THE MEASUREMENT OF BIODIVERSITY

Eric Marcon

UMR Écologie des forêts de Guyane
`http://www.ecofog.gf`

Contents

Part I

Notions and tools

CHAPTER 1

Notions

THE TERM *biodiversity* is attributed[1] to Walter Rosen, a member of the National Research Council of America, who created a portmanteau of the terms *biological diversity* while preparing for a conference, the proceedings of which would be published as "Biodiversity".[2] The question of biological diversity has interested ecologists long before the invention of the term biodiversity, but the neologism, although widely adopted,[3] has become at the same time a fuzzy concept, in which one can find whatever one is looking for, to the point that the term has lost its scientific character.[4] One cause of this shift, linked with conservation biology, is that biodiversity has been named to draw attention to its erosion. As this erosion potentially affects many aspects of the living world, the definition of biodiversity fluctuates according to need: DeLong[5] lists 85 in the first ten years of literature. Indicators of biodiversity can encompass many other things besides the diversity of living things: the number of threatened species (eg the IUCN Red List), the size of populations or the extent of protected ecosystems, the degradation of habitats, the threat to emblematic species ... However, it is possible to construct a rigorous and coherent measure of diversity which can clarify many (but not all) of the concepts which biodiversity encompasses.

In the introduction to the first chapter of the Proceedings of what had become the "Forum on Biodiversity", Wilson uses the word in the narrow sense of numbers of species. The broadening of this sense both towards "natural systems" and towards intraspecific genetic diversity has come from the world of conservation.[6] Michel Loreau, president of the scientific comity of the 2005 Paris conference[7] gave a more complete definition:

> "The Earth is home to an extraordinary biological diversity, which includes not only the species that inhabit our planet, but also the diversity of their genes, the multitude of ecological interactions between them and their physical environment, and the variety of complex ecosystems that they constitute. This biodi-

[1] C. Meine et al. (2006). ""A mission-driven discipline": The growth of conservation biology". In: *Conservation Biology* 20.3, pp. 631–651. DOI: 10.1111/j.1523-1739.2006.00449.x.

[2] E. O. Wilson and F. M. Peter, eds. (1988). *Biodiversity*. Washington, D.C.: The National Academies Press.

[3] P. Blandin (2014). "La diversité du vivant avant (et après) la biodiversité : repères historiques et épistémologiques". In: *La biodiversité en question. Enjeux philosophiques, éthiques et scientifiques*. Ed. by E. Casetta and J. Delord. Paris: Editions Matériologiques. Chap. 1, pp. 31–68.

[4] J. Delord (2014). "La biodiversité : imposture scientifique ou ruse épistémologique ?" In: *La biodiversité en question. Enjeux philosophiques, éthiques et scientifiques*. Ed. by E. Casetta and J. Delord. Paris: Editions Matériologiques. Chap. 3, pp. 83–118. DOI: 10.3917/edmat.delor.2014.01.0083.

[5] D. C. J. DeLong (1996). "Defining Biodiversity". In: *Wildlife Society Bulletin* 24.4, pp. 738–749.

[6] J. G. Speth et al. (1992). "Foreword". In: *Global Biodiversity Strategy*. Ed. by K. Courrier. Washington, D.C.: WRI, IUCN, UNEP, pp. v–vi.

[7] M. Loreau (2005). "Discours de clôture". In: *Actes de la Conférence internationale Biodiversité Science et Gouvernance*. Ed. by R. Barbault and J.-P. Le Duc. Paris, France.: IRD Editions, pp. 254–256.

versity, which is the product of more than 3 billion years of evolution, is a natural heritage and a vital resource on which humanity depends in many ways."

Even today, the term *biodiversity* most often refers to the species richness of an ecosystem. For the sake of clarity, here biodiversity is examined at the level of species (another fuzzy concept[8]). However, accounting for the totality of living beings is almost impossible. The measure of diversity is therefore limited to a taxocene, that is to say a subset of taxonomically related species of a community: butterflies, mammals, trees (the delineation of the subset is not necessarily strictly taxonomic) ...

Since the "Forum on Biodiversity", tropical forests, because of their diversity and importance to conservation, have been an intense focus of biodiversity studies. Most of the examples here will be from studies of tropical forest trees, because they have the advantage of being clearly defined as individuals (and hence easy to count), and because they provide the opportunity to consider, from real data, various methodological problems for the estimation of their diversity.

However, it should always be borne in mind that investigating specific diversity is only one approach. We could of course examine other levels and objects, such as genetic diversity (in terms of different alleles for certain genes or markers) within a population, or even the diversity of interactions between species of a community.[9] The methods presented here apply to the measurement of diversity in general, and not even necessarily biological.

The purpose of this book is to address the measurement of diversity, and not its importance as such. We refer you, for example, to Chapin et al.[10] for a review of this subject, to Cardinale et al.[11] for the consequences of biodiversity loss on ecosystem services, Ceballos et al.[12] for the autocatalytic properties of biodiversity.

The measurement of diversity is an important topic as such,[13] to help formalize concepts and apply them to reality. The question is far from exhausted, and is still the subject of active research and controversy autoconte Ricotta2005b.

1.1 Components

Intuitively, a community a community comprised of many species, seems, but with one dominant species, seems less diverse than a community with fewer species whose numbers are similar (Figure 1.1, left column) Therefore it is necessary to take into account both of the two components of diversity: richness and evenness.[14]

1.1.1 Richness

Richness (a term first used by Mcintosh[15]) is the number (or

[8] J. Hey (2001). "The mind of the species problem". In: *Trends in Ecology & Evolution* 16.7, pp. 326–329. DOI: 10.1016/S0169-5347(01)02145-0.

[9] Z. Jizhong et al. (1991). "An index of ecosystem diversity". In: *Ecological Modelling* 59, pp. 151–163. DOI: 10.1016/0304-3800(91)90176-2.

[10] F. S. I. Chapin et al. (2000). "Consequences of changing biodiversity". In: *Nature* 405.6783, pp. 234–242. DOI: 10.1038/35012241.

[11] B. J. Cardinale et al. (2012). "Biodiversity loss and its impact on humanity". In: *Nature* 486.7401, pp. 59–67. DOI: 10.1038/nature11148.

[12] G. Ceballos et al. (2017). "Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines". In: *Proceedings of the National Academy of Sciences*, p. 201704949. DOI: 10.1073/pnas.1704949114.

[13] A. Purvis and A. Hector (2000). "Getting the measure of biodiversity." In: *Nature* 405.6783, pp. 212–9. DOI: 10.1038/35012221.

