UNIVERSITAT DE BARCELONA

GRAU DE MATEMÀTIQUES

Treball final de grau

RELATIONS BETWEEN ECOLOGICAL DIVERSITY INDICES

Autora: Gisela Espigulé

Tutor: Dr. Jan Graffelman,
Departament d'Estadística
i Investigació Operativa, UPC

Tutor UB: Dr. Josep Fortiana
Departament de
Matemàtiques i Informàtica

Realitzat a: Departament de
Matemàtiques i Informàtica

Barcelona, January 19, 2020

Abstract

The main goal of this work is to investigate if there is any explicit relationship between the Shannon index, widely used as an ecological diversity measure, and other diversity indices. The thesis focuses on defining the Shannon index rigorously and comparing it with two species abundance models: the broken stick model, and the geometric series model. Such relationships leads us to think of new methods to estimate biodiversity indices and define a new diversity measure.

2010 Mathematics Subject Classification. 92D40, 92B05, 92B10, 92B15, 62P10, 62P12

Contents

Chapter 1

Introduction

The research project I present is motivated by the theory one can read in the widely cited book *Measuring Biological Diversity* (Magurran, 2004). However, in this thesis we have done simulations to better enlight the topics discussed. We will start by introducing rigorously the population and estimation of the well known Shannon index and its properties. Then we will try to find explicit relationship between this index and two of the most popular biological diversity models: the geometric series and the broken stick model.

The motivation for choosing this topic for my bachelor thesis comes from my stay at the University of Tromsø (UiT, Norway), where I took a course in Computer Intensive Statistics. It was there that I saw the opportunity to apply mathematics to biological studies - something that has always peaked my interests. Thus, after this Erasmus exchange I took the course Statistics for Biosciences at the Faculty of Mathematics and Statistics (UPC), where I asked professor Jan Graffelman for a topic for my bachelor thesis. I wanted something related to biodiversity from a mathematical point of view. While choosing the topic, some of the questions were "Why is the Shannon index so important in biodiversity studies?" "How does it work with empirical data?" "Are there any explicit relationships between the Shannon index and some biodiversity models worked in class?"

1.1 Objectives

Magurran states that *"most diversity measures are not explicitly associated with named species abundance models"* (Magurran, 2004). This statement motivated our main goal: **Study whether there are any explicit relationship between the Shannon index and other diversity indices.**

In addition, the other tasks of this thesis are:

- Deepen our understanding of the Shannon index and its behaviour with empirical and simulated data.
- Work on the Geometric series and the Broken stick model.
- Estimate parameters of the biodiversity models using the maximum likelihood method.
- Simulate biodiversity data.
- Apply the theory to empirical data.

1.2 Outline

We will start by introducing some history and theory regarding the basic concepts related to biodiversity. In the third chapter we will introduce two different graphics for visualizing empirical biodiversity data which will be used in the following chapters. Chapter four introduces Shannon index and its properties together with a simulation study. This will enable us to better understand this index in order to properly compare it with the biological models. The fifth chapter deals with the analytical task of finding explicit relationship between the Shannon index and the Broken stick and geometric models. In the last chapter, we apply the concepts discussed in the previous chapters to empirical data. This will give us an idea of how and where to apply the knowledge explained to biodiversity studies.

Chapter 2

History and background

2.1 What is Biodiversity?

***Biodiversity** means the variability among living organisms from all sources: inter alia, terrestrial, marine and other aquatic systems and the ecological complexes of which they are part. This includes diversity within and between species and of ecosystems* (United Nations Environment Program, 2019).

The term biodiversity, which has the same meaning of ‘biological diversity’, encompasses the variety of biological life at more than one scale. It is not only the variety of species (both plant and animal), but also the variety of genes within species and the variety of ecosystems in which the species reside. In this project we will be centered in the variety of species or **species diversity**.

The relevance of this subject in the scientific community is illustrated in Figure 2.1, where we have represented the number of papers published using the term biodiversity or biological diversity during the last three decades.

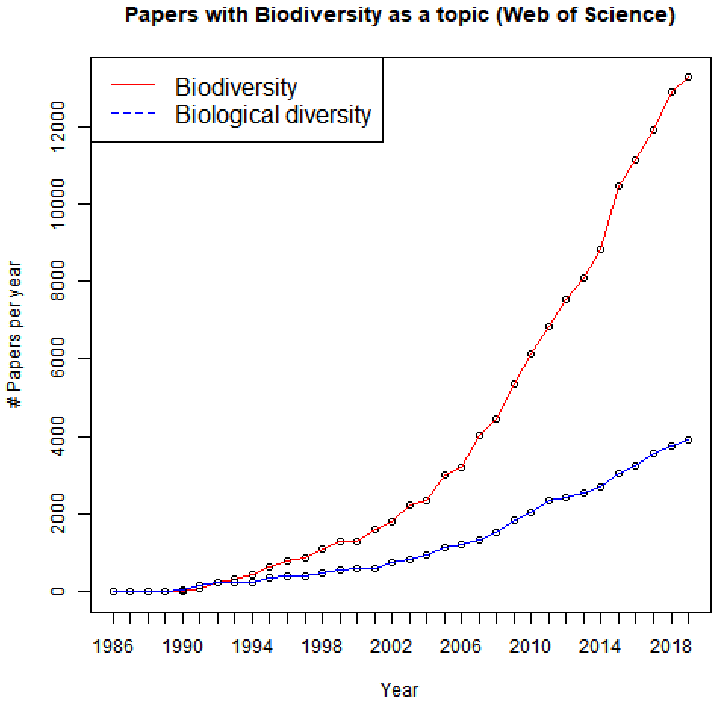


Figure 2.1: The increasing use of the words *biodiversity* and *biological diversity* in published papers indexed in the *Web of Science* between 1985 and 2019.

2.2 How to measure Biodiversity?

Biodiversity studies are often of comparative nature, and the purpose of these studies are often to compare or rank communities, or to assess whether diversity has changed over time (Graffelman, 2018).

General ecology concepts

The ecological concepts that we will encounter while measuring biodiversity are the following :

- **Taxa**: Group where the species are classified into.
- **Species abundance**: The total number of individuals in an area, population, or community. Relative abundance refers to the total number

of individuals of a taxa compared with the total number of individuals in an area, population, or community.

- **Species richness**: The number of species within a given sample, community, or area.
- **Species evenness**: The uniformity of abundance between species in a community.
- **Assemblage**: A collection of species inhabiting a given area, the interactions between the species, if any, being unspecified.
- **Ecosystem**: A dynamic complex of plant, animal and micro-organism communities and their non-living environment interacting as a functional unit.
- **Niche**: The way a species makes a living "profession" role a species plays in a community.
- **Diversity index (or measure)**: A single statistic that incorporates information on richness and evenness.

(United Nations Environment Program, 2019)

Diversity measures

There are many different diversity indices. Each index measures certain components of diversity, such as richness and evenness of a collection of species.

We cite some of the most popular diversity measures, which are not necessarily the best:

- The Shannon index (H') was introduced by Ramon Margalef, who was the first to apply the Communication Theory of Claude Shannon to ecology studies (Margalef, 1957), (Shannon, 1948).
- The Simpson index (D), which is less used than the Shannon index, but considered to be one of the most robust diversity measure available (Simpson, 1949).
- The Margalef diversity index (D_{Mg}) to estimate richness (Margalef, 1974).

- The k parameter of the geometric model (Magurran, 2004).
- The α parameter of the Fisher's log series (Fisher et al., 1943).

In this project we will focus on the Shannon index and the parameter k of the geometric model.

Species abundance models

There is a vast range of species abundance models. They attempt to mathematically describe the relationship between the number of species and the number of individuals in those species. (Magurran, 2004) We cite some of the most popular species abundance models:

- The broken stick model (MacArthur, 1957).
- The geometric model (He and Tang, 2008).
- The Tokeshi's model (Tokeshi, 1993).
- The Fisher's logarithmic series (Fisher et al., 1943).
- The log normal distribution (Preston, 1948).

In this project we will focus on the broken stick model and the geometric model.

Chapter 3

Graphics for biodiversity

In this chapter, we will introduce two graphics for visualizing the species abundance distribution in different ways. These methods are specially useful for showing how abundance models are fitted to empirical data, which will be explained in Chapter 6. Graphics are central for interpreting the data distribution, which is a common first step in any ecological study. Each graphic emphasizes a different characteristic of the species abundance distribution. We subsequently present the rank-abundance plot and the frequency distribution.

3.1 Rank/abundance plot

One of the best known and most informative methods is the **rank/abundance plot** (Krebs, 1989), also referred to as the **whittaker plot**. Species are plotted in sequence from most to least abundant along the x axis. Their abundances are typically displayed in a logarithmic scale, such that species whose abundance span several orders of magnitude can be easily accommodated in the same graph. The shape of the rank-abundance plot is often used to infer which species abundance model best describes the data (Magurran, 2004).

