## Analysis of species associations

**Statistical analysis of species association indices**

**Supplementary material**

# **S.1 A primer on Bayesian analysis**

Bayesian analysis has become a powerful tool for an increasing number of scientific disciplines. Ecology is not an exception and the literature on Bayesian methods in Ecology is growing. Two examples are the books by McCarthy (2007) and King et al. (2010). In this section, we describe the general Bayesian inference mechanism; its application to the case of an association index is discussed in the main text.

In any statistical inference problem, the main source of information is a set of observed data ***x****.* These data are always assumed to be a *random* sample, in the sense that it was generated according to a probability distribution . This model describes the variability that can be observed in the data given the value of the parameter . When its functional form of is known, the only quantity that must be studied to have a complete description of the underlying phenomena is precisely . Thus, in this parametric setting, we have a model that allows us to say what we can expect from the data given the (unknown) value of the parameter, and the goal of the researcher is to provide an answer to the *inverse* question. What can be said about the parameter given the data that have been observed? From a Bayesian point of view, the answer to this question is provided through another probability model, , so that for any set (an interval, for example) the researcher can calculate , the probability that belongs to , given the observed data . The main characteristic of this procedure is that is treated as a random variable even though it is a fixed (but unknown) constant.

The Bayesian theory conceives the probability as a general measure of uncertainty, which can describe both the variability in the random sample and the lack of knowledge about the fixed value of the parameter. As for the specific mechanism to obtain this model, Bayes’ formula provides the solution:

Since does not depend on , it is quite common to write , where ∝ denotes “is proportional to”. The so-called *posterior distribution* describes what the researcher knows about once the data are taken into account, and the calculation of is then feasible for any set . We stress that is an unknown constant. In any case, to produce the posterior distribution we need , the sampling model for the data, and , known as the *prior distribution* of . This prior model describes the knowledge of the researcher about *before* the data are available. The Bayes formula updates the prior distribution into the posterior distribution using the sampling model.

Sometimes, the researcher may have relevant knowledge about before de data are collected. This knowledge can be incorporated into the study by tailoring the prior distribution to represent that information properly. However, in other situations, only a limited amount of prior information may be available, or a posterior distribution mostly based on the sample data may be required for some reason. Under such circumstances, the prior distribution is chosen to produce a posterior distribution mainly shaped by the sample. Prior distributions of this type are usually known as non-informative or reference priors and are commonly used in scientific research.

# **S.2 Some illustrative examples of frequency tables and their corresponding association indices**

**S.2.1 Synthetic data**

Ecological indices measure a type of association that is not the same as stochastic dependence. In this section, we present four examples using synthetic data that illustrate how, for the same table of frequencies, some well-known ecological indices can suggest patterns that are different from those suggested by some popular statistical indices. The ecological indices considered are Ochiai, Dice, and Jaccard, whereas the statistical indices included for comparison are the Pearson correlation coefficient and the index.

Un reloj con números romanos

Descripción generada automáticamente con confianza mediaTable S1. Synthetic data for Example 1.

The first example is shown in Table S1, where the joint relative frequencies equal the product of the corresponding marginal relative frequencies. Consequently, these data strongly support the hypothesis of stochastic independence. In this case, the Pearson index is given by while equals .

On the other hand, the ecological indices take values that suggest a strong positive association. Specifically, (Ochiai), (Dice) and (Jaccard).

Imagen en blanco y negro de un reloj

Descripción generada automáticamente con confianza mediaTable S2. Synthetic data for Example 2.

The second example (Table S2) also satisfies the product rule of stochastic independence, even though the frequencies and have been exchanged for each other. Here again and , as expected. However, now we have , and . Permutation of the frequencies in the main diagonal of the table leads to ecological indices pointing towards no association.

Diagrama

Descripción generada automáticamente con confianza mediaTable S3. Synthetic data for Example 3.

Example 3 (Table S3) concentrates almost all of the cases on the main diagonal, with both and large compared with the other frequencies. Here, from a statistical point of view, we have a strong positive association since . On the other hand, the ecological indices take the following values: , and . This is an instance of agreement between the two types of indices.