[14] R. H. Whittaker (1965). "Dominance and diversity in land plant communities". In: *Science* 147.3655, pp. 250–260.
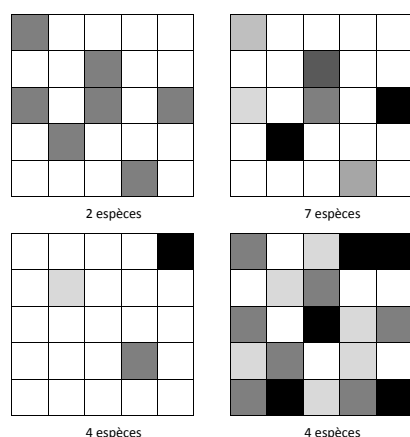
Figure 1.1 – The importance of richness (top) and of evenness (bottom) for defining diversity. Top line: all things being equal, a community of 7 species appears more diverse than a community of 2 species. Bottom line: if richness is equal, a community with lower evenness appears to have lower diversity. Left Column: a community with lower richness (top), can appear more diverse if it has greater evenness. Right column: *idem* for the bottom community.

increasing function of the number) of different classes present in the studied system, for example the number of tree species in a forest.

A number of more or less explicit assumptions are made:

- The classes are well known: counting the number of species makes little sense if the taxonomy is not well established. This is sometimes a major difficulty when working on microorganisms;
- The classes are equidistant: the richness increases by one unit when adding a species, whether this species is close to the previous ones or extremely original.

The simplest and most used index of richness is, quite simply, the number of species S.

1.1.2 Evenness

The regularity of species distribution, called *evenness* or *equitability*, is an important element of diversity. The contribution of a species to an ecosystem in which it is represented abundantly is not the same as the contribution of a species represented by a single individual. In the figure 1.1, the bottom line shows two communities both composed of 4 species, but the one on the right has much greater evenness than the one on the left. Intuitively, the community on the right seems more diverse. With an equal number of species, the presence of very dominant species causes, mathematically, the scarcity of some others. Therefore it is intuited that diversity will be greatest when species have a more regular distribution.

Indices of evenness are independent from numbers of species (thus of richness).

Most common indices, like those of Simpson or Shannon, evaluate both richness and evenness.

1.1.3 Disparity

Traditional measures of diversity, called *species-neutral diversity* or *taxonomic of diversity*, take into account the distance between classes at all. Yet two species of the same genus are obviously closer than two species of different families. Measures of phylogenetic diversity and of functional diversity take this notion into account. This requires some additional definitions.[16]

Although often the difference between two classes is measured as a distance, sometimes not all properties of a distance are met. In such cases, the difference is called a dissimilarity. *Divergence*[17] is the measurement of the dissimilarity between two classes, with or without frequency weighting.

Disparity[18] is the average divergence between any two species (regardless of their frequency) or, equivalently, the total length of the branches of a phylogenetic tree. It is the component which describes how different any one species is from another.

Regularity describes the occupation of niches by a species (functional regularity), or the regularity, over time and between clades, of speciation events represented by a phylogenetic tree. This concept extends that of evenness: diversity increases with richness, with divergence between secies, and with regularity (which is reduced to evenness when all species are equally divergent from one another).

1.1.4 Aggregation

From a broad review of the literature in a number of scientific disciplines interested in diversity (beyond biodiversity), Stirling[19] believes that the three components, which he calls *variety* (richness), *equilibrium* (evenness) and *disparity*, cover all aspects of diversity.

Stirling defines *aggregation* as a measurement of diversity with the capacity to explicitly combine these three components. This is not to say that these components contribute to diversity independently from one another.[20]

1.2 Levels of study

Traditionally, diversity is measured at several nested levels: α, β and γ. This system was first used by Whittaker,[21] who used the Fisher Alpha index (see page **??**)to measure local diversity, and thus named it α diversity, and used the subsequent letters from the Greek alphabet according to his needs.

1.2.1 α, β, and γ diversity

α diversity is local diversity, measured within a delimited system. More precisely, it concerns diversity in a uniform habitat of fixed

[16]D. Mouillot et al. (2005). "Niche overlap estimates based on quantitative functional traits: a new family of non-parametric indices". In: *Oecologia* 145.3, pp. 345–353. DOI: 10.1007/s00442-005-0151-z; C. Ricotta (2007). "A semantic taxonomy for diversity measures". In: *Acta Biotheoretica* 55.1, pp. 23–33. DOI: 10.1007/s10441-007-9008-7.

[17]S. Pavoine and M. B. Bonsall (2011). "Measuring biodiversity to explain community assembly: a unified approach". In: *Biological Reviews* 86.4, pp. 792–812. DOI: 10.1111/j.1469-185X.2010.00171.x.

[18]B. Runnegar (1987). "Rates and Modes of Evolution in the Mollusca". In: *Rates of Evolution*. Ed. by M. Campbell and M. F. Day. London: Allen & Unwin, pp. 39–60.

[19]A. Stirling (2007). "A general framework for analysing diversity in science, technology and society". In: *Journal of the Royal Society, Interface* 4.15, pp. 707–719. DOI: 10.1098/rsif.2007.0213.

[20]L. Jost (2010). "The Relation between Evenness and Diversity". In: *Diversity* 2.2, pp. 207–232. DOI: 10.3390/d2020207.

[21]R. H. Whittaker (1960). "Vegetation of the Siskiyou Mountains, Oregon and California". In: *Ecological Monographs* 30.3, pp. 279–338. DOI: 10.2307/1943563, page 320.
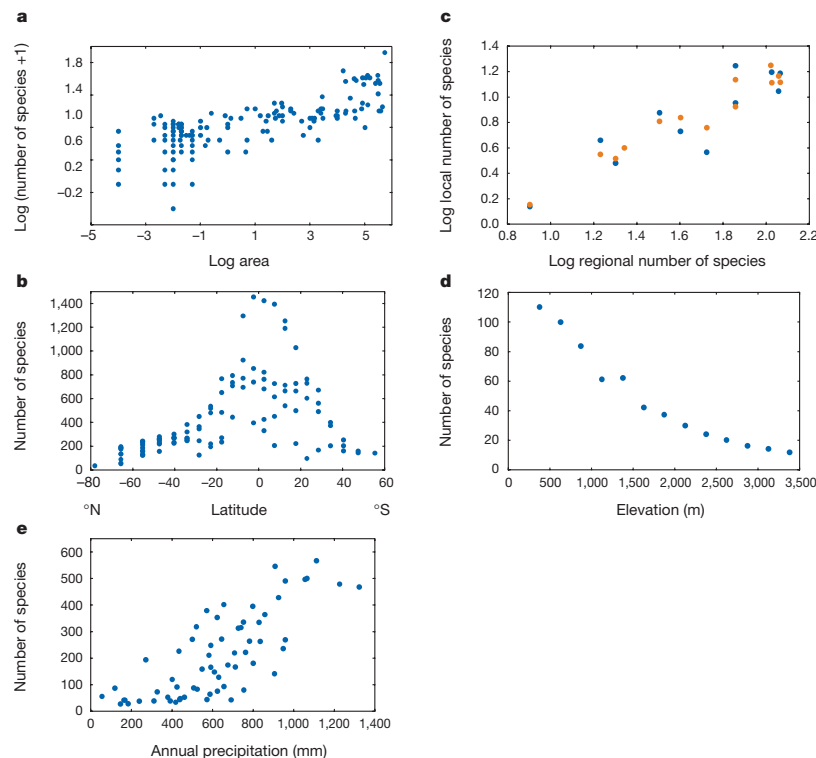
size.