Example 1. *This example illustrates how a rank-abundance plot looks like while using some empirical data. We plot the abundance distribution of 6815 individuals of the insects group of Macrolepidoptera, caught in light traps in the UK. There are 197 different species (Lewis et al., 1967).*

individuals	species	individuals	species	individuals	species
1	37	21	4	69	1
2	22	22	1	73	1
3	12	23	1	75	1
4	12	25	1	83	1
5	11	28	2	87	1
6	11	29	2	88	1
7	6	33	2	105	1
8	4	34	2	115	1
9	3	38	1	131	1
10	5	39	1	139	1
11	2	40	3	173	1
12	4	42	2	200	1
13	2	48	2	223	1
14	3	51	1	232	1
15	2	52	1	294	1
16	2	53	1	323	1
17	4	58	1	603	1
18	2	61	1	1799	1
20	4	64	2		

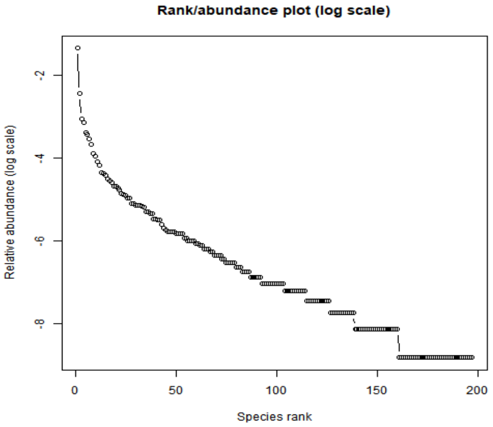
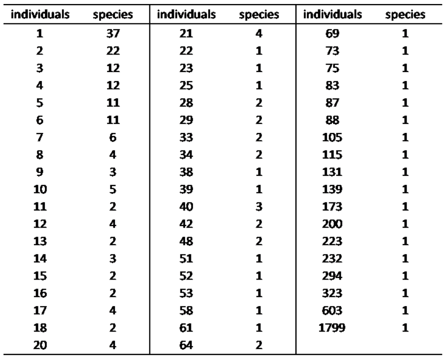


Figure 3.1: Abundance distribution of *Macrolepidoptera* individuals caught in light traps in the UK.

3.2 Frequency distribution

This plotting method is more useful when fitting the log series model to the data. A **frequency distribution** has the number of species on the y axis, that are displayed in relation to the number of individuals per species. A variant of this plot is typically employed when the log normal is chosen. Sometimes the abundance classes of the x axis are presented on a log scale (Magurran, 2004).

Example 2. *We plot now the same data as the example of the Macrolepidoptera in the section before, but in a frequency distribution plot.*



individuals	species	individuals	species	individuals	species
1	37	21	4	69	1
2	22	22	1	73	1
3	12	23	1	75	1
4	12	25	1	83	1
5	11	28	2	87	1
6	11	29	2	88	1
7	6	33	2	105	1
8	4	34	2	115	1
9	3	38	1	131	1
10	5	39	1	139	1
11	2	40	3	173	1
12	4	42	2	200	1
13	2	48	2	223	1
14	3	51	1	232	1
15	2	52	1	294	1
16	2	53	1	323	1
17	4	58	1	603	1
18	2	61	1	1799	1
20	4	64	2		

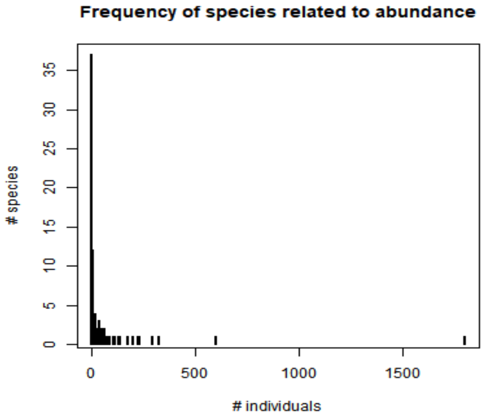


Figure 3.2: Frequency distribution of *Macrolepidoptera* individuals caught in light traps in the UK.

Chapter 4

The Shannon index

In this chapter, we present the Shannon index. We divide it into three parts: 1) Shannon index definition, 2) Estimation of the Shannon index, 3) Shannon index evenness measure, and 4) A simulation study of the Shannon index.

4.1 Shannon index, $\mathcal{H}$

The Shannon index was originally defined to describe information entropy. However, it has been applied in many other fields such as in ecology where it has been widely used for measuring ecological diversity. It was first defined by Claude Shannon in 1948 (Shannon, 1948).

Definition 1 (**Shannon index,** $\mathcal{H}$)**.** *Given a vector of probabilities* $\theta = (\theta_1, \ldots, \theta_S)$*, such that* $\sum_{i=1}^{S} \theta_i = 1$*,* $\mathcal{H}$ *is defined as the following sum:*

$$\mathcal{H} := -\sum_{i=1}^{S} \theta_i ln(\theta_i)$$

4.2 Shannon index estimation, H'

In order to define the estimation of the Shannon index in an ecological context, we have set some variables and assumptions:

Variables

- $N := \#individuals$

- $S := \#species$
- $n_i := \#individuals\ of\ the\ ith\ species\ (abundance\ of\ each\ species)$

Assumptions

The assumptions made while estimating the Shannon index (in this context) are:

1. The abundance of individuals follows a **multinomial** distribution $X = (X_1, \ldots, X_S) \sim MN(N, \theta_1, ..., \theta_S)$, where $0 \leq \theta_i \leq 1$.
2. We estimate the true value of θ_i by the maximum likelihood (ML) method (Pielou, 1969).
3. The true number of species S, present in the assemblage under study, is assumed to be known, and the number N of sampled individuals is determined in advance.

Estimation of θ

Given these assumptions, we calculate the ML estimator of θ by applying the method of *Lagrange multipliers* (Arfken and Weber, 2005):

- **Probability mass function** of the multinomial distribution.

$$P(X_1 = n_1, ..., X_S = n_S) = \frac{N!}{n_1! \ldots n_S!} \theta_1^{n_1} \ldots \theta_S^{n_S},\ 0 \leq \theta_i \leq 1$$

- **Likelihood function**

$$L(\theta | n_1, ..., n_S) = \frac{N!}{n_1! \ldots n_S!} \theta_1^{n_1} \ldots \theta_S^{n_S}$$

- **Calculation of the maximum likelihood estimator of θ**
  We use k to denote the multinomial coefficent:

$$\frac{N!}{n_1! \ldots n_S!}$$

$$\begin{aligned}\log L(\theta|n_1,...,n_S) &= \log(k) + n_1\log(\theta_1) + \cdots + n_S\log(n_S)\\ &= \log(k) + \sum_{i=1}^{S} n_i \log(\theta_i)\end{aligned}$$

We define $f(\theta, n_1, ..., n_S, \lambda) = \log(k) + \sum_{i=1}^{S} n_i\log(\theta_i) + \lambda\left(1 - \sum_{i=1}^{S}\theta_i\right)$

$$\begin{aligned}&\frac{\partial f}{\partial\lambda} = 1 - \sum_{i=1}^{S}\theta_i \Rightarrow \sum_{i=1}^{S}\theta_i = 1\\ &\frac{\partial f}{\partial\theta_i} = \frac{n_i}{\theta_i} - \lambda = 0 \Rightarrow \lambda = \frac{n_i}{\theta_i} \Rightarrow n_i = \lambda\theta_i \;, \forall i = 1, ..., S\end{aligned}$$

Then, since $\sum_{i=1}^{S}\theta_i = 1$ and $N = \sum_{i=1}^{S} n_i$, we have

$$N = \sum_{i=1}^{S} n_i = \sum_{i=1}^{S}\lambda\theta_i = \lambda \Rightarrow \frac{n_i}{\theta_i} - N = 0 \Rightarrow \hat{\theta}_i = \frac{n_i}{N}$$

Now we prove $\hat{\theta}_i = \frac{n_i}{N}$ is a maximum:

$$\frac{\partial^2 f}{\partial\theta_i^2} = -\frac{n_i}{\theta_i^2} \leq 0$$

Finally, we calculate H' by applying the *Principle of functional invariance* (Márquez and Julià, 2011) and that $\hat{\theta}_{iML} = n_i/N$:

$$H' = \hat{\mathcal{H}}_{ML} = -\sum_{i=1}^{S}\hat{\theta}_i ln(\hat{\theta}_i) = -\sum_{i=1}^{S}\frac{n_i}{N}\ln\frac{n_i}{N}$$

Observation. *We observe that $\hat{\theta}_i = \frac{n_i}{N}$ is an unbiased estimator of θ_i:*

$$E(\hat{\theta}_i) = E\left(\frac{n_i}{N}\right) = \frac{E(n_i)}{N} = \frac{N\theta_i}{N} = \theta_i \;,\; \textit{since } X_i \sim B(N,\ \theta_i)\ (\textit{Sanz iSolé},\ 1999)$$

Now, we are ready to define the estimation of the Shannon index broadly used in Ecology.

Definition 2 (Shannon index estimation, H'). *Given a set of N individuals grouped in S species and given the value n_i as the abundance of each species. We can calculate the Shannon index, denoted by H':*

$$H' := -\sum_{i=1}^{S} p_i \ln(p_i)$$

where $p_i := \frac{n_i}{N}$ is the proportion of individuals found in the ith species.

