion of cases in the (i, j) cell in the sample and, thus, can be seen as an estimate of the true proportion of cases in that cell for the entire population. If we were able to fully observe the populations of each of these two species, we would know the true population proportions (probabilities) θ_{11} , θ_{10} , θ_{01} and θ_{00} . However, since the investigation is only conducted through a sample of size N , instead of observing each θ_{ij} we obtain p_{ij} such that $p_{ij} = \hat{\theta}_{ij}$; $i, j = 0, 1$. More importantly, in the same way, since every index I can be calculated as $I = g(\hat{\theta}_{11}, \hat{\theta}_{10}, \hat{\theta}_{01}, \hat{\theta}_{00})$, for some known function g , it can be regarded as an estimate of the true value of the index in the population $\psi = g(\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00})$. If the purpose of the study is to investigate the pattern of association in the population, then the relevant question is: What can be said about ψ , given the information in the sample? The usual answer is to produce the pointwise estimate $\hat{\psi} = I$ and then to say that ψ has an approximate value equal to I . However, at this point, when a frequentist approach to statistical inference is adopted, the problems described in the introduction appear. It is necessary to

evaluate how far $\hat{\psi}$ is from ψ and which other values of ψ are compatible with the sample, in addition to I itself. As discussed above, the procedures used to answer these questions usually make use of interval estimates for ψ . The frequentist techniques rely on the sampling distribution of the estimate I , which is only available in a few cases. So, for many indices, the uncertainty regarding the true strength of the association in the population cannot be properly assessed, although some approximations can be used, such as those based on bootstrap methods. As an alternative, in the following subsection, we introduce a Bayesian solution for this problem, whose calculations are straightforward and provide a general mechanism to produce interval estimates for ψ and thus measure the uncertainty of interest.

Bayesian analysis of association indices

Consider an association index ψ . As we discussed previously, $\psi = g(\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00})$ is a known function of the (unknown) population parameter $\theta = (\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00})$. On the other hand, we have the sampling data $x = (n_{11}, n_{10}, n_{01}, n_{00})$. To produce a Bayesian analysis of ψ , the quantity of interest, we can proceed via a two-step procedure. First, we can get, as described in Section 2.2, the posterior distribution of θ . Then, in the second step, we can use the function $\psi = g(\theta)$ to get the posterior distribution $p(\psi | x)$, via the transformation technique. Recall that if only vague prior information is available regarding θ , the recommendation is to use the reference prior:

$$p(\theta) \propto \prod_{i=0}^1 \prod_{j=0}^1 \theta_{ij}^{-1/2}.$$

This prior leads to the posterior:

$$p(\theta | x) \propto \prod_{i=0}^1 \prod_{j=0}^1 \theta_{ij}^{(n_{ij}-1/2)},$$

a Dirichlet distribution with parameter $\alpha^* = (n_{11} + \frac{1}{2}, n_{10} + \frac{1}{2}, n_{01} + \frac{1}{2}, n_{00} + \frac{1}{2})$. Before discussing the calculation of the posterior distribution of interest, $p(\psi | x)$, let us note some characteristics of this reference prior. It can be shown that $E(\theta_{ij}) = \frac{1}{4}$, $E(\theta_i) = \frac{1}{2}$, $E(\theta_{\cdot j}) = \frac{1}{2}$; $i, j \in \{0, 1\}$ so that $E(\theta_{ij}) = E(\theta_i) \times E(\theta_{\cdot j})$ for every pair (i, j) . Thus, *a priori* the expected value of θ satisfies the condition for stochastic independence, which can be regarded as noninformative since any other condition would imply some particular kind of association and, hence, some specific prior information. If we use this reference prior for θ , any posterior evidence against independence can be attributed solely to the information provided by the sample. In fact, examination of the corresponding prior distribution of ψ may be useful to see the type of results that the index can produce under this scenario of independence. Going back to the posterior distribution of the index, since $\psi = g(\theta)$ we can describe the updated knowledge about the value of ψ just by calculating $p(\psi | x)$. Once we have this posterior distribution, and following De Cáceres et al. (2008), we can produce a probability interval for ψ . There are some indexes for which $p(\psi | x)$ can be obtained from $p(\theta | x)$ very easily using probability calculus. However, in some cases, the required analytic calculations may be cumbersome or may not be possible in closed form. Alternatively, we can use a Monte Carlo approach, simulating a sample $\{\theta_k\}_{k=1}^M$ from the Dirichlet model $p(\theta | x)$ and then obtaining a sample from $p(\psi | x)$ by applying the transformation directly to the simulated values of $(\psi_k = g(\theta_k))$, for each $k = 1, \dots, M$. For