Figure 1.2 – Patterns of biodiversity.[a] (a) The number of species of earthworms increases as a function of the area sampled, from $100\,m^2$ to more than $500\,000\,km^2$ according to the Arrhenius relationship (see page **??**). (b) The number of species of birds as a function of latitude . (c) Relationship between regional richness and local richness. (d) The number of species of bats as a fuction of altitude in a Peruvian nature reserve. (e) The number of ligneous plant species as a function of rainfall in South Africa

[a]K. J. Gaston (2000). "Global patterns in biodiversity". In: *Nature* 405.6783, pp. 220–227. DOI: 10.1038/35012228, figure 1.

Generally speaking,[22] species richness decreases with latitude (diversity is greater in the tropics, and within the tropics increases according to proximity to the equator), see figure 1.2. This pattern is the same for intraspecific genetic diversity.[23] Richness decreases with altitude. Richness is generally lower on islands, and decreases as distance from continental land masses (the source of migrations) increases.

[22]K. J. Gaston (2000). "Global patterns in biodiversity". In: *Nature* 405.6783, pp. 220–227. DOI: 10.1038/35012228.

[23]A. Miraldo et al. (2016). "An Anthropocene map of genetic diversity". In: *Science* 353.6307, pp. 1532–1535. DOI: 10.1126/science.aaf4381.

β diversity measures the differences between local systems. This rather vague definition remains the subject of much debate.[24]

Finally, γ diversity is similar to α diversity, but takes into account the entirety of the system being studied. α and γ diversities are measured in the same way.

[24]C. E. Moreno and P. Rodríguez (2010). "A consistent terminology for quantifying species diversity?" In: *Oecologia* 163.2, pp. 279–82. DOI: 10.1007/s00442-010-1591-7.

1.2.2 Decomposition

Whittaker[25] unsuccessfully proposed a standardization of the scales at which biodiversity is evaluated, suggesting ε for regional diversity (with γ diversity for landscapes and α for habitats) and δ diversity for intra-landscape evaluations. However, only the three original levels have been preserved in the literature, and their scales of observation are not strictly defined.

[25]R. H. Whittaker (1977). "Evolution of species diversity in land communities". In: *Evolutionary Biology* 10. Ed. by M. K. Hecht et al., pp. 1–67.

The distinction between α and β diversities depends upon how strictly habitats are defined. Identifying many habitats decreases α diversity, to the benefit of β diversity. It is therefore important to use a measurement of diversity which does not depend upon

this division, but which uses a cumulative measure (additive or multiplicative) which describes the total diversity, and which can be decomposed into the appropriately weighted sums or products of α diversities of habitats (intra-diversity), and inter-habitat β diversity.

We will call the level at which we define α diversity, *community*. We will call the level at which we will group communities in order to estimate γ diversity, *meta-community*.

1.3 The problem of species

Evaluating species richness supposes that species are clearly defined, which by all evidence is not the case.[26] The first aspect of the problem of the concept of species concerns its nature: is a species a natural reality, or merely an overly-simplistic representation? Richards[27] provides a historic and philosophical analysis of this question. The second aspect of the problem of the concept of species, with immediate and practical consequences, concerns their delimitation. Mayden[28] lists 22 different definitions of the concept of species.

The most common concept is that of the *biological* concept of species[29] : a "group of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups.".[30] As long as the populations of a species are isolated from each other geographically, their ability to reproduce remains theoretical (and moreover is rarely verified experimentally). Allopatric populations have no real gene flow between them and can be considered as distinct species according to the *phylogenetic* concept of species: "the smallest diagnosable cluster of individual organisms within which there is a parental pattern of ancestry and descent".[31] It is the genetic unit detected by the coalescent method for species delineation.[32] The number of phylogenetic species is much higher than the number of biological species. Finally, Van Valen[33] defines species by the ecological niche they occupy (from the example of European white oaks) rather than by gene flow (permanent between distinct species): the *ecological* concept of species is close to that of species complex (a set of neighboring species exchanging genes[34]).

How one chooses to define species modifies considerably the quantification of richness.[35] This creates not merely conceptual problems, but methodological problems:[36] the separation or grouping together of several populations or morphotypes into a greater or lesser number of species is a choice that reflects the knowledge of the moment, and can evolve.[37]

The problem of defining species has an impact upon the measurement of diversity which for the moment has no solution, other than using the same definitions when comparing different commu-

[26] E. Casetta (2014). "Évaluer et conserver la biodiversité face au problème des espèces". In: *La biodiversité en question. Enjeux philosophiques, éthiques et scientifiques*. Ed. by E. Casetta and J. Delord. Paris: Editions Matériologiques. Chap. 5, pp. 139–154.

[27] R. A. Richards (2010). *The Species Problem. A Philosophical Analysis*. Cambridge: Cambridge University Press.

[28] R. L. Mayden (1997). "A hierarchy of species concepts: the denouement in the saga of the species problem". In: *Species. The units of biodiversity*. Ed. by M. F. Claridge et al. London: Chapman and Hall. Chap. 19, pp. 381–424.

[29] T. Dobzhansky (1937). *Genetics and the Origin of Species*. New York: Columbia University Press.

[30] E. Mayr (1942). *Systematics and the origin of species from the viewpoint of a zoologist*. New York: Columbia University Press.

[31] J. Cracraft (1983). "Species Concepts and Speciation Analysis". In: *Current Ornithology Volume 1*. Ed. by R. F. Johnston. Vol. 1. Current Ornithology. Springer US, pp. 159–187. DOI: 10.1007/978-1-4615-6781-3_6.

[32] J. Sukumaran and L. L. Knowles (2017). "Multispecies coalescent delimits structure, not species". In: *Proceedings of the National Academy of Sciences of the United States of America* in press. DOI: 10.1073/PNAS.1607921114.

[33] L. Van Valen (1976). "Ecological Species, Multispecies, and Oaks". In: *Taxon* 25.2/3, pp. 233–239. DOI: 10.2307/1219444.

[34] J. Pernès, ed. (1984). *Gestion des ressources génétiques des plantes. Tome 2 : Manuel*. Paris: Agence de Coopération culturelle et technique.

[35] P. M. Agapow et al. (2004). "The impact of species concept on biodiversity studies". In: *The Quarterly Review of Biology* 79.2, pp. 161–179.

[36] Hey (2001). "The mind of the species problem", see n. 8, p. 4.