Note. *From now on, we will say "Shannon index" to refer to the sample estimate of the Shannon index, H'.*

Properties. *Important properties of the Shannon index:*

1. *$H' \geq 0$.*
2. *$H'_{max} = ln(S)$.*
3. *$E(H') = -\sum_{i=1}^{S} \theta_i \ln(\theta_i) - \frac{S-1}{2N} + \frac{1-\sum \theta_i^{-1}}{12N^2} + \frac{\sum(\theta_i^{-1}-\theta_i^{-2})}{12N^3} + \dots$*
4. *$Var(H') \approx \frac{\sum_{i=1}^{S} p_i(ln(p_i))^2-(\sum_{i=1}^{S} p_i ln(p_i))^2}{N} + \frac{S-1}{N^2}$.*
5. *H' has a normal distribution for large samples.*
6. *H' captures richness and evenness of species.*
7. *For empirical data, H' often varies from 1.5 to 3.5.*

Proof. We will prove the first two properties and make some comments on the rest of them.

1. We know that all $p_i \in [0, 1]$ in the case there are no individuals for a certain species, $p_i = 0$ for some i, the value of the corresponding summand $0 \ln(0)$ is taken to be 0, which is consistent with the limit:

   $$\lim_{p_i \to 0^+} p_i \ln(p_i) = 0$$

   Thus, abundance 0 does not contribute.

2. We want to see that H' reaches its maximum at the value $p_i = \frac{1}{S}$. First, we use *Lagrange multipliers* (Arfken and Weber, 2005) to maximize the Shannon index:

   $$\begin{aligned} f(p_i, \lambda) &= -\sum_{i=1}^{S} p_i \ln(p_i) + \lambda(1 - \sum_{i=1}^{S} p_i) \\ \frac{\partial f}{\partial \lambda} &= 1 - \sum_{i=1}^{S} p_i = 0 \\ \frac{\partial f}{\partial p_i} &= (-1)\ln p_i + (-p_i)\frac{1}{p_i} - \lambda = -\ln p_i - 1 - \lambda = 0 \Rightarrow p_i = e^{-1-\lambda} \\ \frac{\partial^2 f}{\partial p_i^2} &= -\frac{1}{p_i} < 0 \end{aligned}$$

   Then, since $\sum_{i=1}^{S} p_i = 1$, we have

   $$\sum_{i=1}^{S} p_i = \sum_{i=1}^{S} e^{-1-\lambda} = Se^{-1-\lambda} = Sp_i = 1 \Rightarrow p_i = \frac{1}{S} \tag{4.1}$$

   Thus, if $p_i = \frac{1}{S}$, then H'_{max} is:

   $$H'_{max} = -\sum_{i=1}^{S} \frac{1}{S} \ln(\frac{1}{S}) = -S\frac{1}{S} \ln(\frac{1}{S}) = -\ln(\frac{1}{S}) = \ln(S)$$

3. From (Peet, 1974) and (Bowman et al., 1971) we know the expansion of the first moment of H' up to fourth order. We also derived $E(H')$ up to second order. See Appendix A.

4. The $Var(H')$ is known to be $\frac{\sum_{i=1}^{S} p_i(ln(p_i))^2 - (\sum_{i=1}^{S} p_i ln(p_i))^2}{N} + \frac{S-1}{N^2}$ (Hutcheson, 1970) and (Bowman et al., 1971), but it is not derived here.

5. Due to the maximum likelihood estimators which are known to be asymptotically normal and unbiased (Márquez and Julià, 2011).

6. - It captures richness because if there is a higher number of species S, then the value of H' increases.

- It captures evenness because we can divide H' by the maximum value H'_{max} and then we get J' (that we define later) which gives a value of how differently distributed are the species. Communities with higher values of J' means they are more even.

7. Here we have to refer to the ecologist from Barcelona Ramon Margalef, 1972, who this is, in practice, the range. (Margalef, 1957).

□

Observation. *We saw that p_i is the maximum likelihood estimator of the true value of θ_i. This approximation produces a biased result for the Shannon index as we don't know the true value of N and S. Then, H' should be obtained from the $E(H') \approx -\sum \theta_i \ln(\theta_i) - \frac{S-1}{2N} + \frac{1-\sum \theta_i^{-1}}{12N^2}$ (Peet, 1974). In practice, however, this error is rarely significant. A more substantial source of error arises when the sample does not include all the species in the community, as we commented before, it is almost impossible to know the true value of S in an empirical study.*

4.3 The Shannon evenness measure J'

Here, we introduce another widely used measure of biodiversity: the Shannon evenness measure, which captures the evenness of a sample. And it is defined as it follows.

Definition 3 (Shannon evenness measure, J').

$$J' := \frac{H'}{H_{max}} = \frac{H'}{ln(S)}$$

Observation. *Observe the values that J' can take are between 0 and 1, $J' \in [0.1]$.*

Example 3. *This example is to get some idea of what has been explained in this section. In the following table we see five different samples with $N = 50$ individuals differently distributed, which means the Shannon and the Shannon evenness indexes have different values in each case. When we calculate J', S is assumed known.*

Sample	*sp1*	*sp2*	*sp3*	*sp4*	*sp5*	*N*	H'	J'
1	*10*	*10*	*10*	*10*	*10*	*50*	*1.609*	*1.000*
2	*50*	*0*	*0*	*0*	*0*	*50*	*0.000*	*0.000*
3	*49*	*1*	*0*	*0*	*0*	*50*	*0.098*	*0.061*
4	*48*	*1*	*1*	*0*	*0*	*50*	*0.196*	*0.122*
5	*30*	*10*	*5*	*3*	*2*	*50*	*1.156*	*0.718*

$$H' = -\sum_{i=1}^{5} p_i ln(p_i) \ ; \ J' = \frac{H'}{ln(5)}$$

The highest value of H' is reached in sample 1 where each species has the same number of individuals and there are more species than in sample 2, 3 and 4; actually, sample 1 has the maximum H' possible in a 5-species sample. Similarly the one with the lowest value of H' is sample 2 with only one species and $H' = J' = 0$.

4.4 Simulation of the Shannon index in different scenarios

The goal of this simulation study is to investigate the statistical properties of H'. Focusing on the difference between the Shannon index estimations H' and their respective population Shannon indices $\mathcal{H}$.

Here, the simulation methods will be implemented using R programming language (R Core Team, 2016).

Design of the simulation

We are going to simulate three different scenarios that differ in the species abundance distribution. Table 4.1 shows the population distribution in each case, where

- θ_i= the probability of an individual to belong to the ith species, such that $\sum_{i=1}^{S} \theta_i = 1$.
- $S = 10$ (number of species).
- $N = 1000$ (number of individuals).

- $\mathcal{H}$ = the population Shannon index.

Table 4.1: **Three different population distributions**

Case	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}	S	$\mathcal{H}$
1	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	10	2.302585
2	0.25	0.15	0.13	0.11	0.10	0.09	0.07	0.05	0.03	0.02	10	2.105516
3	0.60	0.20	0.10	0.04	0.03	0.01	0.01	0.008	0.001	0.001	10	1.237139

Figure 4.1 and Figure 4.2 represent the rank/abundance plots. The first case is evenly distributed, the second case some species are more dominant than others and the third case half of the species are dominant and the other half are rare.

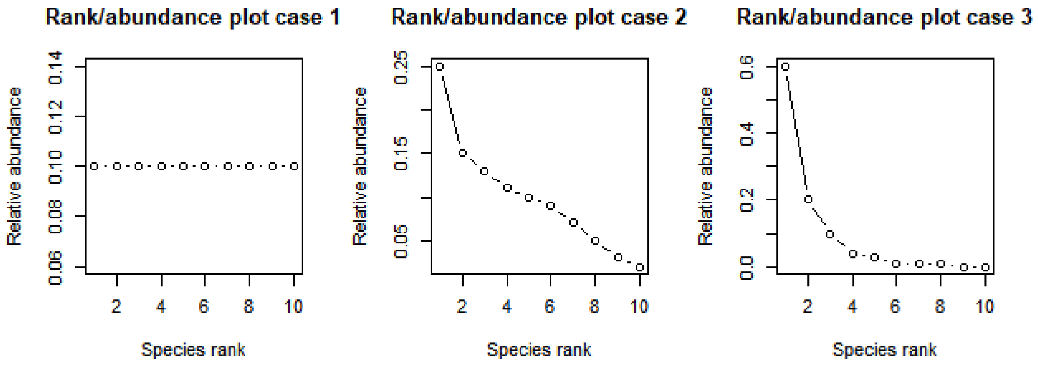


Figure 4.1: Rank/abundance plots.

