Mesures de la Biodiversité

Eric Marcon

16/12/2024

Ce document est réalisé de façon dynamique et reproductible grâce à:

- LaTeX, dans sa distribution Miktex (http://miktex.org/) et la classe memoir (http://www.ctan.org/pkg/memoir).
- R (http://www.r-project.org/) et RStudio (http://www.rstudio.com/)
- bookdown (http://bookdown.org/)

Son code source est sur GitHub: https://github.com/EricMarcon/MesuresBioDiv3/. Le texte mis à jour en continu peut être lu sur https://ericmarcon.github.io/MesuresBioDiv3/. Les versions d'étape sont déposées sur HAL: https://hal-agroparistech.archives-ouvertes.fr/cel-01205813/.

Photographie en couverture: Hadrien Lalagüe

Contents

Motivation

Le terme *biodiversity* est attribué[1] à Walter Rosen, un membre du *National Research Council* américain, qui a commencé à contracter les termes *biological diversity* pendant la préparation d'un colloque dont les actes seront publiés sous le titre "Biodiversity".[2] La question de la diversité biologique intéressait les écologues bien avant l'invention de la biodiversité, mais le néologisme a connu un succès fulgurant[3] en même temps qu'il devenait une notion floue, dans lequel chacun peut placer ce qu'il souhaite y trouver, au point de lui retirer son caractère scientifique.[4] Une cause de ce glissement est que la biodiversité a été nommée pour attirer l'attention sur son érosion, en lien avec la biologie de la conservation. Cette érosion concernant potentiellement de nombreux aspects du monde vivant, la définition de la biodiversité fluctue selon les besoins: DeLong[5] en recense 85 dans les dix premières années de littérature. Les indicateurs de la biodiversité peuvent englober bien d'autres choses que la diversité du vivant: le nombre d'espèces menacées (par exemple la liste rouge de l'IUCN), la taille des populations ou la surface des écosystèmes préservés, la dégradation des habitats, la menace pesant sur des espèces emblématiques… Une mesure rigoureuse et cohérente de la diversité peut pourtant être construite pour clarifier beaucoup (mais pas tous) des concepts qui constituent la biodiversité.

Dans l'introduction du premier chapitre des actes de ce qui était devenu le "Forum sur la Biodiversité", Wilson utilise le mot dans le sens étroit de nombres d'espèces. L'élargissement de la notion aux "systèmes naturels" et à l'opposé à la diversité génétique intraspécifique est venu du monde de la conservation.[6] La déclaration de Michel Loreau, président du du comité scientifique de la conférence de Paris en 2005[7] en donne une définition aboutie:

> La Terre abrite une extraordinaire diversité biologique, qui inclut non seulement les espèces qui habitent notre planète, mais aussi la diversité de leurs gènes, la multitude des interactions écologiques entre elles et avec leur environnement physique, et la variété des écosystèmes complexes

[1] C. Meine et al. ""A Mission-Driven Discipline": The Growth of Conservation Biology." In: *Conservation Biology* 20.3 (2006), pp. 631–651. DOI: 10.1111/j.1523-1739.2006.00449.x.

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

[3] P. Blandin. "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, 2014, pp. 31–68.

[4] J. Delord. "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, 2014, pp. 83–118. DOI: 10.3917/edmat.delor.2014.01.0083.

[5] D. C. J. DeLong. "Defining Biodiversity." In: *Wildlife Society Bulletin* 24.4 (1996), pp. 738–749. JSTOR: 3783168.

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

[7] M. Loreau. "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, 2005, pp. 254–256.

qu'elles constituent. Cette biodiversité, qui est le produit de plus de 3 milliards d'années d'évolution, constitue un patrimoine naturel et une ressource vitale dont l'humanité dépend de multiples façons.

Aujourd'hui encore, le terme *biodiversité* concerne le plus souvent la richesse en espèces d'un écosystème. Pour simplifier la présentation, le niveau d'étude dans ce document sera en général celui des espèces.[8] La prise en compte de la totalité des êtres vivants est généralement hors de portée. La mesure de diversité est alors limitée à un taxocène, c'est-à-dire un sous-ensemble des espèces d'une communauté reliées taxonomiquement: les papillons, les mammifères, les arbres (la délimitation du sous-ensemble n'est pas forcément strictement taxonomique)…

Un objet privilégié des études sur la biodiversité est, depuis le Forum, la forêt tropicale parce qu'elle est très diverse et un enjeu pour la conservation. La plupart des exemples concerneront ici les arbres de la forêt tropicale, qui ont l'avantage d'être clairement définis en tant qu'individus (donc simples à compter) et posent des problèmes méthodologiques considérables pour l'estimation de leur diversité à partir de données réelles.

On peut bien évidemment s'intéresser à d'autres niveaux et d'autres objets, par exemple la diversité génétique (en termes d'allèles différents pour certains gènes ou marqueurs) à l'intérieur d'une population, ou même la diversité des interactions entre espèces d'une communauté.[9] On gardera toujours à l'esprit que la prise en compte de la diversité spécifique n'est pas la seule approche, les méthodes présentées ici s'appliquent à la mesure de la diversité en général, pas même nécessairement biologique.

L'objectif de ce document est de traiter la mesure de la biodiversité, pas son importance en tant que telle. On se référera par exemple à Chapin et al.[10] pour une revue sur cette question, Cardinale et al.[11] pour les conséquences de l'érosion de la biodiversité sur les services écosystémiques ou Ceballos et al.[12] pour les propriétés autocatalytiques de la biodiversité.

La mesure de la diversité est un sujet important en tant que tel,[13] pour permettre de formaliser les concepts et de les appliquer à la réalité. La question est loin d'être épuisée, et fait toujours l'objet d'une recherche active et de controverses.[14]

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

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

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

[11] B. J. Cardinale et al. "Biodiversity Loss and Its Impact on Humanity." In: *Nature* 486.7401 (2012), pp. 59–67. DOI: 10.1038/nature11148.

[12] G. Ceballos et al. "Biological Annihilation via the Ongoing Sixth Mass Extinction Signaled by Vertebrate Population Losses and Declines." In: *Proceedings of the National Academy of Sciences* (2017), p. 201704949. DOI: 10.1073/pnas.1704949114.

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

[14] C. Ricotta. "Through the Jungle of Biological Diversity." In: *Acta Biotheoretica* 53.1 (2005), pp. 29–38. DOI: 10.1007/s10441-005-7001-6.

Calculs et données

La présentation des mesures de diversité est donnée avec un usage intensif du formalisme mathématique. La liste des notations est fournie ci-dessous, on s'y référera autant que nécessaire.

Les calculs sont réalisés dans R,[15] essentiellement avec le package *divent*,[16] qui succède au package *entropart*.[17] L'ensemble du code est disponible sur GitHub[18] où se trouvent les mises à jour de ce document[19].

Les données sont souvent celles de la parcelle 6 de la forêt de Paracou (figure 1 en Guyane française,[20] d'une surface de 6.25 ha. Tous les arbres de plus de 10 cm de diamètre à hauteur de poitrine (DBH: *Diameter at Breast Height*) y ont été inventoriés en 2016. La position de chaque arbre, son espèce et sa surface terrière sont fournis.

D'autres exemples utilisent la parcelle forestière permanente de Barro Colorado Island, souvent abrégée BCI:[21] 50 ha de forêt tropicale dont les arbres de plus de 1 cm de diamètre à hauteur de poitrine (DBH: *Diameter at Breast Height*) ont été inventoriés. Le jeu de données utilisé pour les exemples est une version réduite aux arbres de plus de 10 cm disponible dans le package *vegan*,[22] soient 21457 arbres dans 225 espèces.

Le code R pour réaliser la figure est le suivant:

```
library("divent")
paracou_6_wmppp %>%
autoplot(
  labelSize = expression(paste("Surface terrière (", cm^2, ")")),
  labelColor = "Espèce"
) +
theme(legend.position = "bottom", legend.direction = "vertical", legend.margin=margin())
```

[15] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2024.

[16] **R-divent**.

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

[18] https://github.com/EricMarcon/MesuresBioDiv3/

[19] https://ericmarcon.github.io/MesuresBioDiv3/

[20] S. Gourlet-Fleury et al. *Ecology & Management of a Neotropical Rainforest. Lessons Drawn from Paracou, a Long-Term Experimental Research Site in French Guiana*. Paris: Elsevier, 2004.

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

[22] J. Oksanen et al. "Vegan: Community Ecology Package." In: (2012).

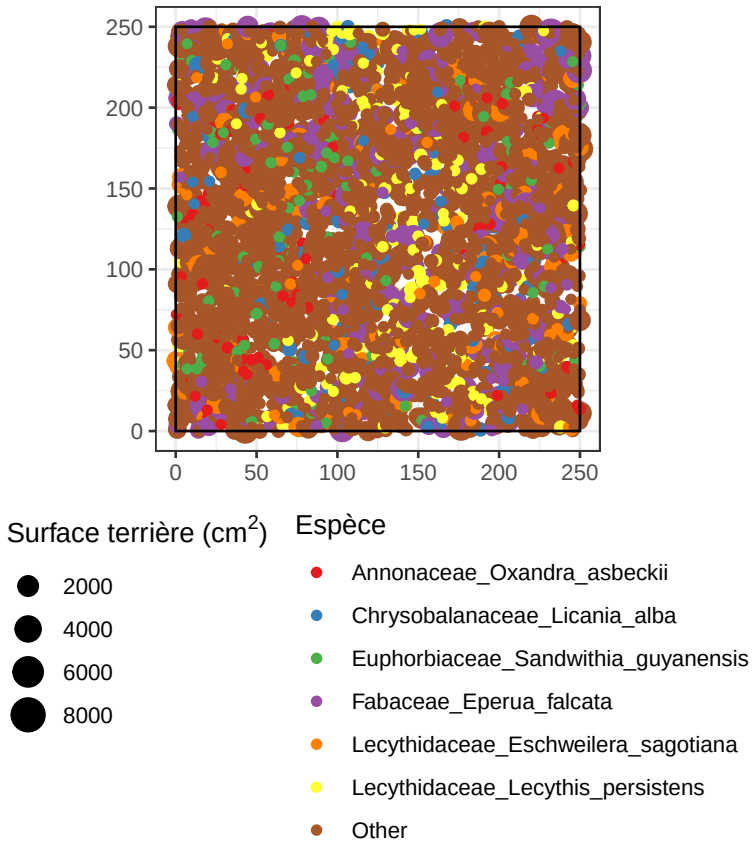


Figure 1: Carte de la parcelle 6 de Paracou. Les points représentent les arbres. Leur taille est proportionnelle à leur surface terrière. Seules les espèces les plus fréquentes sont identifiées sur la carte.

Notations

Les notations mathématiques peuvent différer de celles de la littérature citée pour l'homogénéité de ce document.

Les matrices sont notées en caractères gras et majuscules: $\mathbf{X}$. Les éléments de la matrice $\mathbf{X}$ sont notés $x_{i,j}$.

Les vecteurs sont notés en gras minuscule: $\mathbf{p}$. Les nombres sont notés en minuscules, n, et les variables aléatoires en majuscules: N. Les valeurs maximales des énumérations font exception: elles sont notées en majuscules pour les distinguer des indices: $\sum_{s=1}^{S} p_s = 1$.

Le produit matriciel de $\mathbf{X}$ et $\mathbf{Y}$ est noté $\mathbf{XY}$. Dans les scripts R, l'opérateur est `%*%`. Le produit de Hadamard (terme à terme) est noté $\mathbf{X} \circ \mathbf{Y}$ (opérateur `*` dans R). De même $\mathbf{X}^n$ indique la puissance n au sens du produit matriciel d'une matrice carrée (opérateur `%^%` du package *expm*), alors que $\mathbf{X}^{\circ n}$ est la matrice dont chaque terme est celui de $\mathbf{X}$ à la puissance n (opérateur `^` de R). La matrice transposée de $\mathbf{X}$ est notée $\mathbf{X}'$.

Les notations sont les suivantes:

$\mathbf{1}(\cdot)$: la fonction indicatrice, qui vaut 1 si la condition dans la parenthèse est vraie, 0 sinon.

$\mathbf{1}_s$: le vecteur de longueur s composé uniquement de 1. $\mathbf{1}_s \mathbf{1}_s' = \mathbf{J}_s$ où $\mathbf{J}_s$ est la matrice carré de taille s ne contenant que des 1.

A: l'aire d'étude, et, selon le contexte, sa surface.

C: le taux de couverture de l'échantillon, c'est-à-dire la probabilité qu'un individu de la communauté appartienne à une des espèces échantillonnées. C^n est le taux de couverture correspondant à un échantillon de taille n.

${}^qD$: la diversité vraie (nombre de Hill pour les diversités α et γ), nombre équivalent de communautés pour la diversité β. ${}^q_iD_\alpha$ est la diversité α mesurée dans la communauté i. ${}^q\bar{D}(T)$ est la diversité phylogénétique.

Δ: la matrice de dissimilarité dont les éléments sont $\delta_{s,t}$, la dissimilarité entre l'espèce s et l'espèce t.

$\mathbb{E}(X)$: l'espérance de la variable aléatoire X.

f_ν: le nombre d'espèces observées ν fois dans un échantillon

(qui peut être défini par sa surface ou son nombre d'individus). $f_{>0}$ est le nombre d'espèces observées au moins une fois. f_1 s'appelle le nombre de *singletons* et f_2 le nombre de *doubletons.*

${}^qH$: l'entropie de Tsallis (ou HCDT). ${}^q_iH_\alpha$ est l'entropie α mesurée dans la communauté i. Si nécessaire, le vecteur des probabilités servant au calcul est précisé sous la forme ${}^qH(\mathbf{p})$. ${}^q\bar{H}(T)$ est l'entropie phylogénétique.