[37] A. Barberousse and S. Samadi (2014). "La taxonomie et les collections d'histoire naturelle à l'heure de la sixième extinction". In: *La biodiversité en question. Enjeux philosophiques, éthiques et scientifiques*. Ed. by E. Casetta and J. Delord. Paris: Editions Matériologiques. Chap. 6, pp. 155–182.

nities. The phylogenetic approach (page **??**) may get around the
problem: if two very similar taxa provide barely more diversity
than a single taxon, it is not critical to distinguish them.

Tools

> **Key Points**
>
> Diversity can be described locally by an accumulation curve (SAC) that represents the number of species sampled as a function of effort. At a larger scale, this curve is called the species-area relationship (SAR).
>
> Species abundance distribution (SAD) is represented by a frequency histogram or a rank-abundance diagram.
>
> The coverage rate is the sum of the probabilities of the species observed given the sampling effort. It can be estimated accurately from inventory data. Completeness is the proportion (in number) of species observed.

Some tools are necessary before getting to the heart of the subject. Relationships describing the number of species based on sample size (species-area relationships) and species abundance distribution are important for ecologists. Estimating diversity from real data is based on a measurement of how exhaustive sampling is by the rate of coverage.

2.1 Calculations and data

The measures of diversity here presented require an intense usage of mathematical formalism. A glossary is provided in the preface, and will be referred to here where necessary.

Calculations were made in R,[1] mostly using the package *entropart*.[2] All the code used can be downloaded [3] in addition to updates of this document.

Much of the data used here come from the permanent forestry plot of Barro Colorado Island (BCI):[4] 50 ha of tropical forest for which there is an inventory of all trees of more than 1 cm diameter at breast height (DBH). Examples in the text using these data employ a reduced dataset of trees of more than 10 cm available in the package *vegan*.[5] Some other examples use the inventories

[1] R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

[2] E. Marcon and B. Hérault (2015). "entropart, an R Package to Measure and Partition Diversity". In: *Journal of Statistical Software* 67.8, pp. 1–26. DOI: 10.18637/jss.v067.i08.

[3] http://www.ecofog.gf/spip.php?article427

[4] R. Condit et al. (2012). "Thirty Years of Forest Census at Barro Colorado and the Importance of Immigration in Maintaining Diversity". In: *PLoS ONE* 7.11, e49826. DOI: 10.1371/journal.pone.0049826.

[5] J. Oksanen et al. (2012). *vegan: Community Ecology Package*. URL: http://cran.r-project.org/package=vegan.

Figure 2.1 – Accumulation curve of tree species of the Barro Colorado Island site. The number of species is cumulated in the order of the one-hectare squares of the site.
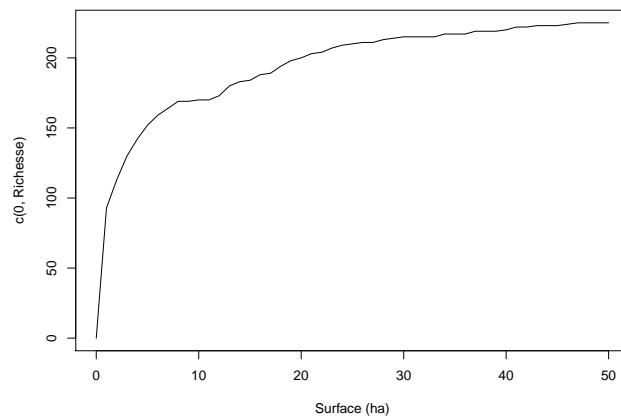
[6] S. Gourlet-Fleury et al. (2004). *Ecology & management of a neotropical rainforest. Lessons drawn from Paracou, a long-term experimental research site in French Guiana.* Paris: Elsevier.

of the Paracou research station, located in the tropical forest of French Guiana.[6] Data comes principally from two one-hectare plots at the site, numbers 6 and 18. This data, again of trees of more than 10 cm DBH, is available in the package *entropart*.

2.2 SAD and SAR

The species area relationship (SAR) curve represents the number of species observed as a function of the area sampled (Figure **??**, page **??**). There are several ways to account for this relationship,[7] but which can be divided into two large groups:[8]

[7] S. M. Scheiner (2003). "Six types of species-area curves". In: *Global Ecology and Biogeography* 12.6, pp. 441–447. DOI: 10.1046/j.1466-822X.2003.00061.x.

[8] J. Dengler (2009). "Which function describes the species-area relationship best? A review and empirical evaluation". In: *Journal of Biogeography* 36.4, pp. 728–744. DOI: 10.1111/j.1365-2699.2008.02038.x.

- In the strictest sense of the term SAR, each point represents a community. The issue is the relationship between the number of species and the size of each community;
- If the points represent a different sampling effort for each community, this is defined as a species accumulation curve (SAC). A rarefaction curve can be calculated by reducing the sampling effort using statistical tools. This provides a theoretical SAC which is not affected by the order of the data.

The Figure 2.1 [9] shows species accumulation for BCI data. A SAC can be plotted against the area, the number of individuals, or the number of sample plots, as needed.

Species abundance distribution (SAD) is the law which provides the expected abundance of each species of a community. Species are not identified individually, but by the number of

[9] R code for the figure:

```r
library("vegan")
data(BCI)
Cumul <- apply(BCI, 2, cumsum)
Richesse <- apply(Cumul, 1, function(x) sum(x > 0))
plot(y = c(0, Richesse), x = 0:50, type = "l", xlab = "Surface (ha)")
```
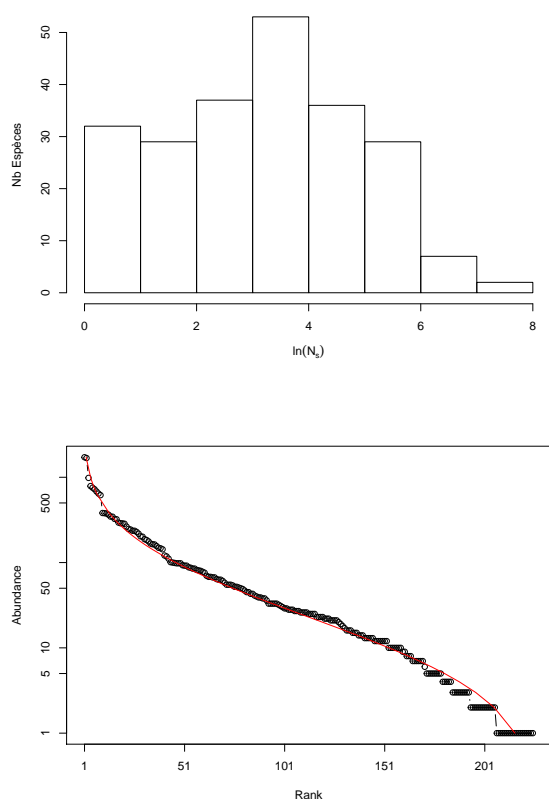
Figure 2.2 – Frequency histogram (Preston plot, above) of trees of the BCI site. X-axis: the number of trees of each species (in logarithm); Y-axis: the number of species. Corresponding rank-abundance curve (Whittaker plot, below). The points are data, the curve is the adjustment of a lognormal distribution.

individuals they have (Figure 2.2 [10]). It can be represented as a frequency histogram (Preston[11] plot, above) or as a rank-abundance curve (RAC or Whittaker[12] plot, below). RACs are often used to recognise known distributions. Izsák and Pavoine[13] studied the properties of RACs for the main SADs.