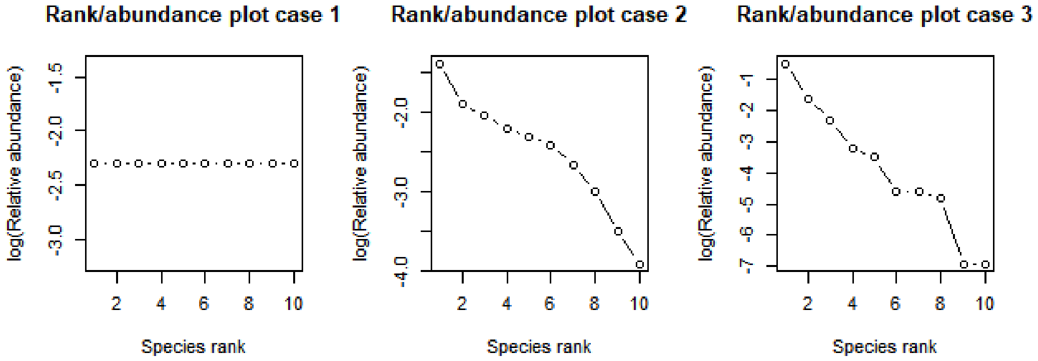


Figure 4.2: Rank/abundance plots in a natural log scale.

Running the simulation

For each distribution, we simulate 10.000 random variations of the distribution using an `R` function called `rmultinom` (R Core Team, 2016). This function is explained in Appendix A.

From each simulation we obtain an estimated Shannon index and we calculate the mean of them, noted by $\overline{H'}$, which will be compared with the population Shannon index, $\mathcal{H}$.

Finally, for each case, we plot the estimated Shannon indices histogram along with the mean and the population Shannon index.

Results

We observe that for all cases the mean, $\overline{H'}$, underestimates the population Shannon index $\mathcal{H}$. We see this in Figure 4.3 , where the histograms are plotted along with the mean in green colour and the population Shannon index in red.

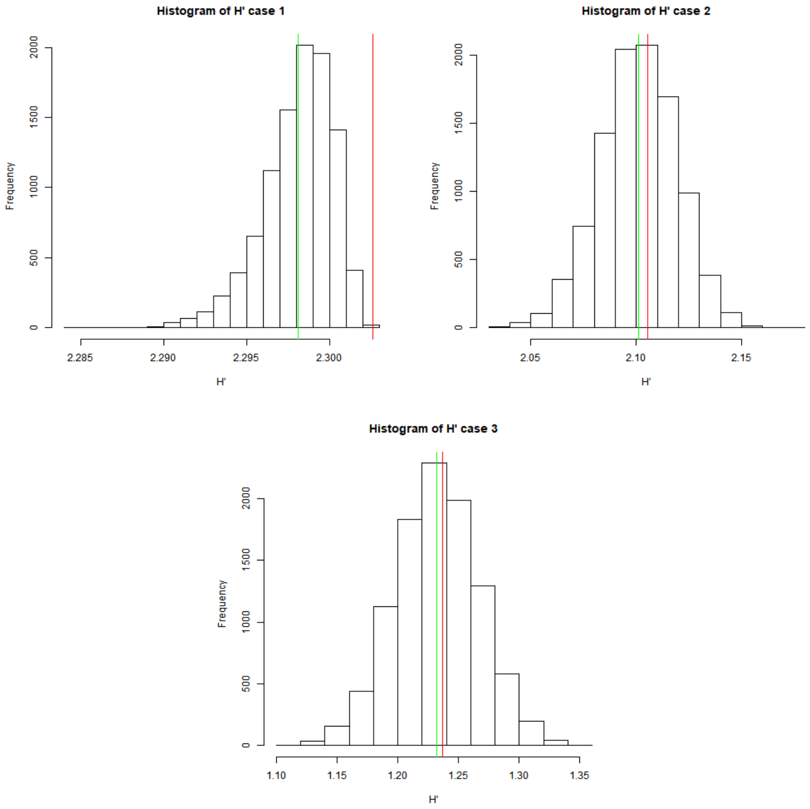


Figure 4.3: Histograms of H' together with the value of $\mathcal{H}$ in red colour and the mean of H' in green colour.

The bias of the Shannon indices mean has been calculated for each case and it is represented in Table 4.2 together with the expected bias obtained from the Expression (4.2).

$$bias = H' - \mathcal{H} \approx -\frac{S-1}{2N} + \frac{1-\sum \theta_i^{-1}}{12N^2} \tag{4.2}$$

Table 4.2: **Shannon indices and bias.**

Case	$\overline{H'}$	$\mathcal{H}$	*bias*	expected bias	1^{st} term	2^{nd} term
1	2.298079	2.302585	$-4.494\ 10^{-3}$	$-4.508\ 10^{-3}$	$-4.500\ 10^{-3}$	$-8.250\ 10^{-6}$
2	2.100841	2.105516	$-4.474\ 10^{-3}$	$-4.513\ 10^{-3}$	$-4.500\ 10^{-3}$	$-1.377\ 10^{-5}$
3	1.231937	1.237139	$-4.932\ 10^{-3}$	$-4.700\ 10^{-3}$	$-4.500\ 10^{-3}$	$-1.999\ 10^{-4}$

The results that we obtain by calculating these biases are two: The first one is that the contribution of the second term is so small that could be omitted for large numbers of N (in or case $N = 1000$, and the second term is irrelevant). The second one is that the estimation of the Shannon index is an accurate estimator, since its bias is of order 10^{-3}.

Chapter 5

Biological models

Biological or **theoretical** models are based on the assumption that an ecological community has a property called niche space that is divided amongst the species that live there. In this section we want to see if there are any explicit relationships between two niche-based species abundance models (Broken stick and Geometric series) and the Shannon index.

5.1 Broken stick model

The broken stick model, sometimes known as the random niche boundary hypothesis, was proposed by MacArthur (MacArthur, 1957). He likened the subdivision of niche space within a community to a stick broken randomly and simultaneously into S species. It is a very uniform distribution and the model may also be viewed as representing a group of S species of equal competitive ability jostling for niche space (Magurran, 2004).

This model is conventionally written in terms of rank order abundance. The number of individuals in the ith species (n_i) is obtained from the term:

$$n_i = \frac{N}{S} \sum_{j=i}^{S} \frac{1}{j}$$

Where, n_i = abundance of the ith species; N = the total number of individuals; and S = the total number of species. The relative abundance of each species is given by the proportion $p_i = n_i/N$, which would be:

$$p_i = \frac{1}{S}\sum_{j=i}^{S}\frac{1}{j}$$

Observation. *Note that the proportions* p_i *only depend on the number of species and we observe there is no parameter to be estimated for the Broken stick model.*

Example 4. *For S=20 we have the following values of* p_i *(approximated):*

p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
0.179	*0.129*	*0.104*	*0.088*	*0.075*	*0.065*	*0.057*	*0.050*	*0.043*	*0.038*

p_{11}	p_{12}	p_{13}	p_{14}	p_{15}	p_{16}	p_{17}	p_{18}	p_{19}	p_{20}
0.033	*0.028*	*0.024*	*0.020*	*0.017*	*0.013*	*0.010*	*0.007*	*0.005*	*0.002*

Now, we suppose $N = 1000$ *and we use the rank/abundance plot (Figure 5.1) to see how is the shape for an assemblage following a broken stick model.*

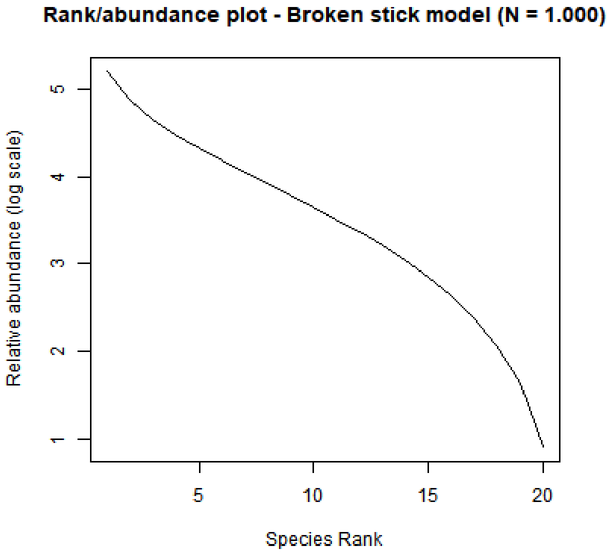


Figure 5.1: Example of a broken stick model rank/abundance plot ($S = 20$ and $N = 1000$).

5.2 Relationship between the broken stick model and the Shannon index

For the broken stick model there are no parameters to be estimated, so given the number of species we can calculate the values of p_i from the formulas. Thus, the estimation of the Shannon index can be directly calculated.

Example 5. *This table shows different values of H'_{bs}, $H'_{max} = ln(S)$ and $J' = H'_{bs}/H'_{max}$ depending on S.*

S	H'_{bs}	H'_{max}	J'
5	*1.34*	*1.61*	*0.83*
10	*1.97*	*2.30*	*0.86*
15	*2.35*	*2.71*	*0.87*
20	*2.62*	*2.99*	*0.87*
25	*2.84*	*3.22*	*0.88*

This gives us an idea of how similar are the resulting indexes while the value of S increases. Note: H'_{bs}= the broken stick model Shannon index.