M large enough, any characteristic of the distribution $p(\psi | x)$ can be approximated with an arbitrary level of precision using the simulated sample $\{\psi_k\}_{k=1}^M$. Simulation from $p(\theta | x)$ is rather easy using ratios of Gamma random variables (see Kotz et al. 2000, chapter 49, for example). Note that, when we simulate from this distribution, we will get a set of scenarios for the true population proportions that are compatible with the observed data. Moreover, the most probable scenarios, according to the data, will appear more frequently in the sequence $\{\theta_k\}_{k=1}^M$. In a similar way, if we calculate $\psi_k = g(\theta_k)$, for $k = 1, \dots, M$ the set $\{\psi_k\}_{k=1}^M$ is a collection of values of the true population index that are compatible with the information in the sample x and the most probable values will appear more frequently in this set. To produce interval estimates, if M is large enough (say several thousand) and we find that 95% of the simulated values lie, for example, in the interval $[0.78, 0.84]$, we will be able to say that the true value of the population index ψ lies in the interval $[0.78, 0.84]$ with probability 0.95. This sort of assertion clarifies the uncertainty that remains *a posteriori* about the value of ψ . It must be clear that the simulation algorithm in our proposal is just a tool to calculate, exactly, the posterior distribution $p(\psi | x)$. We are not using any bootstrap method; we are not resampling the observed data nor applying any form of a null model technique.

Our proposed Bayesian approach produces inferences that are, for all practical purposes, numerically exact for any sample size. These ideas will be illustrated in the following section. To conduct the corresponding analyses, we have developed the R package *basa* (Bayesian Analysis of Association Indices). This package automatically takes care of the simulation step and allows the user to calculate the posterior distribution and the relevant summaries for any association index, not just for those discussed in this paper.

Application to field data

As part of a project conducted in the *Montes Azules* biosphere reserve (Southern Mexico) in 2015, a camera-trap dataset recording the presence-absence of several species was generated. We will only be concerned with a specific pair of mammal species, namely the Central American agouti (*Dasyprocta punctata*) and the white-nosed coati (*Nasua narica*), for which we have a sample of $N = 40$ records (see Table 3). Also, we will focus on two association indices (Ochiai and Pearson); however, the analysis can be replicated for any other index. In all cases, we used a reference prior and a simulated sample of size $M = 10,000$ from the posterior distribution of θ to obtain the results. For the sake of comparison, we searched for a computation resource to produce a frequentist counterpart of our results. In the case of the Ochiai and Pearson indexes, we only could find the R package *spaa* (Zhang and Ma 2014), which only produces pointwise estimates. The comparison is then based on these estimates.

We compare the results obtained *a priori* with those obtained *a posteriori*. In this way, we can assess the impact of the data when we initially assume that there is no association.

(i) Ochiai index.

The prior distribution for this index is shown in Fig. 1(a) (left panel). Apart from a mode at zero, the density is rather flat over the interval $(0, 1)$. The prior mean and median are given by 0.453 and 0.440, respectively. A 0.95 interval estimate is given by $(0, 0.944)$, showing that the Ochiai index can be practically anywhere even though there is a clear mode at zero. Once the 40 observations are

Table 3. *Dasyprocta punctata* and *Nasua narica* camera trap data.