I: le nombre de communautés qui constituent une partition de la méta-communauté dans le cadre de la décomposition de la diversité. Les communautés sont indexées par i.

$I(p_s)$: l'information apportée par l'observation d'un évènement de probabilité p_s. $I(q_s, p_s)$ est le gain d'information apporté par l'expérience (q_s est observé) par rapport aux probabilités p_s attendues.

$\mathbf{I}_s$: la matrice identité de rang s: matrice carrée de taille $s \times s$ dont la diagonale ne comporte que des 1 et les autres éléments sont nuls.

N: le nombre (aléatoire) d'individus se trouvant dans l'aire d'étude. N_s est la même variable aléatoire, mais restreinte aux individus de l'espèce s.

n: le nombre d'individus échantillonnés. $n_{s,i}$ est le nombre d'individus de l'espèce s dans la communauté i. Les effectifs totaux sont n_{s+} (pour l'espèce s), n_{+i} pour la communauté i et n le total général. S'il n'y a qu'une communauté, le nombre d'individus par espèce est n_s.

p_s: la probabilité qu'un individu tiré au hasard appartienne à l'espèce s, autrement dit la probabilité de l'espèce s. Son estimateur le plus simple, $\widehat{p}_s$ est la fréquence observée. Selon le contexte, $\widehat{p}_s$ peut désigner un estimateur plus élaboré. $p_{s|i}$ est la même probabilité dans la communauté i. p_ν est la probabilité d'une espèce observées ν fois dans un échantillon.

$\mathbf{p} = (p_1, p_2, \dots, p_s, \dots, p_S)$: le vecteur décrivant la distribution des probabilités p_s, appelé simplexe en référence à sa représentation dans l'espace à S dimensions.

π_ν: la probabilité qu'une espèce choisie au hasard soit représentée par ν individus, $\sum_{\nu=1}^{n} \pi_\nu = 1$. Si l'espèce est choisie explicitement, la probabilité est notée π_{n_s}.

${}^qR$: l'entropie de Rényi d'ordre q.

S: le nombre d'espèces d'une communauté, considéré comme une variable aléatoire, estimé par $\widehat{S}$.

$S(A)$ et $S(n)$: le nombre d'espèces, considéré comme une fonction de la taille de l'échantillon.

$t^n_{1-\alpha/2}$: le quantile d'une loi de Student à n degrés de liberté au seuil de risque α, classiquement 1,96 pour n grand et $\alpha = 5\%$.

Z: la matrice de similarité entre espèces dont les éléments sont
$z_{s,t}$, la similarité entre l'espèce s et l'espèce t.

$(\cdot)$: la fonction gamma.

$(\cdot)$: la fonction digamma.

$\binom{n}{k}$: le nombre de combinaisons de k éléments parmi n:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

.

Part I

Notions

Chapter 1

Notions de Diversité

1.1 Composantes

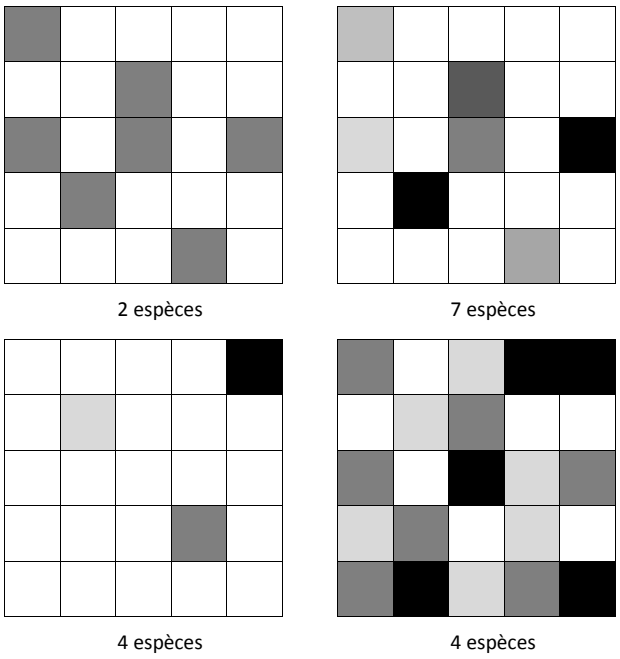


Figure 1.1: Importance de la richesse (en haut) et de l'équitabilité (en bas) pour la définition de la diversité. Ligne du haut: toutes choses égales par ailleurs, une communauté de 7 espèces semble plus diverse qu'une communauté de 2 espèces. Ligne du bas: à richesse égale, une communauté moins équitable (à gauche) semble moins diverse. Colonne de gauche: une communauté moins riche (en haut) peut sembler plus diverse si elle est plus équitable. Colonne de droite: idem pour la communauté du bas.

Une communauté comprenant beaucoup d'espèces mais avec une espèce dominante n'est pas perçue intuitivement comme plus diverse qu'une communauté avec moins d'espèces, mais dont les effectifs sont proches (figure 1.1, colonne de gauche). La prise en compte de deux composantes de la diversité, appelées richesse et équitabilité, est nécessaire.[1]

Richesse

La richesse[2] est le nombre (ou une fonction croissante du nombre) de classes différentes présentes dans le système étudié, par exemple le nombre d'espèces d'arbres dans une forêt.

[1] R. H. Whittaker. "Dominance and Diversity in Land Plant Communities." In: *Science* 147.3655 (1965), pp. 250–260. DOI: 10.1126/science.147.3655.250.

[2] terme introduit par R. P. Mcintosh. "An Index of Diversity and the Relation of Certain Concepts to Diversity." In: *Ecology* 48.3 (1967), pp. 392–404. DOI: 10.2307/1932674.

Un certain nombre d'hypothèses sont assumées plus ou moins explicitement:

- Les classes sont bien connues: compter le nombre d'espèces a peu de sens si la taxonomie n'est pas bien établie. C'est parfois une difficulté majeure quand on travaille sur les micro-organismes;
- Les classes sont équidistantes: la richesse augmente d'une unité quand on rajoute une espèce, que cette espèce soit proche des précédentes ou extrêmement originale.

L'indice de richesse le plus simple et le plus utilisé est tout simplement le nombre d'espèces S.

Équitabilité

La régularité de la distribution des espèces (équitabilité en Français, *evenness* ou *equitability* en anglais) est un élément important de la diversité. Une espèce représentée abondamment ou par un seul individu n'apporte pas la même contribution à l'écosystème. Sur la figure 1.1, la ligne du bas présente deux communautés de 4 espèces, mais celle de droite est beaucoup plus équitable de celle de gauche et semble intuitivement plus diverse. À nombre d'espèces égal, la présence d'espèces très dominantes entraîne mathématiquement la rareté de certaines autres: on comprend donc assez intuitivement que le maximum de diversité sera atteint quand les espèces auront une répartition très régulière.

Un indice d'équitabilité est indépendant du nombre d'espèces (donc de la richesse).

La plupart des mesures de diversité courantes, comme celle de Simpson ou de Shannon, évaluent à la fois la richesse et l'équitabilité.

Disparité

Les mesures classiques de la diversité, dites mesures de diversité neutre (*species-neutral diversity*) ou taxonomique ne prennent pas en compte une quelconque distance entre classes. Pourtant, deux espèces du même genre sont de toute évidence plus proches que deux espèces de familles différentes. Les mesures de diversité non neutres (chapitre **??**) prennent en compte cette notion, qui nécessite quelques définitions supplémentaires.[3]

La mesure de la différence entre deux classes est souvent une distance, mais parfois une mesure qui n'a pas toutes les propriétés d'une distance: une dissimilarité. Les mesures de *divergence*[4] sont construites à partir de la dissimilarité entre

[3] D. Mouillot et al. "Niche Overlap Estimates Based on Quantitative Functional Traits: A New Family of Non-Parametric Indices." In: *Oecologia* 145.3 (2005), pp. 345–353. DOI: 10.1007/s00442-005-0151-z; C. Ricotta. "A Semantic Taxonomy for Diversity Measures." In: *Acta Biotheoretica* 55.1 (2007), pp. 23–33. DOI: 10.1007/s10441-007-9008-7.

[4] S. Pavoine and M. B. Bonsall. "Measuring Biodiversity to Explain Community Assembly: A Unified Approach." In: *Biological Reviews* 86.4 (2011), pp. 792–812. DOI: 10.1111/j.1469-185X.2010.00171.x.

les classes, avec ou sans pondération par la fréquence.

Si la divergence entre espèces est une distance évolutive comme l'âge du plus récent ancêtre commun, la diversité sera dite phylogénétique. Si c'est une distance fonctionnelle, définie par exemple dans l'espace des traits fonctionnels, la diversité sera dite fonctionnelle.

La disparité,[5] divergence moyenne entre deux espèces (indépendamment des fréquences), ou de façon équivalente la longueur totale des branches d'un arbre phylogénétique, est la composante qui décrit à quel point les espèces sont différentes les unes des autres.

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

Les mesures de *régularité* décrivent la façon dont les espèces occupent l'espace des niches (régularité fonctionnelle) ou la régularité dans le temps et entre les clades des évènements de spéciation représentés par un arbre phylogénétique. Ce concept complète celui d'équitabilité dans les mesures classiques: la diversité augmente avec la richesse, la divergence entre espèces, et la régularité (qui se réduit à l'équitabilité quand toutes les espèces sont également divergentes entre elles).

Agrégation

À partir d'une large revue de la littérature dans plusieurs disciplines scientifiques s'intéressant à la diversité (au-delà de la biodiversité), Stirling[6] estime que les trois composantes, qu'il nomme *variété* (richesse), *équilibre* (équitabilité) et *disparité*, recouvrent tous les aspects de la diversité.

[6] A. Stirling. "A General Framework for Analysing Diversity in Science, Technology and Society." In: *Journal of the Royal Society, Interface* 4.15 (2007), pp. 707–719. DOI: 10.1098/rsif.2007.0213.

Stirling définit la propriété d'*agrégation* comme la capacité d'une mesure de diversité à combiner explicitement les trois composantes précédentes. Cela ne signifie pas que les composantes contribuent indépendamment les unes des autres à la diversité.[7]

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

1.2 Niveaux de l'étude

La diversité est classiquement estimée à plusieurs niveaux emboîtés, nommés α, β et γ par Whittaker[8] qui a nommé α la diversité locale qu'il mesurait avec l'indice α de Fisher (voir le chapitre **??**) et a utilisé les lettres suivantes selon ses besoins.

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

Diversité α, β et γ

La diversité α est la diversité locale, mesurée à l'intérieur d'un système délimité. Plus précisément, il s'agit de la diversité dans un habitat uniforme de taille fixe.

De façon générale,[9] la richesse spécifique diminue avec la latitude (la diversité est plus grande dans les zones tropicales, et au sein de celles-ci, quand on se rapproche de l'équateur),

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

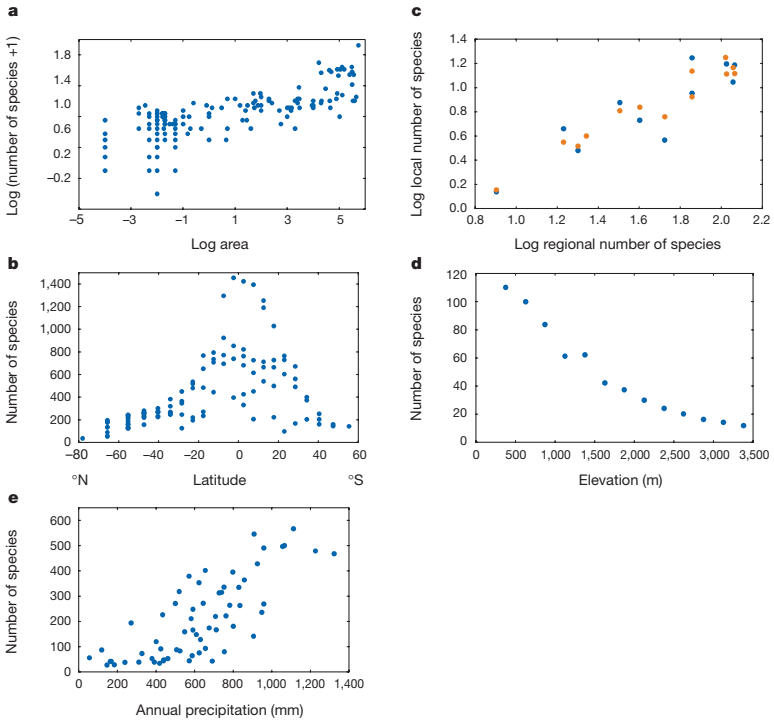


Figure 1.2: Patrons de biodiversité. (a) Le nombre d'espèces de vers de terre augmente en fonction de la surface échantillonnée, de 100 m² à plus de 500000 km² selon la relation d'Arrhenius). (b) Nombre d'espèces d'oiseaux en fonction de la latitude. (c) Relation entre la richesse régionale et la richesse locale. (d) Nombre d'espèces de chauves-souris en fonction de l'altitude dans une réserve au Pérou. (e) Nombre d'espèces de végétaux ligneux en fonction des précipitations en Afrique du Sud.

voir figure 1.2.[10] La tendance est la même pour la diversité génétique intraspécifique.[11] La richesse diminue avec l'altitude. Elle est généralement plus faible sur les îles, où elle décroît avec la distance au continent, source de migrations.