Imagen que contiene objeto, reloj

Descripción generada automáticamenteTable S4. Synthetic data for Example 4.

The last example (Table S4) is quite similar to the previous one, although the large frequencies are now located along the other diagonal of the table and thus suggest a strong negative statistical association. and confirm this since both take the value . In this case, , and , strongly suggesting no association.

**S.2.2 Real data examples**

Here, we discuss two real data examples with a focus on the Ochiai and Pearson indices. These data come from Ludwigand Reynolds (1988, page 139), who study the presence-absence of eight species of trees at ten different locations (See Table S5). Note that we can get a array describing the association between any pair of species. Table S6 displays the Bur oak and Black oak data, whereas Table S7 shows the data for Red oak and American elm. The Ochiai index takes the same value for both tables, namely . On the other hand, the Pearson index is for Table S6 and for Table S7. These values were computed using the spaa R package (see Section S.1.4).

Tabla

Descripción generada automáticamenteTable S5. Ecological data matrix of presence-absences for eight trees in 10 upland forest sampling units, southern Wisconsin. Source: Ludwig and Reynolds (1988, page 139).

Diagrama

Descripción generada automáticamenteTable S6. Bur oak and Black oak, Wisconsin data (see Table S5).

Un reloj con números romanos

Descripción generada automáticamente con confianza media Table S7. Red oak and American elm, Wisconsin data (see Table S5).

The Bayesian analysis of these two cases yields the following results. For the data of Table S6, the point estimate of the Ochiai index is 0.75, with a 95% posterior probability interval given by (0.46, 0.94), while the point estimate of the Pearson index is 0.54 with a 95% posterior probability interval given by (0.09, 0.88). Now, for the data of Table S7, the point estimate of the Ochiai index is 0.78, with a 95% posterior probability interval given by (0.56, 0.94), while the point estimate of the Pearson index is 0 with a 95% posterior probability interval given by (-0.36, 0.55). These values were computed using our *basa* R package (see Section S.1.3).

The main conclusion is that the Ochiai index does not distinguish these two different cases. In fact, when analyzing these data, Ludwig and Reynolds compare with its (estimated) expected value under the hypothesis of independence. As a result, they conclude that since in the first case, the association may be positive between Burk oak and Black oak.

In the second case, given the opposite inequality, they report that the association between Red oak and American elm may be negative. Furthermore, these authors perform a Chi-square test of independence which, incidentally, may not be adequate given the small sample size and the occurrence of zeroes in the tables. They reject the independence hypothesis in the first case but not in the second case. For the sake of comparison, for each of these cases we carried out a Bayesian test of independence based on the mutual information (MI) as described in Gutiérrez-Peña and Mendoza (2017, Sect. 4.2). The results of these Bayesian tests are consistent with those reported by Ludwig and Reynolds (1988). The main point of these examples is to stress that the ecological indices measure some types of association between species that are different from the stochastic dependence described by the statistical indices.

Before closing this section, we note that the mutual information is defined as the Kullback-Leibler divergence between the joint density function and the product of the corresponding marginal density functions of two random variables (Cover and Thomas, 1991). It is always non-negative and is zero if and only if the two variables are stochastically independent. These properties make the MI a suitable statistical measure of association, and a good alternative to the Pearson correlation coefficient. We have implemented the MI in our R package *basa*.

**S.2.3 A short tutorial on the use of the *basa* R package**

Here we illustrate the use of the accompanying software, *basa*, using the Wisconsin tree data shown in Table S5 of this supplementary material. In particular, we show how to reproduce the results (including Figure 3) that support the discussion of Section 3.2 of the article. All the relevant files are available from the website https://lacb-inirena.mx/bayesian-analysis-of-species-associations/

**Data formatting**

We first load the *basa* package

> library(basa)

The file “trees.txt” contains the data shown in Table S5. We load this data into R