		<i>Dasyprocta punctata</i>		
		1	0	
<i>Nasua narica</i>	1	3	2	5
	0	19	16	35
		22	18	40

considered, the posterior distribution is that shown on the right panel of Fig. 1(a). The distribution is now bell-shaped, with a mean of 0.291 and a median of 0.288. Any one of these two values could be used as a point-wise estimate of the true value of the index. If the usual frequentist based Ochiai index is computed with the observed frequencies, we get 0.298, a quite similar value. In comparison, with our method we can say that *a posteriori*, the true value of the index lies in the interval (0.100, 0.498) with probability 0.95, thus providing more information and suggesting a moderate level of association.

(ii) Pearson index.

The prior and posterior distributions for the Pearson index are shown in Fig. 1(b). On the left-hand side, we have the prior distribution. It is symmetric around zero, with mode, mean, and median all equal to this value. The 0.95 prior interval estimate is given by (−0.903, 0.906). On the other hand, the posterior distribution (right-hand side) is a more concentrated bell-shaped curve with a mean of 0.026 and a median of 0.031, both quite like the value of the usual frequentist coefficient (0.038). Again, our method provides more information. We can say that *a posteriori*, the true correlation coefficient lies in the interval (−0.263, 0.285) with a probability of 0.95. This suggests that the two species have a low level of association and, since the value zero belongs to this interval, it can be reasonable to say that the posterior distribution suggests that there is no association between these two species.

Discussion

Ecological indices are designed to measure association (or lack thereof) differently compared to statistical indices, whose purpose is to measure stochastic dependence. The results obtained for the *Dasyprocta punctata* and the *Nasua narica* data make this point very clear. The posterior distribution of the Pearson statistical index shows a mode close to zero and is rather concentrated around this value, thus suggesting stochastic independence. This is confirmed by the Bayesian test of independence mentioned at the end of Section S.2.2 of the Supplementary Material.

On the other hand, the prior-posterior analysis of the Ochiai index shows that the prior mode at zero has been replaced by a positive value in the posterior distribution, and, for example, the probability of the interval (0, 0.1) changes from 0.188 *a priori* to only 0.012 *a posteriori*. Thus, it can be said that global analysis of these two indices points to a moderate level of association.

The posterior distribution of an association index can also be used to answer other questions of interest. For example, if the idea that $\psi \in A$ for a given set A has been entertained before the data are collected, the validity of the hypothesis $H_0 : \psi \in A$ can be evaluated

using its posterior probability, $P(\psi \in A | x)$. Specifically, in the case of our application regarding the Ochiai index, let us denote the population version of O as Ω . Then, if the hypothesis $H_0 : \Omega \geq 0.2$ was of interest, we could decide on its validity by considering that $P(\Omega \geq 0.2 | x) = 0.8$.

Furthermore, if we have two indices that are not functionally related and measure different kinds of association, it could be useful to analyze their *joint* posterior distribution. This type of analysis might shed some light on the complementary nature of the indices and is essentially unfeasible with the frequentist approach. This information can be obtained using the simulated sample $\{\theta_k\}_{k=1}^M$ from the posterior distribution θ . More specifically, if we have two indices, ψ_1 and ψ_2 , and for each θ_k in the sample we calculate $(\psi_1(\theta_k), \psi_2(\theta_k))$, then the set $\{(\psi_1(\theta_k), \psi_2(\theta_k))\}_{k=1}^M$ is a random sample from the joint distribution of (ψ_1, ψ_2) . To illustrate this, Fig. 2 shows the joint posterior distribution of the population version of (R, O) for the *Dasyprocta punctata* and *Nasua narica* data.

This graph shows that when the Pearson index takes relatively high values (suggesting a positive association among species), the Ochiai behaves similarly. On the other hand, when the Pearson index takes low values (evidence of a negative association), the Ochiai index takes values near zero which suggests no association. Finally, if the Pearson index takes values very close to zero, suggesting no association, the Ochiai index can take values between 0.2 and 0.4, which may be interpreted as evidence of a moderate positive association. It is interesting to compare this joint posterior distribution with those obtained for the two pairs of tree species discussed in Section S.2.2 of the Supplementary Material. There we examine the posterior distribution of the Ochiai index and the Pearson index corresponding to the pairs Bur oak and Black oak (Fig. 3a) and Red oak and American elm (Fig. 3b), respectively. The most relevant issue here is that there is more uncertainty regarding the relationship between the Ochiai index and the Pearson index in the case of the two tree datasets compared with our dataset concerning mammal species. This result is not surprising if we note that the tree datasets involve smaller sample sizes (10 observations against 40 in our data set). In any case, the relevant lesson here is that the single value of an index cannot be properly evaluated and interpreted without regard to the corresponding uncertainty which, among other factors, is related to the sample size.