La diversité β mesure à quel point les systèmes locaux sont différents. Cette définition assez vague a fait l'objet de nombreux débats.[12]

Enfin, la diversité γ est similaire à la diversité α, prise en compte sur l'ensemble du système étudié. Les diversités α et γ se mesurent donc de la même façon, mais à différentes échelles.

Décomposition

Whittaker[13] a proposé sans succès une normalisation des échelles d'évaluation de la biodiversité, en introduisant la diversité régionale ε (γ étant réservé au paysage et α à l'habitat) et la diversité δ entre les paysages. Seuls les trois niveaux originaux ont été conservés par la littérature, sans définition stricte des échelles d'observation.

La distinction entre les diversités α et β dépend de la finesse de la définition de l'habitat. La distinction de nombreux habitats diminue la diversité α au profit de la β. Il est donc important de définir une mesure qui ne dépende pas de ce découpage, donc une mesure cumulative (additive ou multiplicative) décrivant la diversité totale, décomposable en la

[10] Gaston, "Global Patterns in Biodiversity," see n. 9, p. 5, figure 1.

[11] A. Miraldo et al. "An Anthropocene Map of Genetic Diversity." In: *Science* 353.6307 (2016), pp. 1532–1535. DOI: 10.1126/science.aaf4381.

[12] C. E. Moreno and P. Rodríguez. "A Consistent Terminology for Quantifying Species Diversity?" In: *Oecologia* 163.2 (2010), pp. 279–82. DOI: 10.1007/s00442-010-1591-7.

[13] R. H. Whittaker. "Evolution of Species Diversity in Land Communities." In: *Evolutionary Biology* 10 (1977). Ed. by M. K. Hecht et al., pp. 1–67.

somme ou le produit convenablement pondérés de toutes les diversités α des habitats (diversité intra) et de la diversité β inter-habitat.

Nous appellerons *communauté* le niveau de découpage concernant la diversité α et *méta-communauté* le niveau de regroupement pour l'estimation de la diversité γ.

1.3 Courbes d'accumulation

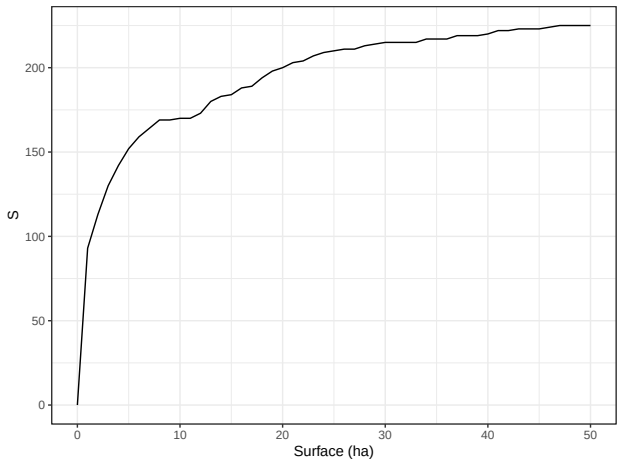


Figure 1.3: Courbe d'accumulation des espèces d'arbres du dispositif de Barro Colorado Island. Le nombre d'espèces est cumulé dans l'ordre des carrés d'un hectare du dispositif.

Evaluer la diversité d'une communauté nécessite en pratique de l'inventorier. Le nombre d'espèces découvertes en fonction de l'effort d'échantillonnage permet de tracer une courbe d'accumulation (SAC: *Species Acumulation Curve*). Une courbe de raréfaction (*Rarefaction Curve*) peut être calculée en réduisant par des outils statistiques l'effort d'échantillonnage réel pour obtenir une SAC théorique, libérée des aléas de l'ordre de prise en compte des données.

La figure 1.3 montre l'accumulation des espèces pour les données de BCI. Une SAC peut être tracée en fonction de la surface, du nombre d'individus ou du nombres de placettes d'échantillonnage, selon les besoins.

Code R pour réaliser la figure 1.3:

```
library("vegan")
data(BCI)
# Chaque parcelle (ligne) cumule ses abondances à la précédente
cumul <- apply(BCI, 2, cumsum)
# Le nombre d'espèces de chaque parcelle est cumulé
Richesse <- apply(cumul, 1, function(x) sum(x > 0))
ggplot(
  data.frame(
    A = 0:50,
    S = c(0, Richesse)
  )
) +
  geom_line(aes(x = A, y = S)) +
  labs(x = "Surface (ha)")
```

Les courbes d'accumulation peuvent aussi concerner la diversité (voir le chapitre **??**), mesurée au-delà du nombre d'espèces.

Plus généralement, une courbe aire-espèces (SAR: *Species Area Relationship*) représente le nombre d'espèces observées en fonction de la surface échantillonnée (figure **??**). Il existe plusieurs façons de prendre en compte cette relation,[14] classables en deux grandes familles:[15]

- Dans une SAR au sens strict, chaque point représente une communauté. La question traitée est la relation entre le nombre d'espèces et la taille de chaque communauté, en lien avec des processus écologiques;
- Une courbe d'accumulation (SAC) ne représente que l'effet statistique de l'échantillonnage. Pour éviter toute confusion, le terme SAR ne doit pas être utilisé pour décrire une SAC.

[14] S. M. Scheiner. "Six Types of Species-Area Curves." In: *Global Ecology and Biogeography* 12.6 (2003), pp. 441–447. DOI: 10.1046/j.1466-822X.2003.00061.x.

[15] J. Dengler. "Which Function Describes the Species-Area Relationship Best? A Review and Empirical Evaluation." In: *Journal of Biogeography* 36.4 (2009), pp. 728–744. DOI: 10.1111/j.1365-2699.2008.02038.x.

1.4 Diversité asymptotique

Augmenter l'effort d'échantillonnage peut permettre d'atteindre un stade où la diversité n'augmente plus: sa valeur est appelée *diversité asymptotique*. Dans des communautés très diverses comme les forêts tropicales, la diversité asymptotique ne peut en général pas être observée sur le terrain à cause de la variabilité de l'environnement: l'augmentation de la surface inventoriée amène à échantillonner dans des communautés différentes avant d'atteindre la diversité asymptotique de la communauté étudiée. La diversité asymptotique est donc celle d'une communauté théorique qui n'existe généralement pas. En d'autres termes, c'est la diversité d'une communauté dont l'inventaire disponible serait un échantillon représentatif.

Evaluer la diversité asymptotique nécessite d'utiliser des estimateurs de diversité, dont la précision dépend de l'exhaustivité de l'échantillonnage. La diversité peut aussi être estimée pour un effort donné: un hectare de forêt ou 5000 arbres par exemple, ou encore un taux de couverture choisi, qui décrit mieux la qualité de l'échantillonnage.

1.5 Couverture

Le taux de couverture (*sample coverage*) de l'échantillon est la proportion des espèces découvertes,

$$C = \sum_{s=1}^{S} \mathbf{1}\left(n_s > 0\right) p_s, \tag{1.1}$$

où $\mathbf{1}(\cdot)$ est la fonction indicatrice. Son complément à 1 est appelé déficit de couverture (*coverage deficit*).

Le déficit de couverture est la probabilité qu'un individu tiré au hasard dans la communauté appartienne à une espèce absente de l'échantillon inventorié. C'est donc aussi la probabilité qu'un individu ajouté à l'inventaire lui ajoute une nouvelle espèce.[16] La pente de la courbe d'accumulation donnant l'espérance du nombre d'espèces en fonction du nombre d'individus (courbe de raréfaction de la figure **??**) est donc égale au déficit de couverture:[17]

$$1 - \mathbb{E}\left[C\left(n\right)\right] = \mathbb{E}\left[S\left(n+1\right)\right] - \mathbb{E}\left[S\left(n\right)\right], \tag{1.2}$$

où $C(n)$ est le taux de couverture d'un échantillon de taille n et $S(n)$ le nombre d'espèces découvertes dans cet échantillon.

Le taux de couverture augmente avec l'effort d'échantillonnage. Plus il est élevé, meilleures seront les estimations de la diversité: la diversité asymptotique est obtenue quand il atteint 1. Les estimateurs de la diversité asymptotique développés plus loin reposent largement sur cette notion pour la correction du biais d'échantillonnage,[18] c'est-à-dire la sous-estimation systématique de la diversité due aux espèces non observées, qui est un des éléments du biais d'estimation.

Pour comparer la diversité non asymptotique de deux communautés avec le même effort d'échantillonnage, Chao and Jost[19] montrent que le taux de couverture plutôt que la taille de l'échantillon doit être identique.

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

[17] A. Chao and L. Jost. "Coverage-Based Rarefaction and Extrapolation: Standardizing Samples by Completeness Rather than Size." In: *Ecology* 93.12 (2012), pp. 2533–2547. DOI: 10.1890/11-1952.1.

[18] G. Dauby and O. J. Hardy. "Sampled-Based Estimation of Diversity Sensu Stricto by Transforming Hurlbert Diversities into Effective Number of Species." In: *Ecography* 35.7 (2012), pp. 661–672. DOI: 10.1111/j.1600-0587.2011.06860.x.

[19] Chao and Jost, see n. 17.

Formule des fréquences de Good-Turing

La relation fondamentale entre les fréquences des espèces dans un échantillon est due à Turing et a été publiée par Good.[20] En absence de toute information sur la loi de distribution des espèces, en supposant seulement que les individus sont tirés indépendamment les uns des autres selon une loi multinomiale respectant ces probabilités, la formule de Good-Turing relie la probabilité attendue $\mathbb{E}(p_\nu)$ d'une espèce représentée par ν individus au rapport entre les nombres d'espèces représentées $\nu+1$ fois et ν fois:

$$\mathbb{E}\left(p_\nu\right) \approx \frac{(\nu+1)}{n}\frac{\mathbb{E}\left(f_{\nu+1}\right)}{\mathbb{E}\left(f_\nu\right)}. \tag{1.3}$$

La variance de p_ν est connue:

$$\mathrm{Var}\left(p_\nu\right) \approx \mathbb{E}\left(p_\nu\right)\left[\mathbb{E}\left(p_{\nu+1}\right) - \mathbb{E}\left(p_\nu\right)\right]. \tag{1.4}$$

Elle est petite en comparaison de l'espérance.

[20] Good, see n. 16.

Le nombres d'espèces observées ν et $\nu+1$ fois varient selon l'échantillonnage. La relation de Good-Turing concerne leur espérance mais comme on ne dispose en général que d'un seul inventaire, les espérances $\mathbb{E}(f_\nu)$ et $\mathbb{E}(f_{\nu+1})$ sont remplacées par les valeurs observées. De même, chacune des espèces observées ν fois a une probabilité différente: la relation ne permet pas de prédire précisément, pour un échantillon, les probabilités de chaque espèce.

Ces relations sont le fondement de plusieurs estimateurs de biodiversité présentés plus loin. Les singletons (f_1: le nombre d'espèces observées une seule fois) et les doubletons (f_2: le nombre d'espèces observées deux fois) ont une importance centrale. Pour $\nu = 1$, on a par exemple $\alpha_1 = 2f_2/(nf_1)$: la fréquence d'une espèce typiquement représentée par un singleton est proportionnelle au rapport entre le nombre des doubletons et des singletons. Cet estimateur de probabilité est meilleur que l'estimateur naïf $1/n$: en d'autres termes, la distribution des fréquences observées permet d'estimer les probabilités de façon non triviale.

La relation a été précisée[21] en limitant les approximations dans les calculs. La seule approximation nécessaire est que les probabilités des espèces représentées le même nombre de fois ν varient peu et puissent être considérées toutes égales à $\mathbb{E}(p_\nu)$, ce qui est acceptable puisque la variance de p_ν est petite. On peut alors estimer directement

$$\hat{p_\nu} = \frac{(\nu+1)\, f_{\nu+1}}{(n-\nu)\, f_\nu + (\nu+1)\, f_{\nu+1}} \tag{1.5}$$

en remplaçant les espérances par les valeurs observées.

Ce nouvel estimateur est à la base de l'estimateur de Chao amélioré et des estimateurs d'entropie de Chao et Jost (sections 3.3 et **??**).

[21] C.-H. Chiu et al. "An Improved Nonparametric Lower Bound of Species Richness via a Modified Good-Turing Frequency Formula." In: *Biometrics* 70.3 (2014), pp. 671–682. DOI: 10.1111/biom.12200. PMID: 24945937, eq. 6 et 7a.

Estimation du taux de couverture

En posant $\nu = 0$ dans l'équation (1.3), $\mathbb{E}(p_0) \times f_0 = \pi_0$, la probabilité totale des espèces non représentées, vaut approximativement f_1/n. C'est l'estimateur de Good ou Good-Turing, parfois appelé abusivement "formule de Turing":[22]

$$\hat{C} = 1 - \frac{f_1}{n}. \tag{1.6}$$

Cet estimateur est biaisé.[23] En réalité,

$$C = 1 - \frac{\mathbb{E}(f_1) - \pi_1}{n}. \tag{1.7}$$

L'estimateur de Good néglige le terme π_1, la somme des probabilités des espèces observées une fois. Ce terme peut être

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

[23] Ibid.

[24] A. Chao et al. "A Generalized Good's Nonparametric Coverage Estimator." In: *Chinese Journal of Mathematics* 16 (1988), pp. 189–199. JSTOR: 43836340.

estimé avec un biais plus petit. Chao et al.[24] puis Z. Zhang and Huang[25] proposent l'estimateur suivant, qui utilise toute l'information disponible et a le plus petit biais possible:

$$\hat{C} = 1 - \sum_{\nu=1}^{n} (-1)^{\nu+1} \binom{n}{\nu}^{-1} f_\nu. \tag{1.8}$$

Les termes de la somme décroissent très vite avec ν. En se limitant à $\nu = 1$, l'estimateur se réduit à celui de Good.