> pre.abs.data <- matrix(scan(file="trees.txt"),nrow=8,ncol=10,byrow=T)

and create a data frame

> pre.abs.df <-

data.frame(BurOak=pre.abs.data[1,],

BlackOak=pre.abs.data[2,],

WhiteOak=pre.abs.data[3,],

RedOak=pre.abs.data[4,],

AmericanElm=pre.abs.data[5,],

Basswood=pre.abs.data[6,],

Ironwood=pre.abs.data[7,],

SugarMaple=pre.abs.data[8,])

We can then display the data frame

> pre.abs.df

BurOak BlackOak WhiteOak RedOak AmericanElm Basswood Ironwood SugarMaple

1 1 1 1 1 1 0 0 0

2 1 1 1 1 1 0 0 0

3 1 1 1 0 1 0 0 0

4 1 1 1 1 1 0 0 0

5 1 0 1 1 1 1 0 0

6 0 0 1 1 0 1 0 1

7 1 0 1 1 1 1 1 1

8 0 0 1 1 0 1 1 1

9 0 0 0 1 1 1 1 1

10 0 0 1 1 1 1 1 1

**Analysis of all of the available indices for the Bur Oak and Black Oak data**

The following command will create a contingency table containing the frequencies corresponding to the Bur Oak and Black Oak species (see Table S6)

> BurOak.BlackOak.data <- con.tab(pre.abs.df["BurOak"],pre.abs.df["BlackOak"])

> BurOak.BlackOak.data

[1] 4 2 0 4

and similarly for the Red Oak and American Elm species (see Table S7).

> RedOak.AmericanElmOak.data <- con.tab(pre.abs.df["RedOak"],pre.abs.df["AmericanElm"])

> RedOak.AmericanElmOak.data

[1] 7 2 1 0

We then define the set of indices we wish to work with

> Indices <- c("dice","hamann","jaccard","mi","ochiai","pearson")

The following command display the posterior summary statistics than be used for the purposes of statistical inference on the association between Bur Oak and Black Oak

> results.table(BurOak.BlackOak.data,Indices)

Posterior summary statistics for each index:

mean median sd mad 2.5% 97.5%

dice 0.734 0.755 0.141 0.139 0.413 0.942

hamann 0.501 0.532 0.241 0.242 -0.040 0.880

jaccard 0.600 0.607 0.167 0.178 0.260 0.890

mi 0.192 0.177 0.121 0.129 0.006 0.457

ochiai 0.752 0.770 0.126 0.123 0.459 0.942

pearson 0.541 0.557 0.202 0.198 0.089 0.875

and similarly for Red Oak and American Elm

> results.table(RedOak.AmericanElm.data,Indices)

Posterior summary statistics for each index:

mean median sd mad 2.5% 97.5%

dice 0.780 0.795 0.104 0.100 0.539 0.939

hamann 0.334 0.359 0.261 0.270 -0.211 0.781

jaccard 0.651 0.659 0.135 0.143 0.369 0.884

mi 0.029 0.014 0.043 0.018 0.000 0.146

ochiai 0.787 0.801 0.100 0.097 0.558 0.938

pearson 0.005 -0.044 0.235 0.205 -0.360 0.552

In what follows, we will only illustrate the use of *basa* using the pair Bur Oak and Black Oak. Similar command can produce the corresponding results for the pair Red Oak and American Elm.

The following command displays the Pearson correlation matrix for pairs of indices

> corrmat.pearson(BurOak.BlackOak.data,Indices)

Pearson correlation matrix:

dice hamann jaccard mi ochiai

hamann 0.866 NA NA NA NA

jaccard 0.991 0.874 NA NA NA

mi 0.801 0.880 0.839 NA NA

ochiai 0.994 0.867 0.986 0.823 NA

pearson 0.853 0.947 0.853 0.941 0.875

and the corresponding (more robust) Spearman correlation matrix

> corrmat.spearman(BurOak.BlackOak.data,Indices)

Spearman correlation matrix:

dice hamann jaccard mi ochiai

hamann 0.883 NA NA NA NA

jaccard 1.000 0.885 NA NA NA

mi 0.848 0.923 0.840 NA NA

ochiai 0.997 0.879 0.997 0.859 NA

pearson 0.873 0.963 0.863 0.990 0.877

**Analysis of all pairs of species with Ochiai index**

*basa* is also able to provide summary statistics that allow one to carry our simultaneous inferences for all the species of interest for a given index. In the following example, we use the Ochiai index.