We can say that ecological association and statistical association are not synonymous. In fact, some ecological indices suggest a high degree of association between species when, in statistical terms, there is independence. Similarly, for some datasets, the degree of ecological association can be low, but the corresponding statistical measures can indicate strong dependence. This fact is extremely interesting from a statistical point of view. It is also important because some of these ecological indices are widely used. See Kalgaotra *et al.* (2020), for example, where the Ochiai index is used to define edges between the nodes of a network in the context of medical research. A careful analysis of each ecological index would allow us to fully understand what kind of patterns it can detect. In any case, and regardless of the type of association that it describes, an ecological association index calculated from a sample is intended to shed light on the relationship between the species under study and is thus part of an inferential process. We believe this study can help promote a better understanding of the statistical basis of species association indices, thus facilitating a better interpretation of their ecological implications.

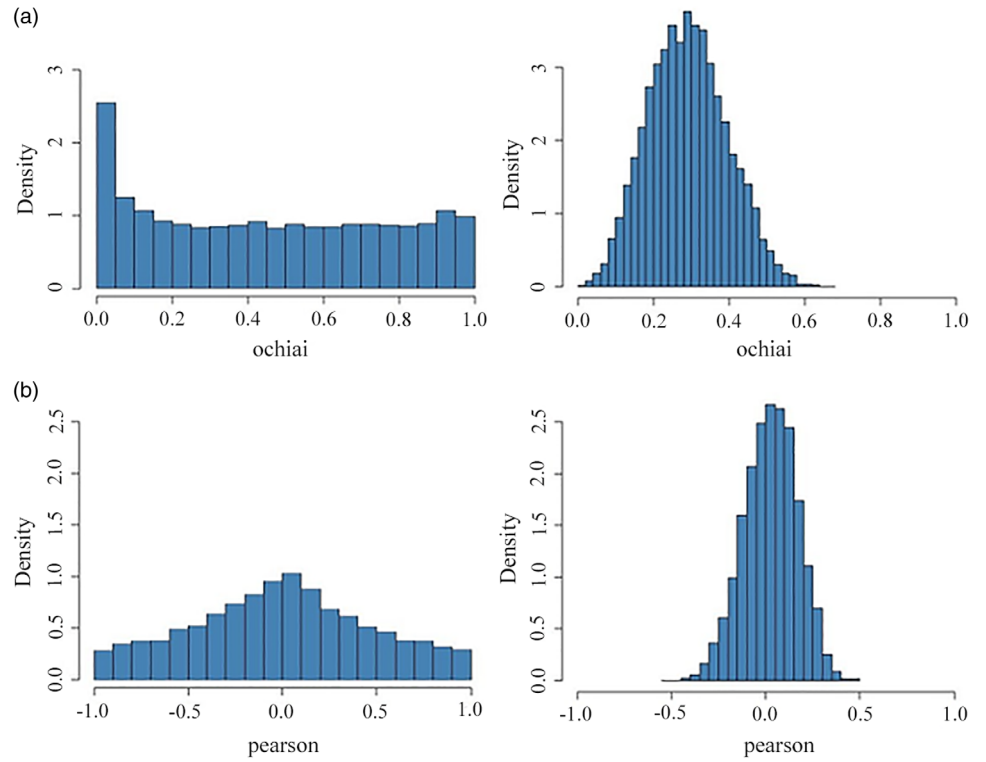


Figure 1. Prior and posterior distributions of the unknown population value of: (a) Ochiai index and (b) Pearson index. *Dasyprocta punctata* and *Nasua narica* data.