Esty,[26] complété par C.-H. Zhang and Z. Zhang,[27] a montré que l'estimateur était asymptotiquement normal et a calculé l'intervalle de confiance de $\hat{C}$:

$$C = \hat{C} \pm t^n_{1-\alpha/2} \frac{\sqrt{f_1 \left(1 - \frac{f_1}{n}\right) + 2f_2}}{n}. \tag{1.9}$$

Où $t^n_{1-\alpha/2}$ est le quantile d'une loi de Student à n degrés de libertés au seuil de risque α, classiquement 1,96 pour n grand et $\alpha = 5\%$.

Un autre estimateur est utilisé dans le logiciel SPADE[28] et son portage sous R, le package *spadeR*.[29] Il est la base des estimateurs d'entropie de Chao et Jost (section **??**). L'estimation de l'équation (1.7) donne la relation

$$\hat{C} = 1 - \frac{f_1 - \hat{\pi}_1}{n}. \tag{1.10}$$

Or, $\hat{\pi}_1 = f_1 \hat{p}_1$. p_1 peut être estimé par la relation de Good-Turing (1.5), en remplaçant f_0 par l'estimateur Chao1 (3.5). Alors:

$$\hat{C} = 1 - \frac{f_1}{n}(1 - \hat{p}_1) = 1 - \frac{f_1}{n}\left[\frac{(n-1)\, f_1}{(n-1)\, f_1 + 2f_2}\right]. \tag{1.11}$$

Dans le package *divent*, la fonction `coverage` calcule les trois estimateurs (celui de Zhang et Huang par défaut):

```
library("divent")
BCI %>%
  colSums() %>%
  coverage()
```

```
## # A tibble: 1 x 2
##   estimator  coverage
##   <chr>         <dbl>
## 1 ZhangHuang    0.999
```

Le taux de couverture de BCI est proche de 1 parce que l'inventaire couvre 50 ha. Il est moindre sur les 6.25 ha de la parcelle 6 de Paracou:

[25] Z. Zhang and Huang, see n. 22.

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

[27] C.-H. Zhang and Z. Zhang. "Asymptotic Normality of a Nonparametric Estimator of Sample Coverage." In: *Annals of Statistics* 37 (5A 2009), pp. 2582–2595. DOI: 10.1214/08-aos658.

[28] A. Chao and T.-J. Shen. *Program SPADE: Species Prediction and Diversity Estimation. Program and User's Guide.* CARE, 2010.

[29] A. Chao et al. "SpadeR: Species Prediction and Diversity Estimation with R." in: (2016).

```
paracou_6_abd %>%
  colSums() %>%
  coverage()
```

```
## # A tibble: 1 x 2
##   estimator  coverage
##   <chr>         <dbl>
## 1 ZhangHuang    0.972
```

Les estimateurs présentés ici supposent une population de taille infinie (de façon équivalente, les individus sont tirés avec remise). Le cas des populations de taille finie est traité par Chao and Lin[30] et Hwang et al.[31]

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

[31] W.-H. Hwang et al. "Good-Turing Frequency Estimation in a Finite Population." In: *Biometrical journal* 57.2 (2014), pp. 321–339. DOI: 10.1002/bimj.201300168.

Complétude

La complétude de l'échantillonnage est la proportion du nombre d'espèces observées: $f_{>0}/S$. Elle compte simplement les espèces et ne doit pas être confondue avec la couverture qui somme leurs probabilités: le taux de complétude est toujours très inférieur au taux de couverture parce que les espèces non échantillonnées sont les plus rares.

La complétude du même échantillon d'arbres de forêt tropicale que dans l'exemple précédent peut être estimée en divisant le nombre d'espèces observées par le nombre d'espèces estimées (voir section 3.1). À BCI:

```
bci_abd <- colSums(BCI)
# Espèces observées
(obs <- div_richness(bci_abd, estimator  = "naive"))
```

```
## # A tibble: 1 x 3
##   estimator order diversity
##   <chr>     <dbl>     <int>
## 1 naive         0       225
```

```
# Richesse estimée
(est <- div_richness(bci_abd))
```

```
## # A tibble: 1 x 3
##   estimator   order diversity
##   <chr>       <dbl>     <dbl>
## 1 Jackknife 1     0       244
```

```
# Complétude
obs$diversity / est$diversity
```

```
## [1] 0.9221311
```

À Paracou:

```
# Espèces observées
(obs <- div_richness(colSums(paracou_6_abd), estimator  = "naive"))
```

```
## # A tibble: 1 x 3
##   estimator order diversity
##   <chr>     <dbl>     <int>
## 1 naive         0       335
```

```
# Richesse estimée
(est <- div_richness(colSums(paracou_6_abd)))
```

```
## # A tibble: 1 x 3
##   estimator   order diversity
##   <chr>       <dbl>     <dbl>
## 1 Jackknife 2     0       473
```

```
# Complétude
obs$diversity / est$diversity
```

```
## [1] 0.7082452
```

1.6 Le problème de l'espèce

Évaluer la richesse spécifique suppose que les espèces soient définies clairement, ce qui n'est de toute évidence pas le cas.[32] Le premier aspect du problème concerne la nature des espèces: réalité naturelle ou seulement représentation simplificatrice. Une analyse historique et philosophique en est faite par Richards.[33] L'autre aspect, avec des conséquences pratiques plus immédiates, concerne leur délimitation. Mayden[34] recense vingt-deux définitions différentes du concept d'espèce. **Wilkins2011<empty citation>**[35] estime qu'il n'y a qu'un seul concept d'espèce mais sept définitions, c'est-à-dire sept façons de les identifier, et vingt-sept variations ou mélanges de ces définitions.

La définition historique est celle de *morphoespèce*, qui classe les espèces selon leurs formes, supposées d'abord immuables. La définition moderne la plus répandue est celle d'espèce *biologique*:[36] un "groupe de populations naturelles isolées reproductivement les unes des autres".[37] Lorsque les populations d'une espèce sont isolées géographiquement, leur capacité à se reproduire ensemble est toute théorique (et rarement vérifiée expérimentalement). Des populations allopatriques n'ont pas de flux de gènes réels entre elles et peuvent être considérées comme des espèces distinctes selon la définition d'espèce *phylogénétique*: "le plus petit groupe identifiable d'individus avec un pattern commun d'ancêtres et de descendants".[38] C'est l'unité génétique détectée par la méthode du coalescent pour la délimitation des espèces.[39] Le nombre d'espèces phylogénétiques est bien supérieur au nombre d'espèces biologiques. Enfin, Van Valen[40] définit les espèces par la niche écologique qu'elles occupent (à partir de l'exemple des chênes blancs européens) plutôt que par les flux de gènes (permanents entre les espèces distinctes):

[32] E. Casetta. "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, 2014, pp. 139–154.

[33] R. A. Richards. *The Species Problem. A Philosophical Analysis*. Cambridge: Cambridge University Press, 2010.

[34] R. L. Mayden. "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, 1997, pp. 381–424.

[35] **<empty citation>**.

[36] T. Dobzhansky. *Genetics and the Origin of Species*. New York: Columbia University Press, 1937.

[37] E. Mayr. *Systematics and the Origin of Species from the Viewpoint of a Zoologist*. New York: Columbia University Press, 1942.

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

[39] J. Sukumaran and L. L. Knowles. "Multispecies Coalescent Delimits Structure, Not Species." In: *Proceedings of the National Academy of Sciences of the United States of America* in press (2017). DOI: 10.1073/PNAS.1607921114.

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

la définition *écologique* d'espèce est proche du concept de complexe d'espèces.[41]

Le choix de la définition modifie considérablement sur la quantification de la richesse.[42] Des problèmes méthodologiques s'ajoutent aux problèmes conceptuels:[43] la séparation ou le regroupement de plusieurs populations ou morphotypes en un nombre plus ou moins grand d'espèces est un choix qui reflète les connaissances du moment et peut évoluer.[44]

L'impact du problème de l'espèce sur la mesure de la diversité reste sans solution à ce stade, si ce n'est d'utiliser les mêmes définitions si des communautés différentes doivent être comparées. L'approche phylogénétique (chapitre **??**) permet de contourner le problème: si deux taxons très semblables apportent à peine plus de diversité qu'un seul taxon, le choix de les distinguer ou non n'est pas critique.

[41] ensemble d'espèces voisines échangeant des gènes, J. Pernès, ed. *Gestion Des Ressources Génétiques Des Plantes. Tome 2 : Manuel.* Paris: Agence de Coopération culturelle et technique, 1984.

[42] P. M. Agapow et al. "The Impact of Species Concept on Biodiversity Studies." In: *The Quarterly Review of Biology* 79.2 (2004), pp. 161–179. DOI: 10.1086/383542.

[43] Hey, "The Mind of the Species Problem," see n. 8, p. vi.

[44] A. Barberousse and S. Samadi. "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, 2014, pp. 155–182.

Chapter 2

Distribution de l'abondance des espèces (SAD)

La distribution de l'abondance des espèces (SAD: *Species Abundance Distribution*) est la loi statistique qui donne l'abondance attendue de chaque espèce d'une communauté. Les espèces ne sont pas identifiées individuellement, mais par le nombre d'individus leur appartenant.

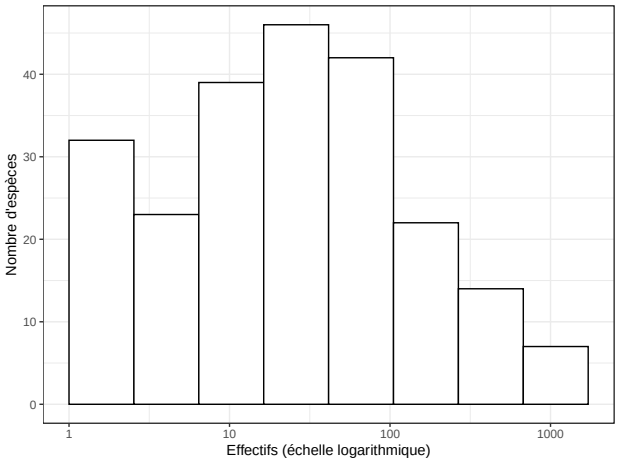


Figure 2.1: Histogramme des fréquences (diagramme de Preston) des arbres du dispositif de Barro Colorado Island. En abscisse: le nombre d'arbres de chaque espèce (en logarithme); en ordonnée: le nombre d'espèces.

Elle peut être représentée sous la forme d'un histogramme des fréquences (diagramme de Preston,[1] figure 2.1) ou bien d'un diagramme rang-abondance (RAC: *Rank Abundance Curve* ou diagramme de Whittaker,[2] figure 2.2). Le RAC est souvent utilisé pour reconnaître des distributions connues. Izsák and Pavoine[3] ont étudié les propriétés des RAC pour les principales SAD.

Code de la figure 2.1:

```
BCI_abd <- sort(colSums(BCI), decreasing = TRUE)
ggplot(data.frame(BCI_abd), aes(BCI_abd)) +
  geom_histogram(
    bins = nclass.Sturges(log(BCI_abd)),
    color = "black",
```

[1] F. W. Preston. "The Commonness, and Rarity, of Species." In: *Ecology* 29.3 (1948), pp. 254–283. DOI: 10.2307/1930989.

[2] Whittaker, "Dominance and Diversity in Land Plant Communities," see n. 1, p. 3.

[3] J. Izsák and S. Pavoine. "Links between the Species Abundance Distribution and the Shape of the Corresponding Rank Abundance Curve." In: *Ecological Indicators* 14.1 (2012), pp. 1–6. DOI: 10.1016/j.ecolind.2011.06.030.

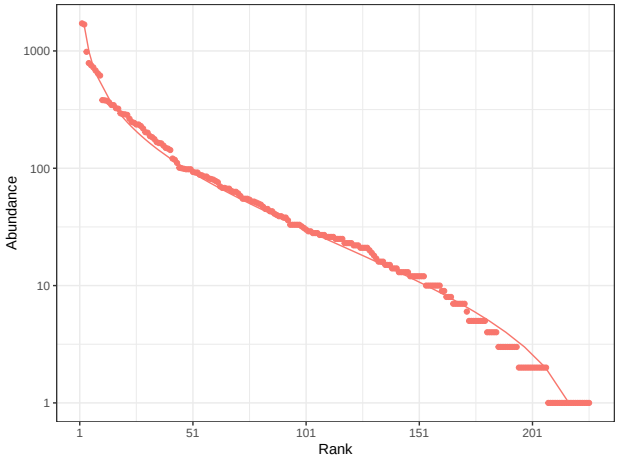


Figure 2.2: Diagramme rang-abondance (diagramme de Whittaker) des arbres du dispositif de Barro Colorado Island. Les points sont les données: en abscisse: le rang de l'espèce, à partir de la plus abondante; en ordonnée: l'abondance de l'espèce. La courbe est l'ajustement d'une distribution log-normale.

```
    fill = "white",
    boundary = 0
  ) +
  scale_x_log10() +
  labs(
    x = "Effectifs (échelle logarithmique)",
    y = "Nombre d'espèces"
  )
```

Code de la figure 2.2:

```
library("divent")
BCI_abd %>%
  as_abundances() %>%
  autoplot(fit_rac = TRUE, distribution = "lnorm")
```

Les SAD sont traitées en détail par Magurran[4] ou McGill et al.[5] Les principales distributions, nécessaires à la compréhension de la suite sont présentées ici:

- La distribution en log-séries de Fisher et al.;[6]
- La distribution géométrique;[7]
- La distribution log-normale;[8]
- Le modèle Broken Stick.[9]

Formellement, la distribution des probabilités des espèces, notées p_s, est à établir.