SADs are not examined in much detail here: please refer to Magurran,[14] McGill et al.,[15] and Izsák and Pavoine. The SADs necessary for comprehension of this document are:

- The geometric distribution;[16]
- The log-series distribution of Fisher et al.;[17]
- The log-normal distribution;[18]
- The broken-stick model.[19]

---

[10]R code:

```
data(BCI)
Ns <- sort(colSums(BCI), decr = TRUE)
N <- sum(Ns)
hist(log(Ns), main = NULL , xlab = expression(ln(N[s])),
ylab="Nb Espèces")

library("entropart")
plot(as.AbdVector(Ns), Distribution = "lnorm")
```

[11]F. W. Preston (1948). "The Commonness, And Rarity, of Species". In: *Ecology* 29.3, pp. 254–283.

[12]Whittaker (1965). "Dominance and diversity in land plant communities", see n. 14, p. 4.

[13]J. Izsák and S. Pavoine (2012). "Links between the species abundance distribution and the shape of the corresponding rank abundance curve". In: *Ecological Indicators* 14.1, pp. 1–6. DOI: 10.1016/j.ecolind.2011.06.030.

[14]A. E. Magurran (1988). *Ecological diversity and its measurement*. Princeton, NJ: Princeton University Press.

[15]B. J. McGill et al. (2007). "Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework". In: *Ecology Letters* 10.10, pp. 995–1015. DOI: 10.1111/j.1461-0248.2007.01094.x.

[16]I. Motomura (1932). "On the statistical treatment of communities". Japanese. In: *Zoological Magazine* 44, pp. 379–383; R. H. Whittaker (1972). "Evolution and Measurement of Species Diversity". In: *Taxon* 21.2/3, pp. 213–251. DOI: 10.2307/1218190.

[17] R. A. Fisher et al. (1943). "The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population". In: *Journal of Animal Ecology* 12, pp. 42–58.

[18] Preston (1948). "The Commonness, And Rarity, of Species", see n. 11, p. 13.

[19] R. H. MacArthur (1957). "On the Relative Abundance of Bird Species". In: *Proceedings of the National Academy of Sciences of the United States of America* 43.3, pp. 293–295.

[20] I. J. Good (1953). "The Population Frequency of Species and the Estimation of Population Parameters". In: *Biometrika* 40.3/4, pp. 237–264. DOI: 10.1093/biomet/40.3-4.237.

2.3 Coverage

2.3.1 Good-Turing frequency formula

The fundamental relationship between species frequencies is due to Turing and was published by Good.[20] In the absence of any information on the distribution of species, and assuming only that individuals are drawn independently of each other according to a multinomial distribution, the Good-Turing formula relates the mean probability α_ν of a species represented ν times (that is to say by ν individuals) to the ratio between the numbers of species represented $\nu + 1$ times and ν times:

$$\alpha_\nu \approx \frac{(\nu + 1)}{n} \frac{s_{\nu+1}^n}{s_\nu^n}. \tag{2.1}$$

Singletons (s_1^n: the number of species observed one time) and doubletons (s_2^n: the number of species observed two times) are of central importance. For $\nu = 1$, we have : $\alpha_1 = {}^{2s_2^n}/s_1^n$: the frequency of a species typically represented by a singleton is approximately equal to twice the ratio between the number of singletons and doubletons. For $\nu = 0$, not knowing the number of non-sampled species s_0^n poses a problem, but the product $\alpha_0 \times s_0^n = \pi_0$, the total probability of unrepresented species can be estimated by ${}^{s_1^n}/n$. These relationships are the foundation of the Chao richness estimators presented below.

[21] C.-H. Chiu et al. (2014). "An Improved Nonparametric Lower Bound of Species Richness via a Modified Good-Turing Frequency Formula". In: *Biometrics* 70.3, pp. 671–682. DOI: 10.1111/biom.12200, eq. 6 et 7a.

The relationship has been improved[21] by limiting the approximations in the calculations. The only requirement is that the propabilities that species represented the same number of times ν vary little and can thus all be considered equal to α_ν. So, α_ν is estimated by

$$\hat{\alpha}_\nu = \frac{(\nu + 1)s_{\nu+1}^n}{(n - \nu)s_\nu^n + (\nu + 1)s_{\nu+1}^n}. \tag{2.2}$$

This new estimator is the foundation of the improved Chao estimator and Chao and Jost entropy estimators (pages **??** and **??**).

2.3.2 Sample coverage and coverage deficit

[22] Good (1953). See n. 20.

Good[22] defines the *coverage* of sampling as the proportion of species discovered,

$$C = \sum_{s=1}^{S} \mathbf{1}\,(n_s > 0)\,p_s,$$

where $\mathbf{1}(\cdot)$ is the indicator function. Its complement to 1 is called the coverage deficit.

Sample coverage increases with sampling effort. The greater it is, the more accurate are the diversity estimates. To compare two

communities using rarefaction curves, Chao and Jost[23] shows that the sample coverage, rather than the sample size, should be the same. The diversity estimators developed below generally take this concept as their foundation for the correction of sampling bias[24] (namely, the systematic underestimation of diversity due to unobserved species, a principal contributor to estimation bias).

According to the frequency relationship mentioned above, the estimator of sample coverage, which Good attributes to Turing is:

$$\hat{C} = 1 - \frac{s_1^n}{n}. \tag{2.3}$$

This estimator is biased.[25] In reality,

$$C = 1 - \frac{\mathbb{E}(S_1^n) - \pi_1}{n}. \tag{2.4}$$

Good's estimator neglects the term π_1, the total of the probabilities of species observed one time. This term can be estimated with a smaller bias. Chao et al.,[26] and also Zhang and Huang,[27] propose the following estimator, which utilizes all available information and which thus has the lowest possible bias:

$$\hat{C} = 1 - \sum_{\nu=1}^{n} (-1)^{\nu+1} \binom{n}{\nu}^{-1} s_\nu^n. \tag{2.5}$$

The terms of the sum decrease very quickly with ν. By limiting it to $\nu = 1$, the estimator is reduced to that of Good.

Esty,[28] augmented by Zhang and Zhang,[29] has shown that the estimator was asymptomatically normal, and calculated the confidence interval of $\hat{C}$:

$$C = \hat{C} \pm t_{1-\alpha/2}^n \frac{\sqrt{s_1^n \left(1 - \frac{s_1^n}{n}\right) + 2s_2^n}}{n}. \tag{2.6}$$

Where $t_{1-\alpha/2}^n$ is the quantile of a Student's t-distribution with n degrees of freedom at the risk threshold α, classically 1,96 for n large and $\alpha = 5\%$.