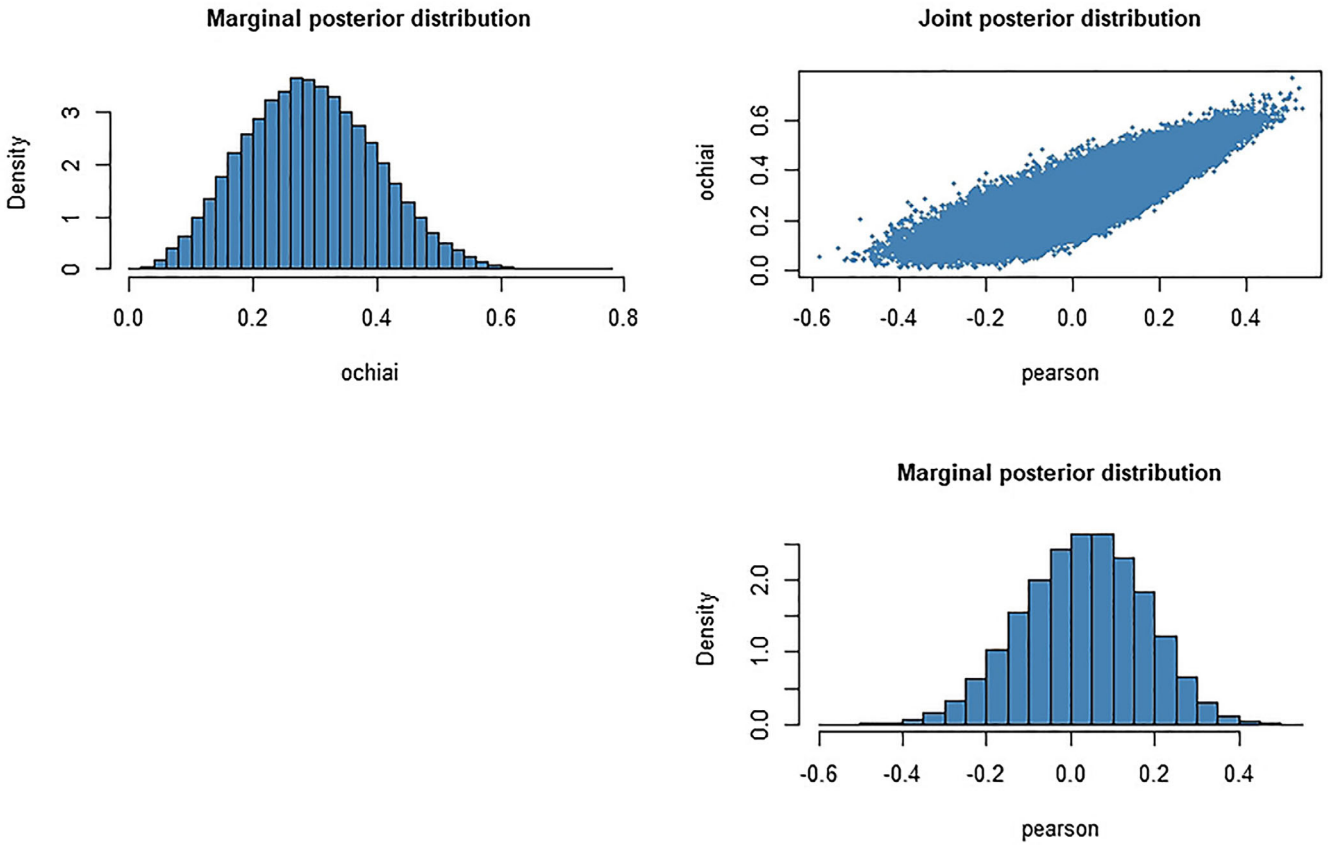


Figure 2. Joint posterior distribution of the unknown population values of the R (Pearson) and O (Ochiai) indices. *Dasyprocta punctata* and *Nasua narica* data.