2.1 La distribution en log-séries

Cette distribution est traitée en détail dans le chapitre **??**.

Le nombre d'espèces est lié au nombre d'individus par la relation $\mathbb{E}(S^n) = \alpha \ln(1 + n/\alpha)$ où S^n indique le nombre d'espèces observées dans un échantillon de n individus. α est le paramètre qui fixe la pente de la partie linéaire de la relation, valide dès que $n \gg \alpha$, où le nombre d'espèces

[4] A. E. Magurran. *Ecological Diversity and Its Measurement*. Princeton, NJ: Princeton University Press, 1988.

[5] B. J. McGill et al. "Species Abundance Distributions: Moving beyond Single Prediction Theories to Integration within an Ecological Framework." In: *Ecology Letters* 10.10 (2007), pp. 995–1015. DOI: 10.1111/j.1461-0248.2007.01094.x.

[6] R. A. Fisher et al. "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 (1943), pp. 42–58. DOI: 10.2307/1411.

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

[8] Preston, "The Commonness, and Rarity, of Species," see n. 1, p. 15.

[9] R. H. MacArthur. "On the Relative Abundance of Bird Species." In: *Proceedings of the National Academy of Sciences of the United States of America* 43.3 (1957), pp. 293–295. DOI: 10.1073/pnas.43.3.293. JSTOR: 89566.

S^n augmente proportionnellement au logarithme du nombre d'individus $\ln(n)$.

La distribution a été obtenue à partir d'inventaires de communautés de papillons en Malaisie et en Angleterre. Le modèle est tombé en désuétude faute de confirmation empirique à l'échelle de la communauté, avant d'être remis en valeur par la théorie neutre[10] dans lequel la distribution de la *méta-communauté* est en log-séries.

[10] S. P. Hubbell. *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, 2001.

2.2 La distribution Broken Stick

Si les espèces se partagent les ressources ou l'espace des niches, représentées par un bâton, par un processus de cassure aléatoire et simultanée (précisément, les $S-1$ cassures du bâton sont distribuées uniformément sur sa longueur) et que leur abondance est proportionnelle à la quantité de ressources ou d'espace de niche obtenus, alors leur distribution suit le modèle Broken Stick de MacArthur.[11]

[11] MacArthur, see n. 9.

Parmi les distributions classiques, c'est la plus équitable: la distribution uniforme des probabilités ($p_s = 1/S$ pour tout s) n'est jamais approchée.

Elle est peu observée empiriquement.

2.3 La distribution log-normale

Dans une distribution log-normale, le logarithme des probabilités des espèces (notées p_s pour l'espèce s) suit une loi normale. L'écart-type σ de cette distribution règle l'équitabilité de la distribution. Son espérance est obtenue à partir du nombre d'espèces et de σ, pour que la somme des probabilités égale 1.

May[12] explique cette distribution par le théorème de la limite centrale: la variable aléatoire valant 1 si un individu est de l'espèce s et 0 sinon est le produit de nombreuses variables de loi inconnues valant 1 en cas de succès (germination d'une graine, survie à de nombreux évènements…). Le logarithme de ce produit est une somme de variables aléatoire dont la loi est forcément normale par application du théorème de la limite centrale.

[12] R. M. May. "Patterns of Species Abundance and Diversity." In: *Ecology and Evolution of Communities*. Ed. by M. L. Cody and J. M. Diamond. Harvard University Press, 1975, pp. 81–120.

La distribution est aussi le résultat d'un algorithme de bâton brisé (*broken stick*) hiérarchique:[13]

[13] M. G. Bulmer. "On Fitting the Poisson Lognormal Distribution to Species-Abundance Data." In: *Biometrics* 30.1 (1974), pp. 101–110. DOI: 10.2307/1939021.

- Si les ressources (représentées par un bâton) sont partagées une première fois aléatoirement, selon une loi quelconque,
- Si chaque bâton obtenu est partagé à nouveau selon le même procédé, et que l'opération est répétée un assez grand nombre de fois,

- Si l'abondance de chaque espèce est proportionnelle aux ressources dont elle dispose,
- Alors la distribution des espèces est log-normale.

Ce mécanisme décrit assez bien un mécanisme de partage successif des ressources, par exemple entre groupes d'espèces de plus en plus petits, correspondant à des niches écologiques de plus en plus étroites.

D'autres arguments existent dans la littérature. Par exemple, Engen and Lande[14] obtiennent une distribution normale à partir d'un modèle de dynamique des populations.

La distribution log-normale décrit assez bien (mais pas exactement) une communauté locale dans le cadre de la théorie neutre[15] comme le montre la figure 2.2. Le nombre d'espèces rares est un peu surestimé. La distribution exacte est donnée par Volkov et al.[16]

2.4 La distribution géométrique

Si les espèces se partagent les ressources selon un algorithme *broken stick* séquentiel (comme dans la distribution log-normale) mais de proportion fixe $0 < k < 1$, alors la distribution obtenue est géométrique. Les abondances successives sont proportionnelles à $k, k(1-k), k(1-k)^2, \ldots, k(1-k)^s, \ldots, k(1-k)^S$.

Ce modèle a été établi par Motomura[17] cité par May.[18] Ses propriétés ont été étudiées par Whittaker.[19]

C'est la distribution qui traduit l'absence de relation entre la taille de l'échantillon et l'abondance des espèces:[20] la distribution du logarithme de ses probabilités est uniforme. En d'autre termes, l'ordre de grandeur de l'abondance d'une espèce est uniformément distribué.

La distribution est observée dans les communautés pionnières,[21] peu diverses, ou les communautés microbiennes.[22]

2.5 Synthèse

La figure 2.3 est inspirée de la figure très connue de Magurran.[23] Elle montre bien une gradation en termes de décroissance de probabilité entre des distributions de même richesse: de la plus équitable (broken stick) à la plus inéquitable (géométrique). Elle doit être nuancée: la proportion k de la distribution géométrique fixe la pente de la droite qui la représente sur la figure. $k = 10\%$ sur la figure: une valeur plus faible diminuerait la pente. De même, l'écart-type de la distribution log-normale décrit sa dispersion. Une valeur supérieure augmenterait sa décroissance.

[14] S. Engen and R. Lande. "Population Dynamic Models Generating the Lognormal Species Abundance Distribution." In: *Mathematical Biosciences* 132.2 (1996), pp. 169–183. DOI: 10.1016/0025-5564(95)00054-2.

[15] Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography*, see n. 10, p. 17.

[16] I. Volkov et al. "Neutral Theory and Relative Species Abundance in Ecology." In: *Nature* 424.6952 (2003), pp. 1035–1037. DOI: 10.1038/nature01883.

[17] Motomura, "On the statistical treatment of communities," see n. 7, p. 16.

[18] May, "Patterns of Species Abundance and Diversity," see n. 12, p. 17.

[19] Whittaker, "Evolution and Measurement of Species Diversity," see n. 7, p. 16.

[20] S. Pueyo et al. "The Maximum Entropy Formalism and the Idiosyncratic Theory of Biodiversity." In: *Ecology letters* 10.11 (2007), pp. 1017–28. DOI: 10.1111/j.1461-0248.2007.01096.x.

[21] **Bazzaz1975**.

[22] B. Haegeman et al. "Robust Estimation of Microbial Diversity in Theory and in Practice." In: *The ISME journal* 7.6 (2013), pp. 1092–101. DOI: 10.1038/ismej.2013.10.

[23] Magurran, *Ecological Diversity and Its Measurement*, see n. 4, p. 16.

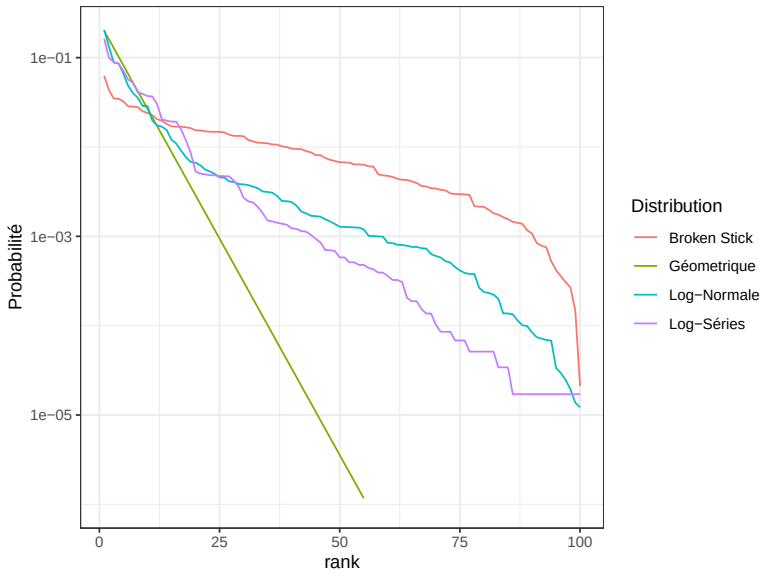


Figure 2.3: Diagramme rang-fréquence des distributions de probabilité classiques. Toutes les distributions sont de 100 espèces. Les probabilités inférieures à 10^{-6} ne sont pas affichées. Les paramètres choisis sont $\alpha = 11$ pour la distribution log-séries, $k = 0,2$ pour la distribution géométrique et $\sigma = 2$ pour la distribution log-normale.

Le code utilisé pour produire la figure 2.3 est le suivant:

```
library("divent")
# Tirage d'une communauté en log-séries
size <- 1E5
alpha <- 11
species_number <- -alpha * log(alpha / (size + alpha))
abd_lseries <- rlseries(species_number, size, alpha)
# Part des ressources accaparées dans la distribution géométrique
prob <- 0.2
# Calcul des probabilités de la distribution géométrique
prob_geom <- prob / (1 - (1 - prob)^species_number) * (1 - prob)^(0:(species_number - 1))
# Dispersion de la loi lognormale
sd <- 2
# Tirage de S valeurs dans une loi lognormale
abd_lnorm <- rlnorm(species_number, meanlog = 0, sdlog = sd)
# Tirage des probabilités de la distribution broken stick
prob_bstick <- c(cuts <- sort(stats::runif(species_number - 1)), 1) - c(0, cuts)
# Graphique
tibble(
  rank = 1:species_number,
  `Log-Séries` = sort(abd_lseries / sum(abd_lseries), decreasing = TRUE),
  `Géometrique` = sort(prob_geom, decreasing = TRUE),
  `Log-Normale` = sort(abd_lnorm / sum(abd_lnorm), decreasing = TRUE),
  `Broken Stick` = sort(prob_bstick, decreasing = TRUE)) %>%
  pivot_longer(cols = -rank) %>%
  ggplot() +
    geom_line(aes(x = rank, y = value, color = name)) +
    scale_y_log10(limits = c(1E-6, NA)) +
    labs(y = "Probabilité", color = "Distribution")
```

La simulation de ces quatre distributions peut être réalisée par la fonction `rcommunity()` du package *divent*, où l'argument `distribution` peut valoir “bstick”, “lnorm”, “geom” ou “lseries”. La simulation des communautés autres que log-séries passe par le tirage des probabilités des espèces (le calcul est déterministe dans le cas de la distribution géométrique) puis le tirage d'un nombre d'individus dans une loi multinomiale respectant ces probabilités et l'effectif total.

La fonction `fit_rac()` permet d'inférer les paramètres d'une distribution à partir d'un vecteur d'abondance. La distribution correspondant au modèle estimé peut être affichée sur la figure Rang-Abondance (figure 2.1).

Le package *sads* fournit les fonctions classiques de R (densité de probabilité, cumulative, quantile, simulation) pour les distributions utiles en écologie, au-delà de celles présentées ici. La distribution de Volkov notamment peut être simulée. Les fonctions `fitxxx()` complètent la fonction `fit_rac()` de *divent*. Enfin, la fonction `radfit()` du package *vegan* ajuste aux données en même temps les distributions broken-stick (désignée par "Null"), géométrique ("Preemption") et lognormale, inclut les distributions de Zipf et de Mandelbrot non traitées ici, mais ignore les log-séries. Les vraisemblances des différents modèles sont comparées pour choisir celui qui s'ajuste le mieux.