Explicit relationship between the broken stick model and the Shannon index

We are looking for a constant of proportionality between the theoretical maximum of the Shannon index $H' = ln(S)$ and the broken stick model Shannon index H'_{bs}.

$$\begin{aligned} H'_{bs} &= -\sum_{i=1}^{S} p_i \ln(p_i) \\ &= -\sum_{i=1}^{S} \left(\frac{1}{S} \sum_{j=i}^{S} \frac{1}{j} \right) \ln \left(\frac{1}{S} \sum_{j=i}^{S} \frac{1}{j} \right) \end{aligned}$$

$$If\ we\ note\ q_i := \sum_{j=i}^{S} \frac{1}{j}\ then\ we\ have$$

$$\begin{aligned} H'_{bs} &= -\sum_{i=1}^{S} \left(\frac{1}{S} q_i \right) \ln \left(\frac{1}{S} q_i \right) \\ &= -\sum_{i=1}^{S} \frac{1}{S} \left(q_i \ln(q_i) - q_i \ln(S) \right) \\ &= H'_{max} \frac{-1}{S} \sum_{i=1}^{S} \left(\frac{1}{ln(S)} q_i \ln(q_i) - q_i \right) \end{aligned}$$

Then, we have: $\boxed{H'_{bs} = H'_{max} f(S)}$

where, $f(S) := -\frac{1}{S} \sum_{i=1}^{S} \left(\frac{1}{ln(S)} \sum_{j=i}^{S} \frac{1}{j} \ln(\sum_{j=i}^{S} \frac{1}{j}) - \sum_{j=i}^{S} \frac{1}{j} \right)$.

Hence, the constant of proportionality we were looking for is this $f(S)$.

Visualization of the relationship between the broken stick model and the Shannon index

From the result above, now we want to compare the broken stick model Shannon index and the maximum Shannon index by representing them in the same plot. On Figure 5.2 we observe this relation and we see H'_{bs} is always lower than H'_{max}, which is never reached by the broken stick Shannon index. Actually, this difference seems to be quite stable from around $S = 200$, where $H'_{bs} - H'_{max} \approx 0.4$.

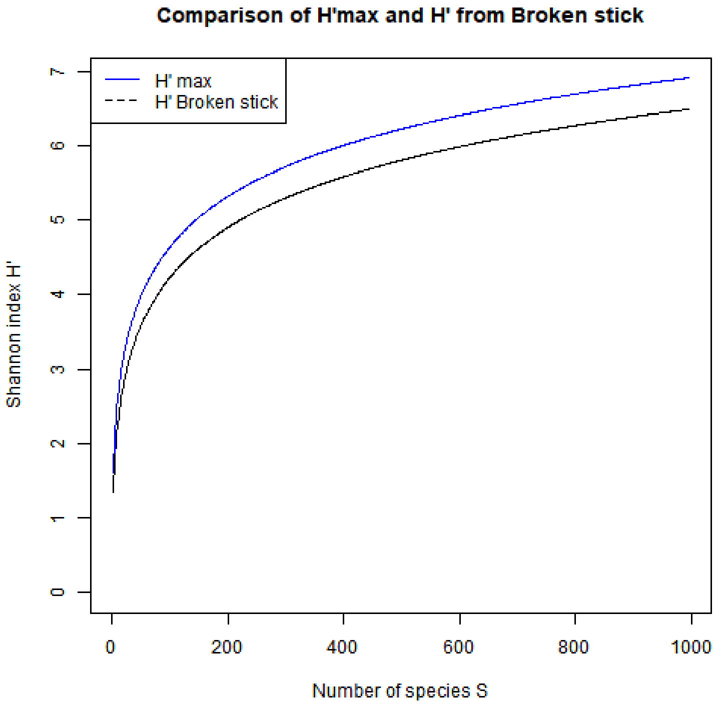


Figure 5.2: Comparing the values of H' from broken stick model proportions and the maximum values $H'_{max} = ln(S)$.

Discussion of the results

We found a clear relationship between the broken stick model and the Shannon index. There is an explicit relationship that can be expressed by the product of the maximum value H'_{max} and a function of S.

In practice, the slow decay in the rank/abundance plot of the broken stick is rarely observed in empirical studies. Therefore, the Shannon index from the broken stick model is always higher than one calculated from empirical data.

This observation leads us to think that it would be more natural to use the broken stick model Shannon index instead of the maximum Shannon index for calculating the Shannon evenness measure, J'.

5.3 Geometric series

Opposite to the broken stick model, the geometric series describes communities of highly uneven species-abundance distribution and low diversity characterized by a few dominant species, thus is a model for poor-species assemblages. The situation is that there is some resource which is limiting and the most dominant species takes fraction k of this resource, then the second most dominant species takes fraction k of the remaining resource, and so on. The geometric series is also called niche preemption model (Magurran, 2004).

The abundance is proportional to the niche a species occupies.

$$ck, ck(1-k), ck(1-k)^2, ..., ck(1-k)^{S-1}, \quad 0 < k < 1$$

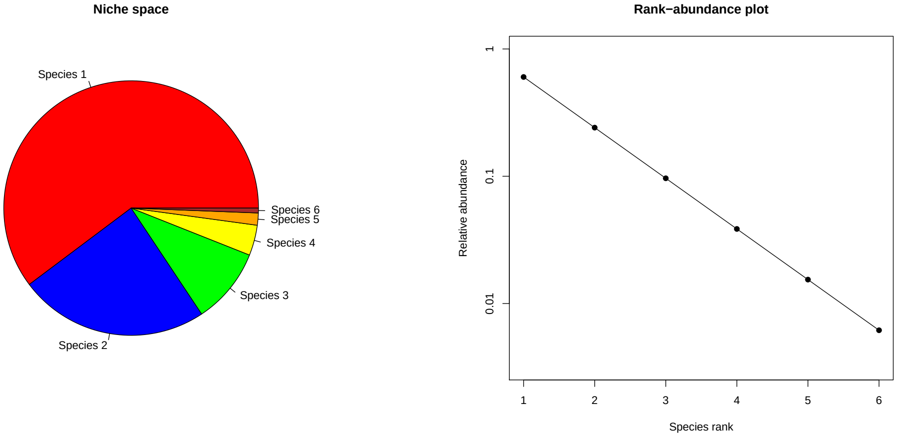


Figure 5.3: On the left Figure we represented the circle as the total resource, where the most dominant species takes a portion k, the next most abundant takes a fraction k of the rest and so on. On the right Figure the rank/abundance plot is presented in a log scale. Graphics are taken from: (Graffelman, 2018).

The number of individuals in the ith species (n_i) is obtained from the term:

$$n_i = NC_k k(1-k)^{i-1}, \; i = 1, \ldots, S$$

Where, N=the total number of individuals, k the proportion of the remaining niche space occupied by each successively colonizing species (k is a constant), and C_k is the constant that insures that $\sum_{i=1}^{S} n_i = N$,

$$C_k = \frac{1}{1-(1-k)^S}$$

As we can observe, the Geometric model has a parameter k to be estimated. Thus, we have to distinguish between the theoretical value of k and the estimated $\hat{k}$, giving the values θ_i for the relative abundance of each species and p_i for the estimation of it. This would be:

$$\theta_i = C_k k(1-k)^{i-1}, \quad p_i = C_{\hat{k}}\hat{k}(1-\hat{k})^{i-1}$$

Note. *Note that* $\ln(n_i)$ *is linear in* i*, with slope* $\ln(1-k)$*, thus,* k *can be estimated by linear regression. But, alternative estimators of* k *have been described in the literature (He and Tang, 2008).*

Because the ratio of the abundance of each species to the abundance of its predecessor is constant through the ranked list of species, the series will appear as a straight line when plotted on a log rank/abundance graph, as we can see in Figure 5.3 and Figure 5.4 (Magurran, 2004).

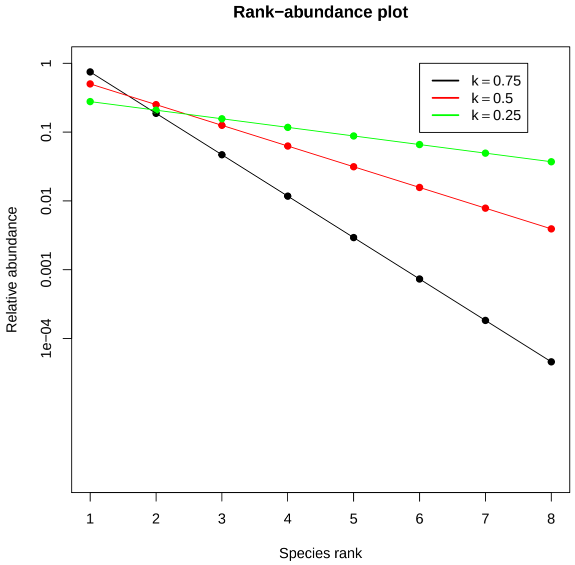


Figure 5.4: Rank/abundance plot in a log scale for different values of k. Graphics from: (Graffelman, 2018)

5.4 Relationship between the geometric series and the Shannon index

The geometric model requires a parameter k to be estimated. Therefore, we will compare the Geometric series with the population Shannon index $\mathcal{H} = -\sum_{i=1}^{S} \theta_i \ln(\theta_i)$, instead of comparing it with the estimation H'.