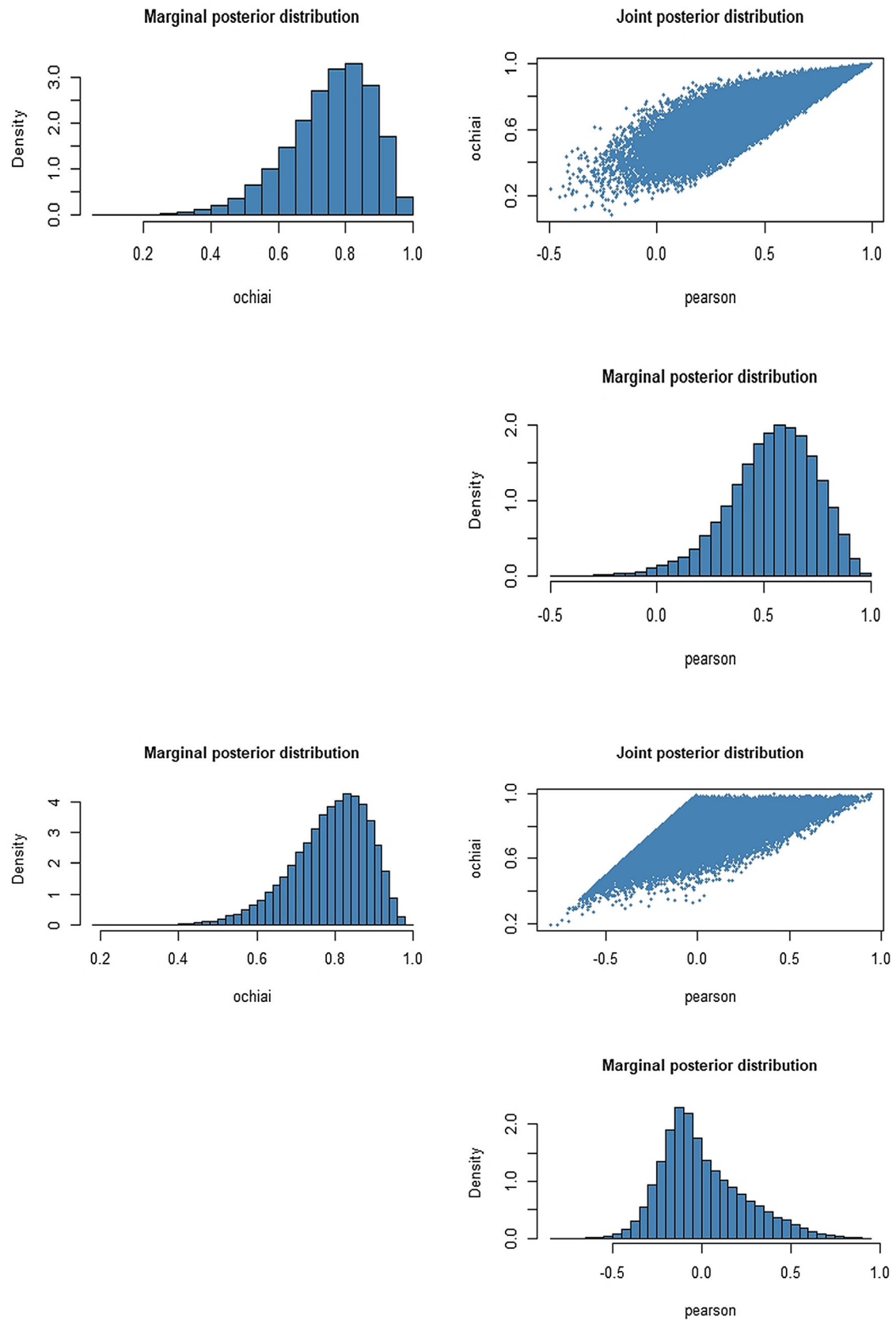


Figure 3. (a) Joint posterior distribution of the unknown population values of the R (Pearson) and O (Ochiai) indices. Bur oak and Black oak data; (b) joint posterior distribution of the unknown population values of the R (Pearson) and O (Ochiai) indices. Red oak and American elm data.