Le code suivant montre comment ajuster une distribution log-normale aux données de BCI avec *divent* ou *sads*.

```
# divent
library("divent")
fit_divent_lnorm <- fit_rac(BCI_abd, distribution = "lnorm")
# Affichage des paramètres estimés
fit_divent_lnorm$parameters
```

```
## # A tibble: 1 x 2
##      mu sigma
##   <dbl> <dbl>
## 1  3.14  1.79
```

```
# sads
library("sads")
# Estimation. Les données sont tronquées: les espèces observées 0 fois ne sont pas comptées
fit_sads_lnorm <- fitlnorm(BCI_abd, trunc=0)
fit_sads_lnorm@fullcoef
```

```
##  meanlog    sdlog
## 3.142695 1.787195
```

```
#vegan
library("vegan")
fit_vegan_lnorm <- radfit(BCI_abd)
fit_vegan_lnorm
```

```
##
## RAD models, family poisson
## No. of species 225, total abundance 21457
##
##             par1      par2     par3    Deviance AIC
## Null                                   10261.14 11387.97
## Preemption  0.034063                    3788.38  4917.21
## Lognormal   3.3569     1.5738            744.30  1875.13
## Zipf        0.14679  -0.94912          4335.50  5466.33
## Mandelbrot  17.014   -2.0064   15.048   988.02  2120.85
##             BIC
## Null        11387.97
## Preemption  4920.63
## Lognormal   1881.96
```

```
## Zipf        5473.16
## Mandelbrot  2131.10
```

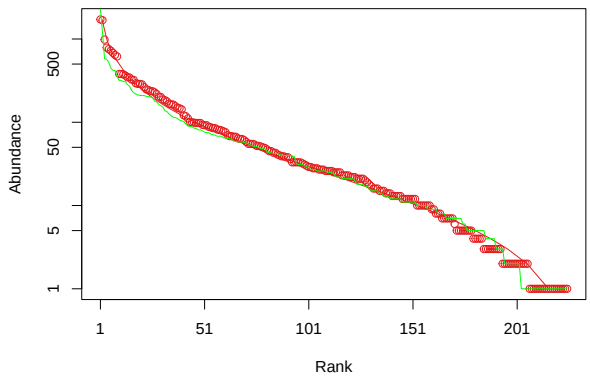
L'ajustement du modèle de Volkov peut être comparé à celui d'une distribution log-normale.

```
# Ajustement du modèle de Volkov
fit_volkov <- fitvolkov(BCI_abd)
fit_volkov@fullcoef
```

```
##        theta            m            J
## 4.796210e+01 9.180496e-02 2.145700e+04
```

Graphiquement, l'ajustement est très proche mais la distribution de Volkov prévoit explicitement des effectifs égaux parce qu'entiers.

```
# Comparaison graphique des deux modèles. Log-normal en rouge.
plot(as_abundances(BCI_abd), fit_rac = TRUE, distribution="lnorm")
# Volkov en vert
lines(
  sort(
    rvolkov(
      length(BCI_abd),
      fit_volkov@fullcoef[1],
      fit_volkov@fullcoef[2],
      fit_volkov@fullcoef[3]
    ),
    decreasing=TRUE
  ),
  col="green"
)
```



Les vraisemblances sont proches.

```
# Comparaison des vraisemblances
fit_sads_lnorm@min
```

```
## [1] 1157.013
```

```
fit_volkov@min
```

```
## [1] 1150.182
```

Les paramètres du modèle de communauté locale de la théorie neutre sont θ, le "nombre fondamental de la biodiversité" égal à deux fois le nombre d'espèces apparaissant par pas de temps dans la méta-communauté, m, le taux de migration, et J, la taille de la communauté locale (qui n'est pas à proprement parler un paramètre mais une statistique décrivant les données).

La différence entre les logarithmes de vraisemblance des deux modèles en faveur du modèle de Volkov, alors que le nombre de paramètres est le même. L'ajustement est donc meilleur mais la différence est petite et la complexité du modèle et des calculs pour l'estimer ne se justifient pas en général: le modèle de Volkov est très peu utilisé en pratique.

Part II

Diversité neutre d'une communauté

Chapter 3

Mesures classiques de la diversité α ou γ

L'essentiel

Les indices classiques de diversité sont ceux de Shannon et de Simpson, et la richesse spécifique. Ils peuvent être estimés à partir des données d'inventaire. L'estimation de la richesse est particulièrement difficile et fait l'objet d'une abondante littérature: les estimateurs non-paramétriques (Chao et Jackknife) sont les plus utilisés.

Les mesures classiques[1] considèrent que chaque classe d'objets est différente de toutes les autres, sans que certaines soient plus ou moins semblables. Dans ce chapitre, les classes seront des espèces. Les mesures sont qualifiées de neutres (*species-neutral*) au sens où elles ne prennent en compte aucune caractéristique propre des espèces. La diversité neutre est souvent appelée diversité taxonomique,[2] même si le terme peut prêter à confusion avec la diversité phylogénétique, quand la phylogénie se réduit à une taxonomie.[3]

Ce type de mesure n'a de sens qu'à l'intérieur d'un taxocène bien défini: sommer un nombre d'espèces d'insectes à un nombre d'espèces de mammifères a peu d'intérêt. Ces méthodes ne sont donc pas forcément les plus adaptées à la conservation: à grande échelle, des indicateurs de biodiversité[4] peuvent être plus pertinents. D'autre part, les communautés sont considérées comme limitées, avec un nombre d'espèces fini: la courbe d'accumulation des espèces atteint donc théoriquement une asymptote quand l'effort d'inventaire est suffisant. Cette approche est opposée à celle, traitée dans les chapitres **??** et suivants, qui considère que la diversité augmente indéfiniment avec la surface,[5] que ce soit par changement d'échelle (élargir l'inventaire ajoute de nouvelles communautés) ou, plus

[1] R. K. Peet. "The Measurement of Species Diversity." In: *Annual review of ecology and systematics* 5 (1974), pp. 285–307. DOI: 10.1146/annurev.es.05.110174.001441.

[2] V. Devictor et al. "Spatial Mismatch and Congruence between Taxonomic, Phylogenetic and Functional Diversity: The Need for Integrative Conservation Strategies in a Changing World." In: *Ecology letters* 13.8 (2010), pp. 1030–40. DOI: 10.1111/j.1461-0248.2010.01493.x. PMID: 20545736; J. C. Stegen and A. H. Hurlbert. "Inferring Ecological Processes from Taxonomic, Phylogenetic and Functional Trait -Diversity." In: *PloS one* 6.6 (2011), e20906. DOI: 10.1371/journal.pone.0020906.

[3] K. R. Clarke and R. M. Warwick. "A Further Biodiversity Index Applicable to Species Lists: Variation in Taxonomic Distinctness." In: *Marine Ecology-Progress Series* 216 (2001), pp. 265–278. DOI: 10.3354/meps216265; C. Ricotta and G. C. Avena. "An Information-Theoretical Measure of Taxonomic Diversity." In: *Acta biotheoretica* 25.51 (2003), pp. 35–41. DOI: 10.1023/A:1023000322071.

[4] A. Balmford et al. "Measuring the Changing State of Nature." In: *Trends in Ecology & Evolution* 18.7 (2003), pp. 326–330. DOI: 10.1016/S0169-5347(03)00067-3.

[5] M. Williamson et al. "The Species-Area Relationship Does Not Have an Asymptote!" In: *Journal of Biogeography* 28.7 (2001), pp. 827–830. DOI: 10.1046/j.1365-2699.2001.00603.x.

théoriquement, parce que les communautés réelles sont considérées comme un tirage aléatoire parmi une infinités d'espèces.[6]

Les mesures présentées ici sont les plus utilisées: richesse, indices de Shannon et de Simpson, et l'indice de Hurlbert. Elles sont sujettes à des biais d'estimation,[7] notamment (mais pas seulement) à cause des espèces non échantillonnées.

Au chapitre suivant, l'entropie HCDT permettra d'unifier ces mesures et les nombres de Hill, et de leur donner un sens intuitif.

3.1 Richesse spécifique

La richesse est tout simplement le nombre d'espèces présentes dans le taxocène considéré. C'est la mesure conceptuellement la plus simple mais pratiquement la plus délicate dans des systèmes très riches comme les forêts tropicales: même avec des efforts d'inventaire considérables, il n'est en général pas possible de relever toutes les espèces rares, ce qui implique de recourir à des modèles mathématiques pour en estimer le nombre.

On ne fait pas de supposition sur la forme de la SAD (voir section 2) quand on utilise des méthodes d'estimation non paramétriques. Les estimateurs les plus connus sont ceux de Chao[8] et le *jackknife*.[9]

Une alternative consiste à inférer à partir des données les paramètres d'une SAD choisie, et particulièrement le nombre total d'espèces. Cette approche est bien moins répandue parce qu'elle suppose le bon choix du modèle et est beaucoup plus intensive en calcul. Il n'existe pas de meilleur estimateur universel[10] et il peut être efficace d'utiliser plusieurs méthodes d'estimation de façon concurrente sur les mêmes données.[11]

Techniques d'estimation non paramétrique

Dans le cadre d'un échantillonnage de n individus, on observe $f_{>0}$ espèces différentes parmi les S existantes. Chaque individu a une probabilité p_s d'appartenir à l'espèce s.

On ne sait rien sur la loi des p_s. On sait seulement, comme les individus sont tirés indépendamment les uns des autres, que l'espérance du nombre n_s d'individus de l'espèce s observé dans l'échantillon est np_s. La probabilité de ne pas observer l'espèce est $(1-p_s)^n$.

Pour les espèces fréquentes, np_s est grand, et les espèces sont observées systématiquement. La difficulté est due aux espèces pour lesquelles np_s, l'espérance du nombre d'observations, est petit. La probabilité de les observer

[6] Fisher et al., "The Relation between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population," see n. 6, p. 16.

[7] D. Mouillot and A. Leprêtre. "A Comparison of Species Diversity Estimators." In: *Researches on Population Ecology* 41.2 (1999), pp. 203–215. DOI: 10.1007/s101440050024.

[8] A. Chao. "Nonparametric Estimation of the Number of Classes in a Population." In: *Scandinavian Journal of Statistics* 11.4 (1984), pp. 265–270. JSTOR: 4615964.

[9] K. P. Burnham and W. S. Overton. "Robust Estimation of Population Size When Capture Probabilities Vary among Animals." In: *Ecology* 60.5 (1979), pp. 927–936. DOI: 10.2307/1936861.

[10] R. B. O'Hara. "Species Richness Estimators: How Many Species Can Dance on the Head of a Pin?" In: *Journal of Animal Ecology* 74 (2005), pp. 375–386. DOI: 10.1111/j.1365-2656.2005.00940.x.

[11] Y. Basset et al. "Arthropod Diversity in a Tropical Forest." In: *Science* 338.6113 (2012), pp. 1481–1484. DOI: 10.1126/science.1226727.

est donnée par la loi binomiale: si np_s est proche de 0, la probabilité d'observer un individu est faible.

Les estimateurs non paramétriques cherchent à tirer le maximum d'information de la distribution des abondances n_s pour estimer le nombre d'espèces non observées. Une présentation détaillée du problème et des limites à sa résolution est fournie par Mao and Colwell[12] qui concluent notamment que les estimateurs ne peuvent fournir qu'une borne inférieure de l'intervalle des possibles valeurs du nombre réel d'espèces.

[12] C. X. Mao and R. K. Colwell. "Estimation of Species Richness: Mixture Models, the Role of Rare Species, and Inferential Challenges." In: *Ecology* 86.5 (2005), pp. 1143–1153. DOI: 10.1890/04-1078.

Chao1 et Chao2

Chao[13] estime le nombre d'espèces non observées à partir de celles observées 1 ou 2 fois.

[13] Chao, see n. 8.

Dans un échantillon de taille n résultant d'un tirage indépendant des individus, la probabilité que l'espèce s soit observée ν fois est obtenue en écrivant la probabilité de tirer dans l'ordre ν fois l'espèce s puis $n-\nu$ fois une autre espèce, multiplié par le nombre de combinaisons possible pour prendre en compte l'ordre des tirages:

$$p_{s,\nu}(n) = \binom{n}{\nu} p_s^{\nu} \left(1-p_s\right)^{n-\nu}. \tag{3.1}$$

L'espérance du nombre d'espèces observées ν fois, $\mathbb{E}(f_\nu)$, est obtenue en sommant pour toutes les espèces la probabilité de les observer ν fois:

$$\mathbb{E}\left(f_\nu\right) = \binom{n}{\nu} \sum_s p_s^{\nu} \left(1-p_s\right)^{n-\nu}. \tag{3.2}$$

Le carré de la norme du vecteur en S dimensions dont les coordonnées sont $(1-p_s)^{n/2}$ est

$$\sum_s (1-p_s)^n,$$

c'est-à-dire $\mathbb{E}(f_0)$, l'espérance du nombre d'espèces non observées. Celui du vecteur de coordonnées $p_s(1-p_s)^{n/2-1}$ est

$$\sum_s p_s^2 (1-p_s)^{n-2} = \frac{2}{n(n-1)} \mathbb{E}(f_2).$$

Enfin, le produit scalaire des deux vecteurs vaut

$$\sum_s p_s (1-p_s)^{n-1} = \frac{1}{n} \mathbb{E}(f_1).$$

L'inégalité de Cauchy-Schwarz (le produit scalaire est inférieur au produit des normes des vecteurs) peut être appliquée

aux deux vecteurs (tous les termes sont au carré):

$$\left[\sum_s (1-p_s)^n\right]\left[\sum_s p_s^2(1-p_s)^{n-2}\right] \geq \left[\sum_s p_s(1-p_s)^{n-1}\right]^2, \tag{3.3}$$

d'où

$$\mathbb{E}(f_0) \geq \frac{n-1}{n}\frac{\left[\mathbb{E}(f_1)\right]^2}{2\mathbb{E}(f_2)}. \tag{3.4}$$

L'estimateur est obtenu en remplaçant les espérances par les valeurs observées:

$$\hat{S}_{Chao1} = f_{>0} + \frac{(n-1)\,(f_1)^2}{2nf_2}, \tag{3.5}$$

où $f_{>0}$ est le nombre d'espèces différentes observé.

Il s'agit d'un estimateur minimum: l'espérance du nombre d'espèces est supérieure ou égale au nombre estimé.

Béguinot[14] a montré que l'estimateur est sans biais si le nombre d'espèces non observées décroît exponentiellement avec la taille de l'échantillon:

$$f_0 = Se^{-kn}, \tag{3.6}$$

où k est un réel strictement positif. Cette relation est cohérente avec un échantillonnage poissonien dans lequel la densité des individus est constante: voir le chapitre **??**.