Explicit relationship between the geometric series and the Shannon index

We are looking for a estimation of the Shannon index given data that follow a geometric series model. Hence we would like to find an explicit relationship between the parameter k and the value $\mathcal{H}$ and S. If we could know the theoretical value of k, then:

$$\begin{aligned}
\mathcal{H} &= -\sum_{i=1}^{S} \theta_i \ln(\theta_i) \\
&= -\sum_{i=1}^{S} Ck(1-k)^{i-1} \ln\left(Ck(1-k)^{i-1}\right) \\
&= -C\sum_{i=1}^{S} k(1-k)^{i-1} \left(\ln(C) + \ln(k) - \ln(1-k) + i\ln(1-k)\right)
\end{aligned}$$

Since we know $\sum_{i=1}^{S} k(1-k)^{i-1} = 1-(1-k)^S = C^{-1}$, then we have:

$$\mathcal{H} = -\ln(C) - \ln(k) + \ln(1-k) - C\ln(1-k)\sum_{i=1}^{S} ik(1-k)^{i-1} \tag{5.1}$$

If we calculate $\sum_{i=1}^{S} ik(1-k)^{i-1} = \frac{-kS(1-k)^S-(1-k)^S+1}{k}$, then:

$$\boxed{\mathcal{H} = -\ln(C) - \ln(k) + \ln(1-k) - C\ln(1-k)\frac{-kS(1-k)^S-(1-k)^S+1}{k}}$$

Note. *We noted $C := C_k$ defined before and we used that $k \in (0,1)$.*

Observation. *From the expression above it is not possible to find k as a closed expression of $\mathcal{H}$ and S.*

Visualization of the relationship between the geometric series and the Shannon index

Here, we want to visualize the relationship between the Shannon index and the k parameter, so in Figure 5.5 we have plotted the Shannon index for different values of S, on the vertical axis, and parameter k, on the horizontal axis.

In order to explore the relationship between the maximum value of the Shannon index H'_{max} and the values of H' for different S and k, we have plotted the Figure 5.6

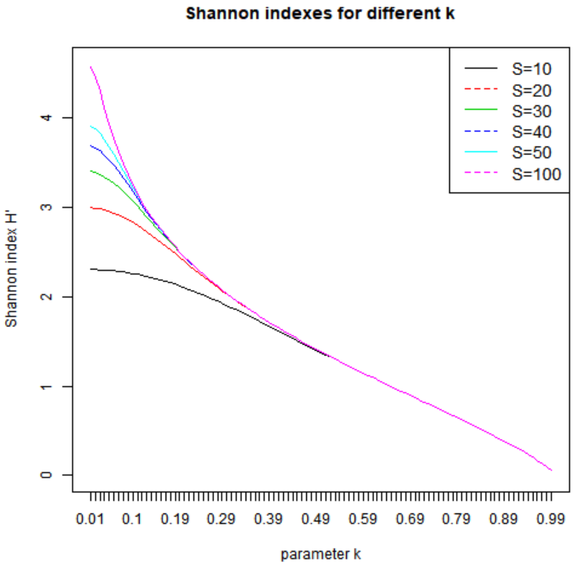


Figure 5.5: Shannon indexes for different values of S.

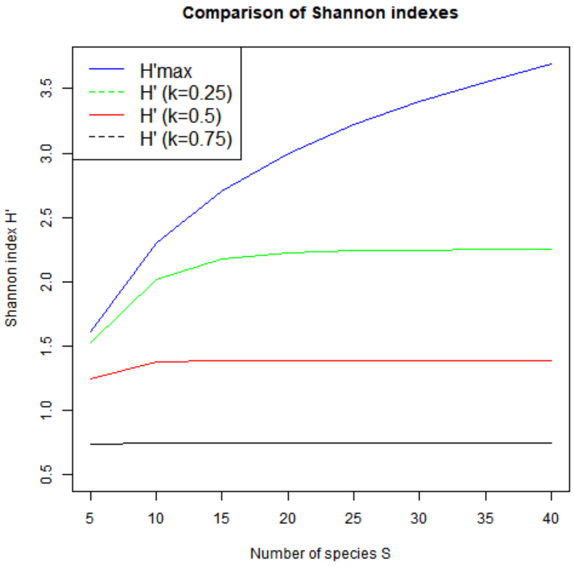


Figure 5.6: Shannon indexes for different values of k together with H'_{max}.

Discussion of the results

We have found an explicit relationship between the Shannon index and the parameter k of the geometric series, but not a closed expression that could allow us to estimate k directly from the Shannon index.

This observation leads us to think that, for empirical data, we can estimate the k parameter from the Shannon index by using numerical methods.Then, we could estimate the Shannon index, given k, or calculate k, given the Shannon index. Therefore, this could lead to a new method for estimating k added to the maximum likelihood or linear regression method, among other methods.

From Figure 5.5 we see an inverse relationship between H' and k that looks almost linear above $k \approx 0.4$. Such relation is monotone, so then we can say k and H' are consistent measures of diversity. If a community is dominated by a few highly abundant species, k will tend to be larger, and correspondingly, the Shannon index will be smaller.

Chapter 6

Application to empirical data

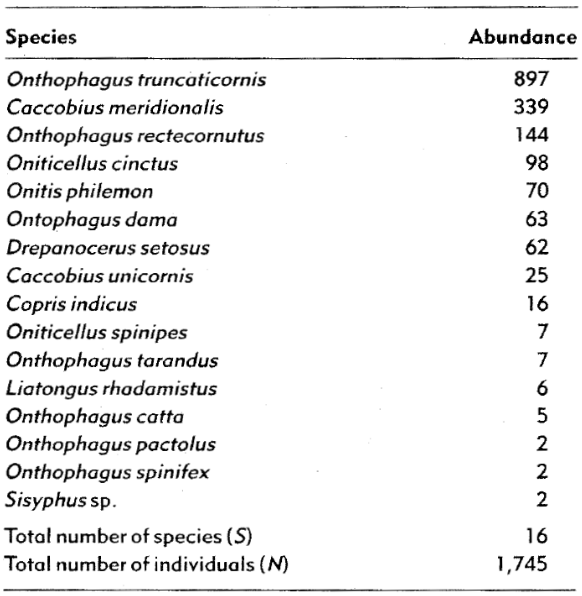
In this section, an example of a species community has been studied focusing on its abundance distribution. The **geometric model** has been fitted to data and the Shannon index calculated. Here we present the procedure we use in order to apply the results of Chapter 4 and Chapter 5.

Procedure

1. Fit the geometric model to data.
2. Estimate the k parameter by linear regression.
3. Estimate the k parameter from the geometric model Shannon index.
4. Compare the two estimators.

Example: Dung beetles

Here we have the abundance distribution of 16 different species in a community of 1745 individuals of dung beetles found around Bangalore in the Western Ghats, India. Data taken from (Magurran, 2004). In Figure 6.1 the relative abundance of each species are represented in a table together with the values of $S = 16$ an $N = 1745$.



Species	Abundance
Onthophagus truncaticornis	897
Caccobius meridionalis	339
Onthophagus rectecornutus	144
Oniticellus cinctus	98
Onitis philemon	70
Ontophagus dama	63
Drepanocerus setosus	62
Caccobius unicornis	25
Copris indicus	16
Oniticellus spinipes	7
Onthophagus tarandus	7
Liatongus rhadamistus	6
Onthophagus catta	5
Onthophagus pactolus	2
Onthophagus spinifex	2
Sisyphus sp.	2
Total number of species (S)	16
Total number of individuals (N)	1,745

Figure 6.1: Beetles abundances.(Magurran, 2004)

First, we calculate the Shannon index and its bias:

- $H' = 1.603149$
- $bias = -\frac{S-1}{2N} = \frac{15}{3490} \approx -4.3\ 10^{-3}$

Then, before fitting the geometric model to data, we visualize the abundance species distribution using the rank/abundance plots in Figure 6.2.

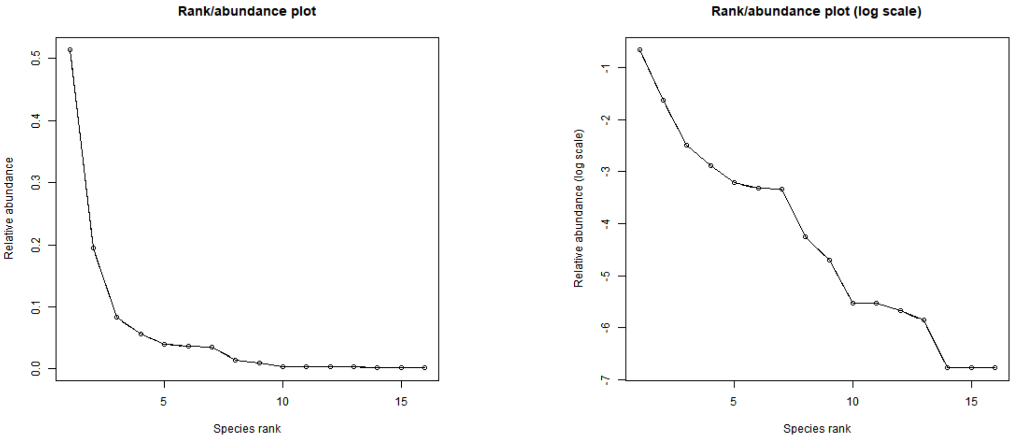


Figure 6.2: Rank/abundance plots of dung beetles.