A different estimator is used in the SPADE software package[30] and its translation into R, creating the package *spadeR*.[31] It is the basis of Chao and Jost's entropy estimators (page **??**). The estimation of the equation (2.4) gives the relationship

$$\hat{C} = 1 - \frac{s_1^n - \hat{\pi}_1}{n}. \tag{2.7}$$

Yet, $\hat{\pi}_1 = s_1^n \hat{\alpha}_1$. α_1 can be estimated by the Good-Turing frequency estimation (2.2), by replacing s_0^n with the Chao1 estimate (équation **??**). So:

[24]G. Dauby and O. J. Hardy (2012). "Sampled-based estimation of diversity sensu stricto by transforming Hurlbert diversities into effective number of species". In: *Ecography* 35.7, pp. 661–672. DOI: 10.1111/j.1600-0587.2011.06860.x.

[25]Z. Zhang and H. Huang (2007). "Turing's formula revisited". In: *Journal of Quantitative Linguistics* 14.2-3, pp. 222–241. DOI: 10.1080/09296170701514189.

[26]A. Chao et al. (1988). "A generalized Good's nonparametric coverage estimator". In: *Chinese Journal of Mathematics* 16, pp. 189–199.

[27]Zhang and Huang (2007). See n. 25.

[28]W. W. Esty (1983). "A Normal Limit Law for a Nonparametric Estimator of the Coverage of a Random Sample". In: *The Annals of Statistics* 11.3, pp. 905–912. DOI: 10.2307/2240652.

[29]C.-H. Zhang and Z. Zhang (2009). "Asymptotic Normality of a Nonparametric Estimator of Sample Coverage". In: *Annals of Statistics* 37.5A, pp. 2582–2595. DOI: 10.1214/08-aos658.

[30]A. Chao and T.-J. Shen (2010). *Program SPADE: Species Prediction And Diversity Estimation. Program and user's guide.* Hsin-Chu, Taiwan. URL: http://chao.stat.nthu.edu.tw/wordpress/wp-content/uploads/software/SPADE{_}UserGuide.pdf.

[31]A. Chao et al. (2016). *SpadeR: Species Prediction and Diversity Estimation with R.* URL: http://chao.stat.nthu.edu.tw/blog/software-download/.

$$\hat{C} = 1 - \frac{s_1^n}{n}(1 - \hat{\alpha}_1) = 1 - \frac{s_1^n}{n}\left[\frac{(n-1)s_1^n}{(n-1)s_1^n + 2s_2^n}\right]. \qquad (2.8)$$

In the package *entropart*, the `Coverage` function calculates three estimators (that of Zhang and Huang by default):

```
data(Paracou618)
Coverage(Paracou618.MC$Ns)

## ZhangHuang
## 0.9226675
```

[32] Chao and Jost (2012). "Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size", see n. **??**, p. **??**.

Chao and Jost[32] shows that the slope of the accumulation curve giving the expectation of the number of species as a function of the number of individuals (rarefaction curve in Figure **??**) is equal to the coverage deficit,

$$1 - \mathbb{E}\left(C^n\right) = \mathbb{E}\left(S^{n+1}\right) - \mathbb{E}\left(S^n\right), \qquad (2.9)$$

where C^n is the coverage of a sampling of size n, and S^n is the number of species discovered in this sampling.

The estimators presented here suppose a population of infinite size (equivalently, individuals are drawn randomly without replacement). For populations of finite size, see Chao and Lin[33] and Hwang et al.[34]

[33] A. Chao and C.-W. Lin (2012). "Nonparametric Lower Bounds for Species Richness and Shared Species Richness under Sampling without Replacement". In: *Biometrics* 68.3, pp. 912–921. DOI: 10.1111/j.1541-0420.2011.01739.x.

[34] W.-H. Hwang et al. (2014). "Good-Turing frequency estimation in a finite population". In: *Biometrical journal* 57.2, pp. 321–339. DOI: 10.1002/bimj.201300168.

2.3.3 Completeness

The completeness of a sampling effort is the proportion of the number of observed species: $s_{\neq 0}^n / S$. It is only a count of the number of species and must not be confused with coverage, which is the sum of their probabilities: the completeness rate is always much lower than the coverage rate because those species which are not sampled are the rarest.

The completeness of the sampling effort of tropical forest trees in the example just mentioned can be estimated by dividing the number of species observed by the the estimated number of species (see page **??**):

```
# Species observed
(Obs <- Richness(Paracou618.MC$Ps))

## None
## 229

# Estimated richness
(Est <- Richness(Paracou618.MC$Ns, Correction = "Jackknife"))

## Jackknife 2
## 359

# Completeness
as.numeric(Obs/Est)

## [1] 0.637883
```

Bibliography

Agapow, P. M., O. R. P. Binindal-Emonds, K. A. Crandall, J. L. Gittleman, G. M. Mace, J. C. Marshall, and A. Purvis (2004). "The impact of species concept on biodiversity studies". In: *The Quarterly Review of Biology* 79.2, pp. 161–179 (cit. on p. 8).

Barberousse, A. and S. Samadi (2014). "La taxonomie et les collections d'histoire naturelle à l'heure de la sixième extinction". In: *La biodiversité en question. Enjeux philosophiques, éthiques et scientifiques.* Ed. by E. Casetta and J. Delord. Paris: Editions Matériologiques. Chap. 6, pp. 155–182 (cit. on p. 8).

Blandin, P. (2014). "La diversité du vivant avant (et après) la biodiversité : repères historiques et épistémologiques". In: *La biodiversité en question. Enjeux philosophiques, éthiques et scientifiques.* Ed. by E. Casetta and J. Delord. Paris: Editions Matériologiques. Chap. 1, pp. 31–68 (cit. on p. 3).

Cardinale, B. J., J. E. Duffy, A. Gonzalez, D. U. Hooper, C. Perrings, P. Venail, A. Narwani, G. M. Mace, D. Tilman, D. A. Wardle, A. P. Kinzig, G. C. Daily, M. Loreau, J. B. Grace, A. Larigauderie, D. S. Srivastava, and S. Naeem (2012). "Biodiversity loss and its impact on humanity". In: *Nature* 486.7401, pp. 59–67. DOI: 10.1038/nature11148 (cit. on p. 4).

Casetta, E. (2014). "Évaluer et conserver la biodiversité face au problème des espèces". In: *La biodiversité en question. Enjeux philosophiques, éthiques et scientifiques.* Ed. by E. Casetta and J. Delord. Paris: Editions Matériologiques. Chap. 5, pp. 139–154 (cit. on p. 8).

Ceballos, G., P. R. Ehrlich, and R. Dirzo (2017). "Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines". In: *Proceedings of the National Academy of Sciences*, p. 201704949. DOI: 10.1073/pnas.1704949114 (cit. on p. 4).