Si aucune espèce n'est observée deux fois, l'estimateur est remplacé par

$$\hat{S}_{Chao1} = f_{>0} + \frac{(n-1)\,f_1\,(f_1-1)}{2n}. \tag{3.7}$$

Si n n'est pas trop petit, les approximations suivantes sont possibles:

$$\hat{S}_{Chao1} = f_{>0} + \frac{(f_1)^2}{2f_2}. \tag{3.8}$$

Si aucune espèce n'est observée deux fois, l'estimateur est remplacé[15] par

$$\hat{S}_{Chao1} = f_{>0} + f_1\,(f_1-1)/2. \tag{3.9}$$

La variance de l'estimateur est connue, mais pas sa distribution:

$$\mathrm{Var}\left(\hat{S}_{Chao1}\right) = f_2\left[\frac{1}{2}\left(\frac{f_1}{f_2}\right)^2 + \left(\frac{f_1}{f_2}\right)^3 + \frac{1}{4}\left(\frac{f_1}{f_2}\right)^4\right]. \tag{3.10}$$

[14] J. Béguinot. "An Algebraic Derivation of Chao's Estimator of the Number of Species in a Community Highlights the Condition Allowing Chao to Deliver Centered Estimates." In: *International Scholarly Research Notices* 2014 (Article ID 847328 2014). DOI: 10.1155/2014/847328.

[15] A. Chao. "Species Richness Estimation." In: *Encyclopedia of Statistical Sciences*. Ed. by N. Balakrishnan et al. 2nd ed. New York: Wiley, 2004.

Si aucune espèce n'est observée deux fois:

$$\operatorname{Var}\left(\hat{S}_{Chao1}\right) = \frac{f_1\left(f_1-1\right)}{2} + \frac{f_1(2f_1-1)^2}{4} + \frac{(f_1)^4}{4f_{>0}}. \quad (3.11)$$

Chao[16] donne une approximation de l'intervalle de confiance à 95% en assumant une distribution normale:

$$f_{>0} + \frac{\hat{S}_{Chao1} - f_{>0}}{c} \leq S \leq f_{>0} + \left(\hat{S}_{Chao1} - f_{>0}\right) c, \quad (3.12)$$

où

$$c = e^{t^n_{1-\alpha/2}\sqrt{\ln\left(1+\frac{\operatorname{Var}(\hat{S}_{Chao1})}{\left(\hat{S}_{Chao1}-f_{>0}\right)^2}\right)}}. \quad (3.13)$$

Eren et al.[17] calculent un intervalle de confiance qui est plus petit quand la valeur maximum théorique du nombre d'espèces est connue, ce qui est rarement le cas en écologie.

Chao[18] propose un estimateur du nombre d'espèces appliqué aux données de présence-absence (un certain nombre de relevés contiennent seulement l'information de présence ou absence de chaque espèce), appelé Chao2. Il est identique à Chao1 mais n est le nombre de relevés, en général trop petit pour appliquer l'approximation de Chao1.

Chiu et al.[19] améliorent l'estimateur en reprenant la démarche originale de Chao mais en utilisant un estimateur plus précis du taux de couverture, (1.11) au lieu de (1.6):

$$\hat{S}_{iChao1} = \hat{S}_{Chao1} + \frac{f_3}{4f_4}\max\left(f_1 - \frac{f_2 f_3}{2f_4}; 0\right). \quad (3.14)$$

Chao et al.[20] étendent l'applicabilité de l'estimateur Chao2 à des données dans lesquelles les espèces sont notées uniquement comme singletons ou doubletons et plus, sans distinction entre doubletons et espèces plus fréquentes. Une relation entre le nombre de doubletons et les données disponibles est fournie; sa résolution numérique (le code R nécessaire est disponible avec l'article) permet d'estimer f_2 et de l'injecter dans l'estimateur Chao2.

[16] A. Chao. "Estimating the Population Size for Capture-Recapture Data with Unequal Catchability." In: *Biometrics* 43.4 (1987), pp. 783–791. DOI: 10.2307/2531532.

[17] M. I. Eren et al. "Estimating the Richness of a Population When the Maximum Number of Classes Is Fixed: A Nonparametric Solution to an Archaeological Problem." In: *Plos One* 7.5 (2012). DOI: 10.1371/journal.pone.0034179, eq. 8.

[18] Chao, see n. 16.

[19] Chiu et al., "An Improved Nonparametric Lower Bound of Species Richness via a Modified Good-Turing Frequency Formula," see n. 21, p. 10.

[20] A. Chao et al. "Seen Once or More than Once: Applying Good-Turing Theory to Estimate Species Richness Using Only Unique Observations and a Species List." In: *Methods in Ecology and Evolution* 8.10 (2017), pp. 1221–1232. DOI: 10.1111/2041-210X.12768.

L'estimateur ACE

Chao and Lee[21] développent l'estimateur ACE (*Abundance-based coverage estimator*) à travers l'estimation du taux de couverture C. L'estimateur ACE utilise toutes les valeurs de f_ν correspondant aux espèces rares: concrètement, la valeur

[21] A. Chao and S.-M. Lee. "Estimating the Number of Classes via Sample Coverage." In: *Journal of the American Statistical Association* 87.417 (1992), pp. 210–217. DOI: 10.1080/01621459.1992.10475194.

limite de ν notée κ est fixée arbitrairement, généralement à 10.

L'estimateur prend en compte le coefficient de variation de la distribution des fréquences ($\hat{p}_s$): plus les probabilités sont hétérogènes, plus le nombre d'espèces non observées sera grand. Finalement:

$$\hat{S}_{ACE} = f_{>\kappa} + \frac{f_{\leq\kappa}}{\hat{C}_{rare}} + \frac{f_1}{\hat{C}_{rare}}\hat{\gamma}_{rare}. \tag{3.15}$$

$f_{>\kappa}$ est le nombre d'espèces dites abondantes, observées plus de κ fois, $f_{\leq\kappa}$ le nombre d'espèces dites rares, observées κ fois ou moins. $\hat{C}_{rare}$ est le taux de couverture ne prenant en compte que les espèces rares.

L'estimateur du coefficient de variation au carré est

$$\hat{\gamma}^2_{rare} = \max\left(\frac{f_{\leq\kappa}\sum_{\nu=1}^{\kappa}\nu\left(\nu-1\right)f_\nu}{\hat{C}_{rare}\left(\sum_{\nu=1}^{\kappa}\nu f_\nu\right)\left(\sum_{\nu=1}^{\kappa}\nu f_\nu - 1\right)} - 1; 0\right). \tag{3.16}$$

Lorsque l'hétérogénéité est très forte, un autre estimateur est plus performant:

$$\tilde{\gamma}^2_{rare} = \max\left(\hat{\gamma}^2_{rare}\left(1 + \frac{\left(1-\hat{C}_{rare}\right)\sum_{\nu=1}^{\kappa}\nu\left(\nu-1\right)f_\nu}{\hat{C}_{rare}\left(\sum_{\nu=1}^{\kappa}\nu f_\nu - 1\right)}\right); 0\right). \tag{3.17}$$

Chao and Shen[22] conseillent d'utiliser le deuxième estimateur dès que $\hat{\gamma}^2_{rare}$ dépasse 0,8. L'estimateur ACE donne normalement une valeur plus grande que Chao1. Si ce n'est pas le cas, la limite des espèces rares κ doit être augmentée.

[22] Chao and Shen, *Program SPADE: Species Prediction and Diversity Estimation. Program and User's Guide.* See n. 28, p. 11.

L'estimateur jackknife

La méthode jackknife a pour objectif de réduire le biais d'un estimateur en considérant des jeux de données dans lesquels on a supprimé un certain nombre d'observations (ce nombre est l'*ordre* de la méthode). Burnham et Overton[23] ont utilisé cette technique pour obtenir des estimateurs du nombre d'espèces, appelés jackknife à l'ordre j, prenant en compte les valeurs de f_1 à f_j. Les estimateurs du premier et du deuxième ordre sont les plus utilisés en pratique:

$$\hat{S}_{J1} = f_{>0} + \frac{\left(n-1\right)f_1}{n}, \tag{3.18}$$

$$\hat{S}_{J2} = f_{>0} + \frac{\left(2n-3\right)f_1}{n} - \frac{\left(n-2\right)^2 f_2}{n\left(n-1\right)}. \tag{3.19}$$

[23] K. P. Burnham and W. S. Overton. "Estimation of the Size of a Closed Population When Capture Probabilities Vary among Animals." In: *Biometrika* 65.3 (1978), pp. 625–633. DOI: 10.2307/2335915; Burnham and Overton, "Robust Estimation of Population Size When Capture Probabilities Vary among Animals," see n. 9, p. 26.

$$\hat{S}_{J3} = f_{>0} + \frac{(3n-6)\, f_1}{n} - \frac{(3n^2-15n+19)\, f_2}{n\,(n-1)} + \frac{(n-3)^3 f_3}{n\,(n-1)\,(n-2)}. \tag{3.20}$$

Augmenter l'ordre du jackknife diminue le biais mais augmente la variance de l'estimateur.

Chao[24] a montré que les estimateurs jackknife pouvaient être retrouvés par approximation de l'indice Chao1.

La variance du jackknife d'ordre 1 est[25]

$$\mathrm{Var}\left(\hat{S}_{J1}\right) = \frac{n-1}{n}\left(\sum_{j=1}^{n} j^2 f_j - \frac{f_{>0}^2}{n}\right). \tag{3.21}$$

L'estimateur est construit à partir de l'hypothèse selon laquelle le nombre d'espèces non observées est de la forme

$$f_0(n) = \sum_{i=1}^{\infty} \frac{a_i}{n^i},$$

où la notation $f_0(n)$ est utilisée pour expliciter sa dépendance à l'effort d'échantillonnage.

Pour cette raison, Cormack[26] affirme qu'il n'a pas de support théorique solide. L'espérance du nombre d'espèces non observées est (eq. (3.2)) $\sum_s (1-p_s)^n$, qui décroît beaucoup plus rapidement que $\sum_i a_i/n^i$: l'hypothèse est bien fausse. En revanche, pour une gamme de n fixée (de la taille de l'inventaire à une taille suffisante pour approcher la richesse asymptotique par exemple), il est toujours possible d'écrire le nombre d'espèces non observées sous la forme d'une série de puissances négatives de n, comme dans l'illustration ci-dessous.

Une communauté log-normale, similaire à BCI (300 espèces, écart-type égal à 2) est simulée et un échantillon de 1000 individus est tiré.

```
# Ecart-type
sdlog <- 2
# Nombre d'espèces
S <- 300
# Tirage des probabilités log-normales
lnorm_abd <- rlnorm(S, 0, sdlog)
lnorm_prob <- lnorm_abd / sum(lnorm_abd)
# Taille de l'échantillon
n <- 1000
# Tirage d'un échantillon
library("divent")
abundances <- rcommunity(1, size = n, prob = lnorm_prob)
```

L'échantillon est présenté en figure 3.1.

Code de la figure 3.1:

[24] Chao, "Nonparametric Estimation of the Number of Classes in a Population," see n. 8, p. 26.

[25] J. F. Heltshe and N. E. Forrester. "Estimating Species Richness Using the Jackknife Procedure." In: *Biometrics* 39.1 (1983), pp. 1–11. DOI: 10.2307/2530802. JSTOR: 2530802.

[26] R. M. Cormack. "Log-Linear Models for Capture-Recapture." In: *Biometrics* 45.2 (1989), pp. 395–413. DOI: 10.2307/2531485.

```
autoplot(abundances, fit_rac = TRUE, distribution="lnorm")
```

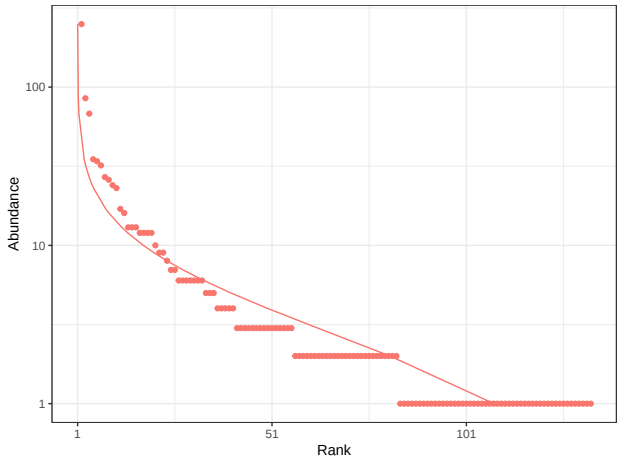


Figure 3.1: Echantillon de 1000 individus tiré dans une communauté log-normale.

Il est possible de vérifier que l'espérance du nombre d'espèces non observées correspond bien à la moyenne des observations.

```
# Espérance des espèces non vues
E0 <- (1 - lnorm_prob)^n
(f0 <- sum(E0))
```

```
## [1] 169.9725
```

```
# Tirage de 1000 échantillons, nombre moyen d'espèces observées
(s_obs <- mean(colSums(rmultinom(1000, size = n, prob = lnorm_prob) > 0)))
```

```
## [1] 129.987
```

```
# Vérification: nombre d'espèces observées en moyenne et non observées
s_obs + f0
```

```
## [1] 299.9595
```

```
# Nombre total d'espèces dans la communauté
(S <- length(lnorm_prob))
```

```
## [1] 300
```

Le nombre d'espèces non observées peut être écrit sous la forme d'une série de puissances négatives de n, comme le prévoit le jackknife, entre deux valeurs de n fixées.

```
# Echantillonnage de 500 à 5000 individus
n_seq <- 500:5000
# Calcul du nombre d'espèces non observées
bias <- sapply(n_seq, function(n) sum((1 - lnorm_prob)^n))
```

Le nombre d'espèces non observées, qui est le biais de l'estimateur de la richesse, est présenté en figure 3.2.