Concluding remarks

We have shown that a Dirichlet prior distribution for the probabilities of a multinomial model provides a suitable framework to easily obtain an adequate description of the uncertainty about any of the association indices. Our proposal uses a reference prior that does not depend on the index under study, but informative priors can be easily accommodated. In addition, the method proposed here also allows the researcher to obtain the joint posterior distribution of any set of indices. This is relevant because examination of the distribution for a pair of indices can be useful to understand what the indices are measuring, both individually and together. For example, the joint posterior distribution allows us to compute conditional interval estimates for one index given a fixed value of another index. In any case, the researchers who wish to explore the advantages of our proposal can do so using the R package *basa* available from: <http://labc-inirena.mx/bayesian-analysis-of-species-associations/>. A tutorial on the use of this package is included in the Supplementary Material (Section S.2.3).

Supplementary material. To view supplementary material for this article, please visit <https://doi.org/10.1017/S0266467424000105>

Availability of data and material. All data generated or analyzed during this study are included in this article.

Acknowledgments. The authors wish to thank the Sistema Nacional de Investigadores, México. The first author is also grateful to the Asociación Mexicana de Cultura, A.C.

Funding statement. This work was supported by Project IN106114–3 of the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (DGAPA-UNAM, México).

Competing interests. The authors declare no conflict of interest.

Code availability. The R package *basa* is available from <https://labc-inirena.mx/bayesian-analysis-of-species-associations/>

References

- Berger JO and Bernardo JM** (1992) Ordered group reference priors, with applications to multinomial problems. *Biometrika* **79**, 25–37.
- Box GEP and Tiao GC** (1973) *Bayesian Inference in Statistical Analysis*. Reading, MI: Addison-Wesley.
- Calvert PP** (1922) Methods for expressing the associations of different species. *Ecology* **3**, 163–165.
- de Cáceres M, Font X and Oliva F** (2008) Assessing species diagnostic value in large data sets: a comparison between Phi-coefficient and Ochiai index. *Journal of Vegetation Science* **19**, 779–788.
- de Cáceres M and Legendre P** (2009) Associations between species and groups of sites: indices and statistical inference. *Ecology* **90**, 3566–3574.
- Dice LR** (1945) Measures of the amount of ecologic association between species. *Ecology* **26**, 297–302.
- Forbes SA** (1907) On the local distribution of certain Illinois fishes: an essay on Statistical Ecology. *Bulletin of the Illinois State Laboratory of Natural History* **VII**, 273–351.
- Griffith DM, Veech JA and Marsh CJ** (2016) Cooccur: probabilistic species co-occurrence analysis in R. *Journal of Statistical Software* **69**, 1–17. <https://doi.org/10.18637/jss.v069.c02>
- Gutiérrez-Peña E and Mendoza M** (2017) *Bayesian methods for categorical data*. In *Wiley StatsRef: Statistics Reference Online*. <https://doi.org/10.1002/9781118445112.stat06134.pub2>
- Hubálek Z** (1982) Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. *Biological Reviews* **57**, 669–689.
- Kalgarota P, Sharda R and Luse A** (2020) Which similarity measure to use in network analysis: impact of sample size on phi correlation and Ochiai index. *International Journal of Information Management* **55**, 102–229.
- Kotz S, Balakrishnan N and Johnson NL** (2000) *Continuous Multivariate Distributions. Vol 1. Models and Applications*. New York: Wiley.
- Lavender TM, Schamp BS, Arnott SE and Rusak JA** (2019) A comparative evaluation of five common pairwise tests of species association. *Ecology* **100**, 1–8.
- Legendre P and Legendre L** (1998) *Numerical Ecology*. Amsterdam: Elsevier.
- Michael EL** (1920) Marine ecology and the coefficient of association: a plea in behalf of quantitative biology. *Journal of Ecology* **8**, 54–59.
- Pielou EC** (1967) The detection of different degrees of coexistence. *Journal of Theoretical Biology* **16**, 427–437.
- Pielou EC** (1972) 2^k Contingency tables in ecology. *Journal of Theoretical Biology* **34**, 337–352.
- Zhang JL and Ma KP** (2014) *spsaa*: an R package for computing species association and niche overlap. *Research Progress of Biodiversity Conservation in China* **X**, 165–174 (in Chinese).