Chao, A. and L. Jost (2012). "Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size". In: *Ecology* 93.12, pp. 2533–2547. DOI: 10.1890/11-1952.1 (cit. on pp. 15, 16).

Chao, A. and C.-W. Lin (2012). "Nonparametric Lower Bounds for Species Richness and Shared Species Richness under Sampling without Replacement". In: *Biometrics* 68.3, pp. 912–921. DOI: 10.1111/j.1541-0420.2011.01739.x (cit. on p. 16).

Chao, A. and T.-J. Shen (2010). *Program SPADE: Species Prediction And Diversity Estimation. Program and user's guide.* Hsin-Chu, Taiwan. URL: http://chao.stat.nthu.edu.tw/wordpress/wp-content/uploads/software/SPADE{_}UserGuide.pdf (cit. on p. 15).

Chao, A., S.-M. Lee, and T.-C. Chen (1988). "A generalized Good's nonparametric coverage estimator". In: *Chinese Journal of Mathematics* 16, pp. 189–199 (cit. on p. 15).

Chao, A., K. H. Ma, T. C. Hsieh, and C.-H. Chiu (2016). *SpadeR: Species Prediction and Diversity Estimation with R.* URL: http://chao.stat.nthu.edu.tw/blog/software-download/ (cit. on p. 15).

Chapin, F. S. I., E. S. Zavaleta, V. T. Eviner, R. L. Naylor, P. M. Vitousek, H. L. Reynolds, D. U. Hooper, S. Lavorel, O. E. Sala, S. E. Hobbie, M. C. Mack, and S. Díaz (2000). "Consequences of changing biodiversity". In: *Nature* 405.6783, pp. 234–242. DOI: 10.1038/35012241 (cit. on p. 4).

Chiu, C.-H., Y.-T. Wang, B. A. Walther, and A. Chao (2014). "An Improved Nonparametric Lower Bound of Species Richness via a Modified Good-Turing Frequency Formula". In: *Biometrics* 70.3, pp. 671–682. DOI: 10.1111/biom.12200 (cit. on p. 14).

Condit, R., R. A. Chisholm, and S. P. Hubbell (2012). "Thirty Years of Forest Census at Barro Colorado and the Importance of Immigration in Maintaining Diversity". In: *PLoS ONE* 7.11, e49826. DOI: 10.1371/journal.pone.0049826 (cit. on p. 11).

Cracraft, J. (1983). "Species Concepts and Speciation Analysis". In: *Current Ornithology Volume 1.* Ed. by R. F. Johnston. Vol. 1. Current Ornithology. Springer US, pp. 159–187. DOI: 10.1007/978-1-4615-6781-3_6 (cit. on p. 8).

Dauby, G. and O. J. Hardy (2012). "Sampled-based estimation of diversity sensu stricto by transforming Hurlbert diversities into effective number of species". In: *Ecography* 35.7, pp. 661–672. DOI: 10.1111/j.1600-0587.2011.06860.x (cit. on p. 15).

DeLong, D. C. J. (1996). "Defining Biodiversity". In: *Wildlife Society Bulletin* 24.4, pp. 738–749 (cit. on p. 3).

Delord, J. (2014). "La biodiversité : imposture scientifique ou ruse épistémologique ?" In: *La biodiversité en question. Enjeux philosophiques, éthiques et scientifiques.* Ed. by E. Casetta and J. Delord. Paris: Editions Matériologiques. Chap. 3, pp. 83–118. DOI: 10.3917/edmat.delor.2014.01.0083 (cit. on p. 3).

Dengler, J. (2009). "Which function describes the species-area relationship best? A review and empirical evaluation". In: *Journal of Biogeography* 36.4, pp. 728–744. DOI: 10.1111/j.1365-2699.2008.02038.x (cit. on p. 12).

Dobzhansky, T. (1937). *Genetics and the Origin of Species.* New York: Columbia University Press (cit. on p. 8).

Esty, W. W. (1983). "A Normal Limit Law for a Nonparametric Estimator of the Coverage of a Random Sample". In: *The Annals of Statistics* 11.3, pp. 905–912. DOI: 10.2307/2240652 (cit. on p. 15).

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". In: *Journal of Animal Ecology* 12, pp. 42–58 (cit. on pp. 13, 14).

Gaston, K. J. (2000). "Global patterns in biodiversity". In: *Nature* 405.6783, pp. 220–227. DOI: 10.1038/35012228 (cit. on p. 7).

Good, I. J. (1953). "The Population Frequency of Species and the Estimation of Population Parameters". In: *Biometrika* 40.3/4, pp. 237–264. DOI: 10.1093/biomet/40.3-4.237 (cit. on p. 14).

Gourlet-Fleury, S., J. M. Guehl, and O. Laroussinie (2004). *Ecology & management of a neotropical rainforest. Lessons drawn from Paracou, a long-term experimental research site in French Guiana.* Paris: Elsevier (cit. on p. 12).

Hey, J. (2001). "The mind of the species problem". In: *Trends in Ecology & Evolution* 16.7, pp. 326–329. DOI: 10.1016/S0169-5347(01)02145-0 (cit. on pp. 4, 8).

Hwang, W.-H., C.-W. Lin, and T.-J. Shen (2014). "Good-Turing frequency estimation in a finite population". In: *Biometrical journal* 57.2, pp. 321–339. DOI: 10.1002/bimj.201300168 (cit. on p. 16).

Izsák, J. and S. Pavoine (2012). "Links between the species abundance distribution and the shape of the corresponding rank abundance curve". In: *Ecological Indicators* 14.1, pp. 1–6. DOI: 10.1016/j.ecolind.2011.06.030 (cit. on p. 13).

Jizhong, Z., M. Shijun, and C. Changming (1991). "An index of ecosystem diversity". In: *Ecological Modelling* 59, pp. 151–163. DOI: 10.1016/0304-3800(91)90176-2 (cit. on p. 4).

Jost, L. (2010). "The Relation between Evenness and Diversity". In: *Diversity* 2.2, pp. 207–232. DOI: 10.3390/d2020207 (cit. on p. 6).

Loreau, M. (2005). "Discours de clôture". In: *Actes de la Conférence internationale Biodiversité Science et Gouvernance.* Ed. by R. Barbault and J.-P. Le Duc. Paris, France.: IRD Editions, pp. 254–256 (cit. on p. 3).

MacArthur, R. H. (1957). "On the Relative Abundance of Bird Species". In: *Proceedings of the National Academy of Sciences of the United States of America* 43.3, pp. 293–295 (cit. on p. 14).

Magurran, A. E. (1988). *Ecological diversity and its measurement.* Princeton, NJ: Princeton University Press (cit. on p. 13).

Marcon, E. and B. Hérault (2015). "entropart, an R Package to Measure and Partition Diversity". In: *Journal of Statistical Software* 67.8, pp. 1–26. DOI: 10.18637/jss.v067.i08 (cit. on p. 11).