Posterior means and standard deviations of the Ochiai index for all species.

> posterior.means(pre.abs.df,"ochiai")

Posterior means for all species:

BurOak BlackOak WhiteOak RedOak AmericanElm Basswood Ironwood

BlackOak 0.755 NA NA NA NA NA NA

WhiteOak 0.770 0.626 NA NA NA NA NA

RedOak 0.651 0.484 0.846 NA NA NA NA

AmericanElm 0.813 0.662 0.786 0.786 NA NA NA

Basswood 0.348 0.082 0.650 0.771 0.559 NA NA

Ironwood 0.246 0.097 0.486 0.627 0.510 0.752 NA

SugarMaple 0.226 0.088 0.570 0.703 0.472 0.844 0.816

> posterior.sds(pre.abs.df,"ochiai")

Posterior standard deviations for all species:

BurOak BlackOak WhiteOak RedOak AmericanElm Basswood Ironwood

BlackOak 0.127 NA NA NA NA NA NA

WhiteOak 0.104 0.127 NA NA NA NA NA

RedOak 0.126 0.145 0.085 NA NA NA NA

AmericanElm 0.099 0.128 0.100 0.100 NA NA NA

Basswood 0.153 0.098 0.126 0.102 0.141 NA NA

Ironwood 0.151 0.110 0.143 0.127 0.148 0.128 NA

SugarMaple 0.140 0.106 0.136 0.117 0.147 0.104 0.120

Posterior medians and median absolute deviations of the Ochiai index for all species. These provide robust versions of the means and standard deviations displayed above.

> posterior.medians(pre.abs.df,"ochiai")

Posterior medians for all species:

BurOak BlackOak WhiteOak RedOak AmericanElm Basswood Ironwood

BlackOak 0.770 NA NA NA NA NA NA

WhiteOak 0.784 0.639 NA NA NA NA NA

RedOak 0.661 0.491 0.861 NA NA NA NA

AmericanElm 0.829 0.676 0.800 0.800 NA NA NA

Basswood 0.340 0.043 0.663 0.783 0.568 NA NA

Ironwood 0.225 0.054 0.490 0.637 0.520 0.771 NA

SugarMaple 0.202 0.048 0.585 0.714 0.471 0.866 0.838

> posterior.mads(pre.abs.df,"ochiai")

Posterior median absolute deviations for all species:

BurOak BlackOak WhiteOak RedOak AmericanElm Basswood Ironwood

BlackOak 0.1245 NA NA NA NA NA NA

WhiteOak 0.0983 0.129 NA NA NA NA NA

RedOak 0.1287 0.152 0.079 NA NA NA NA

AmericanElm 0.0939 0.132 0.097 0.095 NA NA NA

Basswood 0.1620 0.059 0.127 0.100 0.146 NA NA

Ironwood 0.1602 0.071 0.152 0.126 0.153 0.126 NA

SugarMaple 0.1473 0.064 0.143 0.117 0.160 0.094 0.111

**Joint analysis of the Ochiai and Pearson indices for the BurOak vs BlackOak data**

Finally, we show how to carry out the joint analysis of the Ochiai and Pearson indices using the *basa* function *psai2.*

The following command will compute and display marginal summary statistics (mean, standard deviation, median, median absolute deviation, credible intervals for each of the Ochiai and Pearson indices. In addition, it will display both Pearson and Spearman correlation coefficients between those indices and produce a graph that includes the histograms of the marginal posterior distributions of each index and a scatter plot representing the joint posterior distribution of both indices.