After this, we represent the log rank/abundance plot together with the linear regression straight line (see Figure 6.3). And then, we estimate the k parameter by using an `R` function called `lm` (R Core Team, 2016).

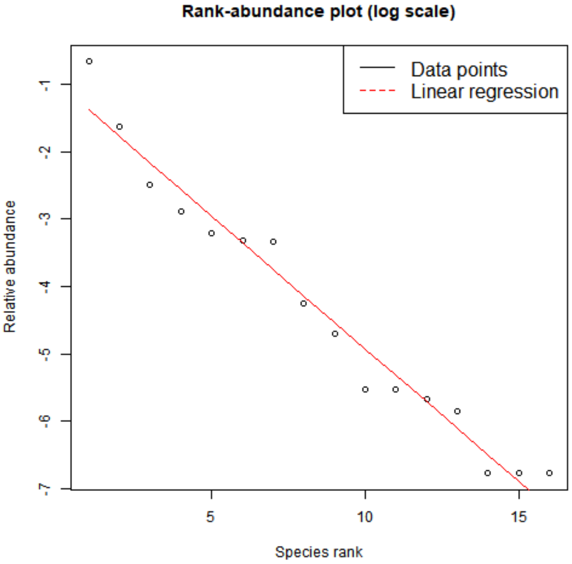


Figure 6.3: Linear regression for dung beetles data.

- The k estimation by linear regression: $\hat{k}_{lr} = 0.3250428$.

Finally, we estimate the k parameter by using the Formula (5.1) and a numerical approximation to the solutions. This was solved by using *Wolfram Language* (??, Mat).

- The k estimation by Formula (5.1): $\hat{k}_{H'} = 0.425896$

In Figure 6.4 we have plotted the geometric Shannon indices depending on $k \in (0, 1)$ and given $S = 16$ together with $H'_{\hat{k}} = 1.92654$ and $H' = 1.603149$.

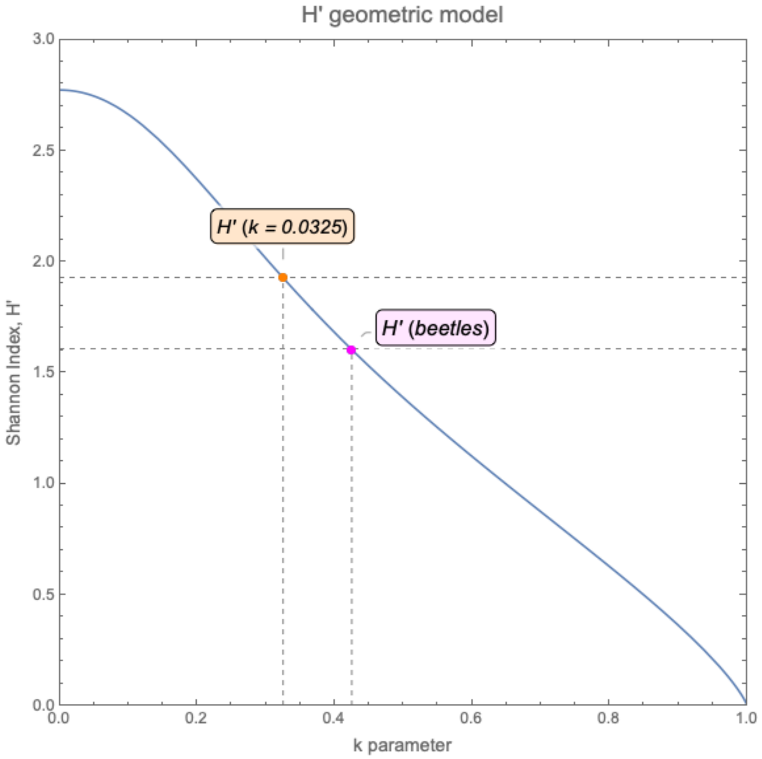


Figure 6.4: The geometric series Shannon indices, for different values of k and given $S = 16$ plotted together with the beetles Shannon index in blue and the geometric Shannon index for $\hat{k} = 0.325$, in red.

Discussion of the results

We have calculated both k by linear regression and by the Expression (5.1) of the Shannon index. The difference between them is $(0.425896 - 0.3250428 = 0.1008532)$ is of order 10^{-1}.

This result, together with the small bias for the Shannon index, lead us to think the linear regression estimation could be worse than the Shannon index estimation.

Hence, we propose the geometric series Shannon index as another method to estimate the k parameter.

Chapter 7

Concluding remarks

Here, we want to emphasize that we successfully achieved a formal definition of the Shannon index. Most studies in the literature do not distinguish between the estimation Shannon index and the population Shannon index.

The simulation study of the estimated Shannon index allowed us to effectively calculate the first moment of the estimation Shannon index. Hence, we obtain the bias and we observe that the second term of the expectation can be omitted when the size of the sample is sufficiently large.

Then, we made the comparison between the Shannon index and the broken stick model. This allowed us to explicitly calculate a Shannon index derived from the broken stick model (H'_{bs}). Furthermore, we suggest that an evenness diversity measure derived from the ratio between the Shannon index (H') and the broken stick model Shannon index (H'_{bs}) would be more realistic than the one widely used in the literature, J'.

We also made the comparison between the Shannon index and a geometric model. This comparison lead us to an explicit relationship between two biodiversity indices; the estimation Shannon index (H') and the k parameter of the geometric model. This result along with the example using empirical data and the small bias of the estimation Shannon index, H', made us realize that a new estimation method based on the estimation Shannon index using numerical methods could be also suitable to estimate the k parameter. In the literature, we found several alternative methods to the one that we described (He and Tang, 2008). So it would be interesting to compare them with our newly defined method.

Bibliography

Arfken, G. B. and H. J. Weber (2005). Mathematical methods for physicists.

Bowman, K., K. Hutcheson, E. Odum, and L. Shenton (1971). *Statistical ecology. Volume 3. Many species populations, ecosystems. Article: Comments on the distribution of indices of diversity. 315-366.* The Penn State Statistics Series, Pennsylvania State.

Fisher, R. A., A. S. Corbet, and C. B. Williams (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, 42–58.

Graffelman, J. (2018). *Species Diversity.* Curs d'Estadistica per les biociencies. UPC.

He, F. and D. Tang (2008). Estimating the niche preemption parameter of the geometric series. *acta oecologica 33*(1), 105–107.

Hutcheson, K. (1970). A test for comparing diversities based on the shannon formula. *Journal of theoretical Biology 29*(1), 151–154.

Krebs, C. (1989). *Ecological Methodology.* Harper & Row.

Lewis, T., L. R. Taylor, et al. (1967). *Introduction to experimental ecology*.

MacArthur, R. H. (1957). On the relative abundance of bird species. pp. 293–295.

Magurran, A. (2004). *Measuring Biological Diversity.* Wiley.

Margalef, R. (1957). La teoria de la informacion en ecologia.

Margalef, R. (1974). *Ecologia.*

Márquez, D. and O. Julià (2011). *Un primer curs d'estadística.* Universitat (Universitat de Barcelona). Publicacions i Edicions de la Universitat de Barcelona.

Peet, R. K. (1974). The measurement of species diversity. *Annual review of ecology and systematics 5*(1), 285–307.

Pielou, E. (1969). *An introduction to mathematical ecology.* Wiley-Interscience.

Preston, F. W. (1948). The commonness, and rarity, of species. *Ecology 29*(3), 254–283.

R Core Team (2016). R: a language and environment for statistical computing.

Sanz i Solé, M. (1999). *Probabilitats*, Volume 28. Edicions Universitat Barcelona.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal 27*(3), 379–423.

Simpson, E. H. (1949). Measurement of diversity. *nature 163*(4148), 688–688.

Tokeshi, M. (1993). Species abundance patterns and community structure. In *Advances in ecological research*, Volume 24, pp. 111–186. Elsevier.

United Nations Environment Program (2019). Biodiversity. https://www.biodiversitya-z.org/content/biodiversity.