Mayden, R. L. (1997). "A hierarchy of species concepts: the denouement in the saga of the species problem". In: *Species. The units of biodiversity.* Ed. by M. F. Claridge, H. A. Dawah, and M. R. Wilson. London: Chapman and Hall. Chap. 19, pp. 381–424 (cit. on p. 8).

Mayr, E. (1942). *Systematics and the origin of species from the viewpoint of a zoologist.* New York: Columbia University Press (cit. on p. 8).

McGill, B. J., R. S. Etienne, J. S. Gray, D. Alonso, M. J. Anderson, H. K. Benecha, M. Dornelas, B. J. Enquist, J. L. Green, F. He, A. H. Hurlbert, A. E. Magurran, P. A. Marquet, B. A. Maurer, A. Ostling, C. U. Soykan, K. I. Ugland, and E. P. White (2007). "Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework". In: *Ecology Letters* 10.10, pp. 995–1015. DOI: 10.1111/j.1461-0248.2007.01094.x (cit. on p. 13).

Mcintosh, R. P. (1967). "An Index of Diversity and the Relation of Certain Concepts to Diversity". In: *Ecology* 48.3, pp. 392–404 (cit. on p. 4).

Meine, C., M. Soulé, and R. F. Noss (2006). ""A mission-driven discipline": The growth of conservation biology". In: *Conservation Biology* 20.3, pp. 631–651. DOI: 10.1111/j.1523-1739.2006.00449.x (cit. on p. 3).

Miraldo, A., S. Li, M. K. Borregaard, A. Florez-Rodriguez, S. Gopalakrishnan, M. Rizvanovic, Z. Wang, C. Rahbek, K. A. Marske, and D. Nogues-Bravo (2016). "An

Anthropocene map of genetic diversity". In: *Science* 353.6307, pp. 1532–1535. DOI: 10.1126/science.aaf4381 (cit. on p. 7).

Moreno, C. E. and P. Rodríguez (2010). "A consistent terminology for quantifying species diversity?" In: *Oecologia* 163.2, pp. 279–82. DOI: 10.1007/s00442-010-1591-7 (cit. on p. 7).

Motomura, I. (1932). "On the statistical treatment of communities". Japanese. In: *Zoological Magazine* 44, pp. 379–383 (cit. on p. 13).

Mouillot, D., W. Stubbs, M. Faure, O. Dumay, J.-A. Tomasini, J. B. Wilson, and T. D. Chi (2005). "Niche overlap estimates based on quantitative functional traits: a new family of non-parametric indices". In: *Oecologia* 145.3, pp. 345–353. DOI: 10.1007/s00442-005-0151-z (cit. on p. 6).

Oksanen, J., F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, and H. Wagner (2012). *vegan: Community Ecology Package.* URL: http://cran.r-project.org/package=vegan (cit. on p. 11).

Pavoine, S. and M. B. Bonsall (2011). "Measuring biodiversity to explain community assembly: a unified approach". In: *Biological Reviews* 86.4, pp. 792–812. DOI: 10.1111/j.1469-185X.2010.00171.x (cit. on p. 6).

Pernès, J., ed. (1984). *Gestion des ressources génétiques des plantes. Tome 2 : Manuel.* Paris: Agence de Coopération culturelle et technique (cit. on p. 8).

Preston, F. W. (1948). "The Commonness, And Rarity, of Species". In: *Ecology* 29.3, pp. 254–283 (cit. on pp. 13, 14).

Purvis, A. and A. Hector (2000). "Getting the measure of biodiversity." In: *Nature* 405.6783, pp. 212–9. DOI: 10.1038/35012221 (cit. on p. 4).

R Core Team (2017). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing (cit. on p. 11).

Richards, R. A. (2010). *The Species Problem. A Philosophical Analysis.* Cambridge: Cambridge University Press (cit. on p. 8).

Ricotta, C. (2007). "A semantic taxonomy for diversity measures". In: *Acta Biotheoretica* 55.1, pp. 23–33. DOI: 10.1007/s10441-007-9008-7 (cit. on p. 6).

Runnegar, B. (1987). "Rates and Modes of Evolution in the Mollusca". In: *Rates of Evolution.* Ed. by M. Campbell and M. F. Day. London: Allen & Unwin, pp. 39–60 (cit. on p. 6).

Scheiner, S. M. (2003). "Six types of species-area curves". In: *Global Ecology and Biogeography* 12.6, pp. 441–447. DOI: 10.1046/j.1466-822X.2003.00061.x (cit. on p. 12).

Speth, J. G., M. W. Holdgate, and M. K. Tolba (1992). "Foreword". In: *Global Biodiversity Strategy.* Ed. by K. Courrier. Washington, D.C.: WRI, IUCN, UNEP, pp. v–vi (cit. on p. 3).

Stirling, A. (2007). "A general framework for analysing diversity in science, technology and society". In: *Journal of the Royal Society, Interface* 4.15, pp. 707–719. DOI: 10.1098/rsif.2007.0213 (cit. on p. 6).

Sukumaran, J. and L. L. Knowles (2017). "Multispecies coalescent delimits structure, not species". In: *Proceedings of the National Academy of Sciences of the United States of America* in press. DOI: 10.1073/PNAS.1607921114 (cit. on p. 8).

Van Valen, L. (1976). "Ecological Species, Multispecies, and Oaks". In: *Taxon* 25.2/3, pp. 233–239. DOI: 10.2307/1219444 (cit. on p. 8).

Whittaker, R. H. (1960). "Vegetation of the Siskiyou Mountains, Oregon and California". In: *Ecological Mono-*

graphs 30.3, pp. 279–338. DOI: 10.2307/1943563 (cit. on p. 6).

— (1965). "Dominance and diversity in land plant communities". In: *Science* 147.3655, pp. 250–260 (cit. on pp. 4, 13).

— (1972). "Evolution and Measurement of Species Diversity". In: *Taxon* 21.2/3, pp. 213–251. DOI: 10.2307/1218190 (cit. on p. 13).

— (1977). "Evolution of species diversity in land communities". In: *Evolutionary Biology* 10. Ed. by M. K. Hecht, W. C. Steere, and B. Wallace, pp. 1–67 (cit. on p. 7).

Wilson, E. O. and F. M. Peter, eds. (1988). *Biodiversity*. Washington, D.C.: The National Academies Press (cit. on p. 3).

Zhang, C.-H. and Z. Zhang (2009). "Asymptotic Normality of a Nonparametric Estimator of Sample Coverage". In: *Annals of Statistics* 37.5A, pp. 2582–2595. DOI: 10.1214/08-aos658 (cit. on p. 15).

Zhang, Z. and H. Huang (2007). "Turing's formula revisited". In: *Journal of Quantitative Linguistics* 14.2-3, pp. 222–241. DOI: 10.1080/09296170701514189 (cit. on p. 15).