> psai2(ochiai,pearson,BurOak.BlackOak.data)

$Index1

$Index1$name

[1] "ochiai"

$Index1$mean

[1] 0.750

$Index1$median

[1] 0.768

$Index1$sd

[1] 0.128

$Index1$mad

[1] 0.127

$Index1$interval

2.5% 97.5%

0.455 0.943

$Index2

$Index2$name

[1] "pearson"

$Index2$mean

[1] 0.537

$Index2$median

[1] 0.554

$Index2$sd

[1] 0.202

$Index2$mad

[1] 0.201

$Index2$interval

2.5% 97.5%

0.0829 0.8652

$Correlation

$Correlation$pearson

[1] 0.871

$Correlation$spearman

[1] 0.880



Further details of these and all available functions can be found the *basa* manual, available at https://lacb-inirena.mx/bayesian-analysis-of-species-associations/.

**S.2.4 Comparison with a frequentist analysis using the *spaa* R package**

In this section, we re-analyze the Wisconsin tree data (shown in Table S5 of this supplementary material) using the *spaa* package (Zhang and Ma 2014).

The following command computes frequentist point estimates of various associations indices between each pair of species, including the V ratio, Jaccard, Ochiai, Dice and Pearson indices, among others. For the sake of comparison with the results of Section S.1.3, here we only display the results for the Ochiai and Pearson indices.

> sp.pair(pre.abs.df)

$Ochiai

BurOak BlackOak WhiteOak RedOak AmericanElm Basswood Ironwood SugarMaple

BurOak 1.000 0.816 0.816 0.680 0.866 0.333 0.204 0.183

BlackOak 0.816 1.000 0.667 0.500 0.707 0.000 0.000 0.000

WhiteOak 0.816 0.667 1.000 0.889 0.825 0.680 0.500 0.596

RedOak 0.680 0.500 0.889 1.000 0.825 0.816 0.667 0.745

AmericanElm 0.866 0.707 0.825 0.825 1.000 0.577 0.530 0.474

Basswood 0.333 0.000 0.680 0.816 0.577 1.000 0.816 0.913

Ironwood 0.204 0.000 0.500 0.667 0.530 0.816 1.000 0.894

SugarMaple 0.183 0.000 0.596 0.745 0.474 0.913 0.894 1.000

$Pearson

BurOak BlackOak WhiteOak RedOak AmericanElm Basswood Ironwood SugarMaple

BurOak 1.000 0.667 0.408 -0.272 0.612 -0.667 -0.583 -0.816

BlackOak 0.667 1.000 0.272 -0.408 0.408 -1.000 -0.667 -0.816

WhiteOak 0.408 0.272 1.000 -0.111 -0.167 -0.272 -0.408 -0.333

RedOak -0.272 -0.408 -0.111 1.000 -0.167 0.408 0.272 0.333

AmericanElm 0.612 0.408 -0.167 -0.167 1.000 -0.408 -0.102 -0.500

Basswood -0.667 -1.000 -0.272 0.408 -0.408 1.000 0.667 0.816

Ironwood -0.583 -0.667 -0.408 0.272 -0.102 0.667 1.000 0.816

SugarMaple -0.816 -0.816 -0.333 0.333 -0.500 0.816 0.816 1.000

It must be pointed out that *spaa* only provides point estimates of the association indices. No assessment of the variability of these values (standard deviations or confidence intervals) is provided by the software.

**References**

**Cover TM and Thomas JA** (1991). *Elements of Information Theory*. Wiley, New York.

**Gutiérrez-Peña E and Mendoza M** (2017). Bayesian Methods for Categorical Data. In *Wiley StatsRef: Statistics Reference Online*. DOI:10.1002/9781118445112.stat06134.pub2.

**King R, Morgan BJT, Gimenez O and Brooks SP** (2010). *Bayesian Analysis for Population Ecology*. Chapman & Hall, Boca Raton.

**Ludwig JA and Reynolds JF** (1988). *Statistical Ecology. A Primer on Methods and Computing*. Wiley, New York

**McCarthy MA** (2007). *Bayesian Methods for Ecology*. Cambridge University Press, Cambridge.

**R Core Team** (2020). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL [https://www.R-project.org/](http://www.R-project.org/).

**Zhang JL and Ma KP** (2014) *spaa*: An R package for computing species association andniche overlap. *Research Progress of Biodiversity Conservation in China* **X**, 165–174 (in Chinese)