Appendix A

Calculation of $E(H')$

We perform the Taylor expansion for $f(x)$ at $x = a$:

$$f(x) \approx f(a) + \frac{f'(a)(x-a)}{1!} + \frac{f''(a)(x-a)^2}{2!} + \frac{f''(a)(x-a)^3}{3!} + ... = \sum_{n=0}^{\infty} \frac{f^n(a)(x-a)}{n!}$$

We use p_i to estimate θ_i where $p_i = \frac{k_i}{N}$.
We consider the transformation $p_i \ln p_i$ (or $\theta_i \ln \theta_i$)

$$f(\theta) = \theta \ln \theta$$

$$f'(\theta) = \ln \theta + \theta \frac{1}{\theta} = \ln \theta + 1$$

$$f''(\theta) = \frac{1}{\theta}$$

$$f'''(\theta) = -\frac{1}{\theta^2}$$

Taylor expansion for $f(p_i)$ around $p_i = \theta_i$

$$p_i \ln p_i \approx \theta_i = \ln \theta_i + (\ln \theta_i + 1)(p_i - \theta_i) + \frac{1}{2\theta}(p_i - \theta_i)^2 - \frac{1}{6\theta^2}(p_i - \theta_i)^3 + ...$$

We obtain $E(H')$ by taking sums and expectations

$$E(H') = E(\sum_{i=1}^{S} p_i \ln p_i) = \sum_{i=1}^{S} E(p_i \ln p_i)$$

First term:

$$-\sum_{i=1}^{S} E(\theta_i \ln \theta_i) = -\sum \theta_i \ln \theta_i = \mathcal{H}$$

Second term:

$$E((\ln \theta_i + 1)(p_i - \theta_i)) = (\ln \theta_i + 1)E(p_i - \theta_i) = (\ln \theta_i + 1)(\theta_i - \theta_i) = 0$$

Third term:

$$E(\frac{1}{2\theta_i})(p_i - \theta_i)^2) = \frac{1}{2\theta_i}E(p_i - \theta_i)^2 = \frac{1}{2\theta_i}E(p_i - E(p_i))^2 = \frac{1}{2\theta_i}V(p_i)$$

Since $p_i = \frac{n_i}{N}$ and $n_i \sim Bin(N, \theta_i)$ we have:

$$E(n_i) = N\theta_i$$
$$V(n_i) = N\theta_i(1 - \theta_i)$$
$$V(p_i) = \frac{1}{N^2}V(n_i) = \frac{\theta_i(1 - \theta_i}{N}$$

Third term summand and $x - 1$:

$$-\sum_{i=1}^{S} \frac{(1 - \theta_i)}{2N} = \frac{-1}{2N}(S - 1) = -\frac{S - 1}{2N}$$

Coincides with Peet p. 292. (Peet, 1974)

Appendix B

R code

In this Appendix we present R code used in Chapter 4, Chapter 5 and Chapter 6 that we consider important to be commented.

Chapter 4

Simulation study

```
# Simulation study - Monte Carlo - using rmultinom
# ==================================================

#Shannon function
shannon <- function(x) {
  S<-length(x)
  N<-sum(x)
  p<-x/N
  p<-p[p>0]
  return(-sum(p*log(p, base = exp(1))))
}

S=10
N<-1000
nsimul<-10000

# case 1
# ======
theta1<-rep(0.1,10)
H1<-shannon(theta1)
# simulation
set.seed(123)
x <- rmultinom(nsimul,N,prob=theta1)
X<-matrix(data=x, nrow = S, ncol = nsimul)
# calculating H'
H1.prima<-numeric(nsimul)
for (i in 1:nsimul) {
  H1.prima[i]<-shannon(X[,i])
}
#mean of H'
m1<-mean(H1.prima)
bias1<-m1-H1


# case 2
# ======
theta2 <- c(0.25,0.15,0.13,0.11,0.10,0.09,0.07,0.05,0.03,0.02)
H2<-shannon(theta2)
# simulation
set.seed(123)
x <- rmultinom(nsimul,N,prob=theta2)
X<-matrix(data=x, nrow = S, ncol = nsimul)
# calculating H'
H2.prima<-numeric(nsimul)
for (i in 1:nsimul) {
  H2.prima[i]<-shannon(X[,i])
}
#mean of H'
m2<-mean(H2.prima)
bias2<-m2-H2


# case 3
# ======
theta3<-c(0.6, 0.2, 0.1, 0.04, 0.03, 0.01, 0.01, 0.008, 0.001, 0.001)
H3<-shannon(theta3)
# simulation
set.seed(123)
```

```
x <- rmultinom(nsimul,N,prob=theta3)
X<-matrix(data=x, nrow = S, ncol = nsimul)
# calculating H'
H3.prima<-numeric(nsimul)
for (i in 1:nsimul) {
  H3.prima[i]<-shannon(X[,i])
}
#mean of H'
m3<-mean(H3.prima)
bias3<-m3-H3
```

Observation. *In order to calculate the Shannon index, if we don't choose the implemented function in the vegan package of R, we note there is a problem defining the function for species with 0 individuals. Then, we have to consider only the positive values of the data set.*

Chapter 5

The code for the Figure 5.2 and the Figure 5.5 is presented here.

Broken stick model, Figure 5.2

```
# Comparing the broken stick model with H'
#=========================================

shannon <- function(x) {
  x<-x[x>0]
  return(-sum(x*log(x, base = exp(1))))
}

sp<-1000
p <- matrix( rep(0), nrow = sp, ncol = sp)
H.prima<-rep(0,sp)
H.max<-rep(0,sp)
for (S in 5:sp) {
  for (i in 1:S) {
    p[S,i]<-0
    for (j in i:S) {
      p[S,i]<-p[S,i]+1/j
    }
    p[S,i]<-p[S,i]*1/S
  }
  x<-p[S,]
  H.prima[S] <- shannon(x)
  H.max[S] <- log(S)
}
H.prima<-H.prima[H.prima>0]
H.max<-H.max[H.max>0]
```

Geometric model, Figure 5.5

```
shannon <- function(x) {
  S<-length(x)
  N<-sum(x)
  p<-x/N
  p<-p[p>0]
  return(-sum(p*log(p, base = exp(1))))
}

# Geometric model
# ===============
N<-1000
k<-seq(0.01,0.99,by=0.01)

# 1. S = 10
# ---------
S<-10
H1<-numeric(length(k))
n<-numeric(S)
for (j in 1:length(k)) {
  C<-1/(1-(1-k[j])^S)
  for (i in 1:S) {
    n[i]<-(N*C*k[j]*(1-k[j])^(i-1))
  }
  H1[j]<-shannon(n)
}

# 2. S = 20
# ---------
S<-20
H2<-numeric(length(k))
n<-numeric(S)
for (j in 1:length(k)) {
  C<-1/(1-(1-k[j])^S)
  for (i in 1:S) {
    n[i]<-(N*C*k[j]*(1-k[j])^(i-1))
  }
  H2[j]<-shannon(n)
}
```

```
# 3. S = 30
# ---------
S<-30
H3<-numeric(length(k))
n<-numeric(S)
for (j in 1:length(k)) {
  C<-1/(1-(1-k[j])^S)
  for (i in 1:S) {
    n[i]<-(N*C*k[j]*(1-k[j])^(i-1))
  }
  H3[j]<-shannon(n)
}

# 4. S = 40
# ---------
S<-40
H4<-numeric(length(k))
n<-numeric(S)
for (j in 1:length(k)) {
  C<-1/(1-(1-k[j])^S)
  for (i in 1:S) {
    n[i]<-(N*C*k[j]*(1-k[j])^(i-1))
  }
  H4[j]<-shannon(n)
}
# 5. S = 50
# ---------
S<-50
H5<-numeric(length(k))
n<-numeric(S)
for (j in 1:length(k)) {
  C<-1/(1-(1-k[j])^S)
  for (i in 1:S) {
    n[i]<-(N*C*k[j]*(1-k[j])^(i-1))
  }
  H5[j]<-shannon(n)
}

# 6. S = 100
# ----------
S<-100
H6<-numeric(length(k))
n<-numeric(S)
for (j in 1:length(k)) {
  C<-1/(1-(1-k[j])^S)
  for (i in 1:S) {
    n[i]<-(N*C*k[j]*(1-k[j])^(i-1))
  }
  H6[j]<-shannon(n)
}
```

Chapter 6

Using `lm` for the linear regression

```r
# DUNG BEETLES
# ============
X<-read.csv("6maguranBeetles.csv", sep = ";")
N<-sum(X$Abundance)
S<-length(X$Species)
species.rank<-1:S
# p = relative abundance
p<-X$Abundance/N
log.p<-log(p)

# REGRESSION RESULTS
# ==================
x<-species.rank
lny<-log.p
out.lny<-lm(formula = lny ~ x)
summary(out.lny)
coefficients(out.lny)
b<- out.lny$coefficients

#estimated line
y<- b[1] + b[2]*x

#estimation of k (linear regression)
#----------------------------------
# ln(1-k)= b[2] then k=1-exp(b[2])
k<-1-exp(b[2]) # k = 0.3250428


# Shannon index Dung Beetles
# ==========================
shannon <- function(x) {
  S<-length(x)
  N<-sum(x)
  p<-x/N
  p<-p[p>0]
  return(-sum(p*log(p, base = exp(1))))
}
H<-shannon(X$Abundance)

# comparing with Shannon index
# =============================

# fixed S = 16
# ------------
n1<-numeric(S)
k1<-seq(0.01,0.99,by=0.01)
H1<-numeric(length(k1))
H1.max<-log(S)
for (j in 1:length(k1)) {
  C<-1/(1-(1-k1[j])^S)
  for (i in 1:S) {
    n1[i]<-(N*C*k1[j]*(1-k1[j])^(i-1))
  }
  H1[j]<-shannon(n1)
}

# shannon index geometric model (S=16 and k.hat=0.3250428)
# ========================================================
k<-0.3250428
n<-numeric(S)
C<-1/(1-(1-k)^S)
for (i in 1:S) {
  n[i]<-(N*C*k*(1-k)^(i-1))
}
H2<-shannon(n)


bias<-H2-H
```