La courbe peut être approchée par une série de puissances négatives de n dont quelques termes sont présentés sur la figure.

```
# Ordre 1
lm1 <- lm(bias ~ 0 + I(1  /n_seq))
# Ordre 2
lm2 <- lm(bias ~ 0 + I(1 / n_seq) + I(1 / n_seq^2))
# Ordre 4
lm4 <- lm(
  bias ~ 0 + I(1 / n_seq) + I(1 / n_seq^2) + I(1 / n_seq^3) + I(1 / n_seq^4)
)
```

Les termes a_i de la série de puissances négatives sont estimées par des modèles linéaires. A l'ordre 1, le modèle *lm1* fournit une approximation grossière du nombre d'espèces non observées avec un seul terme ($f_0(n) \approx a_1/n$). Le modèle s'appproche de plus en plus des données en augmentant le nombre de terme. Le modèle *lm4* contient 4 termes a_1 à a_4 :

```
lm4$coefficients
```

```
##    I(1/n_seq)  I(1/n_seq^2)  I(1/n_seq^3)  I(1/n_seq^4)
##  5.661773e+05 -8.025902e+08  5.183902e+11 -1.180337e+14
```

A partir de 6 termes, les valeurs du biais d'estimation sont presque parfaitement estimées.

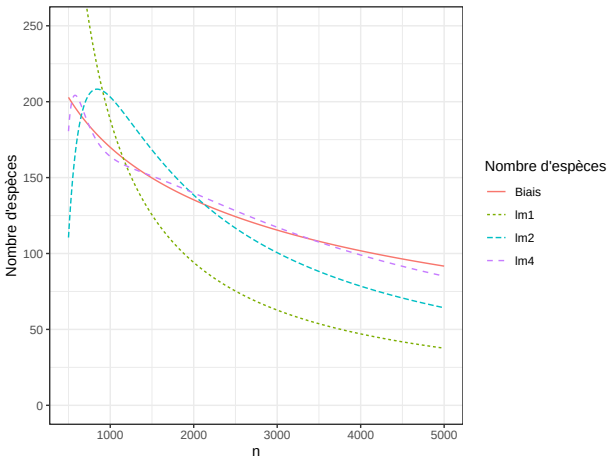


Figure 3.2: Nombre d'espèces non observées dans un échantillon de taille croissante et sa décomposition en séries de puissances négatives de n. Le nombre d'espèces non observées est représenté par la courbe continue. Les séries de puissances négatives d'ordre 1, 2 et 4, notées *lm1* à *lm4*, sont représentées en pointillés. Les courbes d'ordre 6 et plus sont confondues avec la courbe noire.

Code de la figure 3.2:

```
tibble(
  n = n_seq,
  Biais = bias,
  lm1 = predict(lm1),
  lm2 = predict(lm2),
  lm4 = predict(lm4)
) %>%
  pivot_longer(cols = -n) %>%
  ggplot() +
```

```
  geom_line(aes(x = n, y = value, color = name, lty = name)) +
  coord_cartesian(ylim = c(0, 250)) +
  labs(
    color = "Nombre d'espèces",
    lty = "Nombre d'espèces",
    y = "Nombre d'espèces"
  )
```

L'ajustement est possible pour des valeurs de n différentes mais les coefficients a_i sont alors différents: la forme du biais n'est valide que pour une gamme de valeurs de n fixée.

Béguinot[27] apporte un autre argument important en faveur du jackknife. À condition que n soit suffisamment grand, l'estimateur du nombre d'espèces non observées est une fonction linéaire du nombre d'espèces observées ν fois: f_1 pour le jackknife 1, $2f_1 - f_2$ pour le jackknife 2 et ainsi de suite pour les ordres suivants, contrairement à l'estimateur de Chao. Grâce à cette propriété, l'estimateur du jackknife est additif quand plusieurs groupes d'espèces disjoints sont pris en compte: l'estimation du nombre d'espèces de papillons et de scarabées inventoriées ensemble est égale à la somme des estimations des deux groupes inventoriés séparément. Ce n'est pas le cas pour l'estimateur de Chao.

L'estimateur du jackknife est très utilisé parce qu'il est efficace en pratique, notamment parce que son ordre peut être adapté aux données.

[27] J. Béguinot. "Basic Theoretical Arguments Advocating Jackknife-2 as Usually Being the Most Appropriate Nonparametric Estimator of Total Species Richness." In: *Annual Research & Review in Biology* 10.1 (2016), pp. 1–12. DOI: 10.9734/ARRB/2016/25104.

L'estimateur du bootstrap

L'estimateur du bootstrap[28] est

$$\hat{S}_b = f_{>0} + \sum_s (1 - p_s)^n. \tag{3.22}$$

Il est peu utilisé parce que le jackknife est plus performant.[29]

[28] E. P. Smith and G. V. Belle. "Nonparametric Estimation of Species Richness." In: *Biometrics* 40.1 (1984), pp. 119–129. DOI: 10.1002/9780470015902.a0026329.

[29] R. K. Colwell and J. A. Coddington. "Estimating Terrestrial Biodiversity through Extrapolation." In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 345.1311 (1994), pp. 101–118. DOI: 10.1098/rstb.1994.0091.

[30] Chao et al., "SpadeR: Species Prediction and Diversity Estimation with R," see n. 29, p. 11.

Calcul

Ces estimateurs peuvent être calculés de façon relativement simple à l'aide du logiciel SPADE, dans sa version pour R.[30] Le guide de l'utilisateur présente quelques estimateurs supplémentaires et des directives pour choisir. Il est conseillé d'utiliser Chao1 pour une estimation minimale, et ACE pour une estimation moins biaisée de la richesse.

Les intervalles de confiance de chaque estimateur sont calculés par bootstrap: même quand la variance d'un estimateur est connue, sa loi ne l'est généralement pas, et le calcul analytique de l'intervalle de confiance n'est pas possible.

Les estimateurs et leurs intervalles de confiance peuvent également être calculés avec le package *vegan* qui dispose pour cela de deux fonctions: `specpool` et `estimateR`.

`specpool` est basé sur les incidences des espèces dans un ensemble de sites d'observation et donne une estimation unique de la richesse selon les méthodes Chao2, jackknife (ordre 1 et 2) et bootstrap. L'écart-type de l'estimateur est également fourni par la fonction, sauf pour le jackknife d'ordre 2.

`estimateR` est basé sur les abondances des espèces et retourne un estimateur de la richesse spécifique par site et non global comme `specpool`.

Exemple

On utilise les données de Barro Colorado Island (BCI). La parcelle a été divisée en carrés de 1 ha. Le tableau d'entrée est un `dataframe` contenant, pour chaque espèce d'arbres ($DBH \geq 10$ cm), ses effectifs par carré.

On charge le tableau de données:

```
library("vegan")
data(BCI)
```

On utilise la fonction `estimateR` pour calculer la richesse des deux premiers carrés:

```
estimateR(BCI[1:2,])
```

```
##                   1          2
## S.obs      93.000000  84.000000
## S.chao1   117.473684 117.214286
## se.chao1   11.583785  15.918953
## S.ACE     122.848959 117.317307
## se.ACE      5.736054   5.571998
```

Le package *SPECIES*[31] permet de calculer les estimateurs jackknife d'ordre supérieur à 2 et surtout choisit l'ordre qui fournit le meilleur compromis entre biais et variance.

[31] J.-P. Wang. "SPECIES: An R Package for Species Richness Estimation." In: *Journal of Statistical Software* 40.9 (2011), pp. 1–15. DOI: 10.18637/jss.v040.i09.

Comparaison des fonctions sur l'ensemble du dispositif BCI ($f_{>0} = 225$, $f_1 = 19$):

```
specpool(BCI)
```

```
##     Species      chao chao.se  jack1 jack1.se    jack2
## All     225 236.3732 6.54361 245.58 5.650522 247.8722
##         boot  boot.se  n
## All 235.6862 3.468888 50
```

```
library("SPECIES")
# Distribution du nombre d'espèces (vecteur:
# noms = nombre d'individus
# valeurs = nombres d'espèces ayant ce nombre d'individus)
bci_abd <- colSums(BCI)
# Mise au format requis (matrice:
# colonne 1 = nombre d'individus
# colonne 2 = nombres d'espèces ayant ce nombre d'individus)
# par la fonction abd_freq_count dans divent
jackknife(as.matrix(abd_freq_count(bci_abd)))
```

```
##
## Your specified order is larger than that determined by the test,
## Therefore the order from the test is used.
```

```
## $JackknifeOrder
## [1] 1
##
## $Nhat
## [1] 244
##
## $SE
## [1] 6.164414
##
## $CI
##       lb  ub
## [1,] 232 256
```

Comparaison avec la valeur de l'équation (3.18):

```
# Nombre d'espèces par effectif observé
bci_abd_freq_count <- tapply(bci_abd, bci_abd, length)
# Calcul direct de Jack1
sum(bci_abd_freq_count) +
  bci_abd_freq_count[1] * (sum(bci_abd) - 1) / sum(bci_abd)
```

```
##        1
## 243.9991
```

La valeur du jackknife 1 fournie par `specpool` est fausse. La fonction `jackknife` de *SPECIES* donne le bon résultat, avec un intervalle de confiance calculé en supposant que la distribution est normale ($\pm$ 1,96 écart-type au seuil de 95%).

L'estimateur du bootstrap est calculable simplement:

```
# Effectif total
bci_n <- sum(bci_abd)
# Probabilités
bci_prob <- bci_abd / bci_n
# Estimateur du bootstrap
length(bci_prob) + sum((1 - bci_prob)^bci_n)
```

```
## [1] 234.3517
```

Choix de l'estimateur

Des tests empiriques poussés ont été menés par Brose et al.[32] pour permettre le choix du meilleur estimateur de la richesse en fonction de la complétude de l'échantillonnage $f_{>0}/S$. Les auteurs appellent cette proportion couverture (*coverage*). Le terme *completeness* a été proposé par Beck and Schwanghart[33] pour éviter la confusion avec le taux de couverture défini par Good (vu en section 1.5). La complétude est inférieure à la couverture: toutes les espèces ont le même poids alors que les espèces manquantes sont plus rares et pénalisent moins le taux de couverture.

Dans tous les cas, les estimateurs jackknife sont les meilleurs. L'arbre de décision est en figure 3.3.[34] Le choix

[32] U. Brose et al. "Estimating Species Richness: Sensitivity to Sample Coverage and Insensitivity to Spatial Patterns." In: *Ecology* 84.9 (2003), pp. 2364–2377. DOI: 10.1890/02-0558.

[33] J. Beck and W. Schwanghart. "Comparing Measures of Species Diversity from Incomplete Inventories: An Update." In: *Methods in Ecology and Evolution* 1.1 (2010), pp. 38–44. DOI: 10.1111/j.2041-210X.2009.00003.x.

[34] Brose et al., see n. 32, fig. 6.

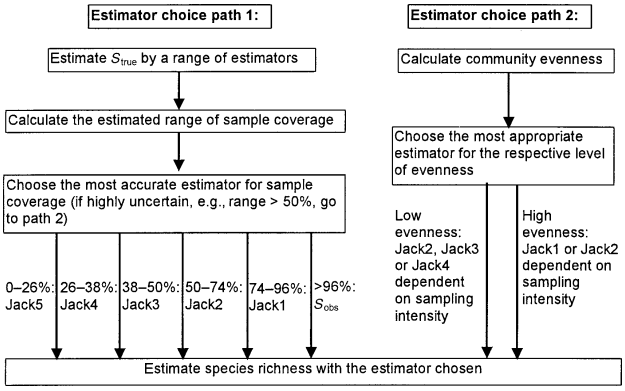


Figure 3.3: Arbre de décision du meilleur estimateur du nombre d'espèces.

dépend principalement de la complétude (*coverage* sur la figure). Une première estimation est nécessaire par plusieurs estimateurs. Si les résultats sont cohérents, choisir un estimateur jackknife d'ordre d'autant plus faible que la complétude est grande. Au-delà de 96%, le nombre d'espèces observé est plus performant parce que les jackknifes surestiment S. S'ils sont incohérents (intervalle des estimations supérieur à 50% de leur moyenne), le critère majeur est l'équitabilité (voir section **??**). Si elle est faible (de l'ordre de 0,5 à 0,6), les estimateurs jackknife 2 à 4 sont performants, l'ordre diminuant avec l'intensité d'échantillonnage (forte: 10%, faible: 0,5% de la communauté). Pour une forte équitabilité (0,8 à 0,9), on préférera jackknife 1 ou 2.

Pour BCI, le nombre d'espèces estimé par jackknife 1 est 244. La complétude est 225/244 = 92%, dans le domaine de validité du jackknife 1 (74% à 96%) qui est donc le bon estimateur.

La parcelle 6 de Paracou nécessite l'estimateur jackknife 2:

```
library("divent")
div_richness(colSums(paracou_6_abd))
```

```
## # A tibble: 1 x 3
##   estimator   order diversity
##   <chr>       <dbl>     <dbl>
## 1 Jackknife 2     0       473
```

```
# Complétude
div_richness(colSums(paracou_6_abd), estimator = "naive", as_numeric = TRUE) /
  div_richness(colSums(paracou_6_abd), as_numeric = TRUE)
```

```
## [1] 0.7082452
```

Chiu et al.,[35] à partir d'autres simulations, préfèrent l'utilisation de l'estimateur *iChao1*. Quand l'échantillonnage est suffisant, les estimateurs de Chao ont l'avantage de posséder une base théorique solide et de fournir une borne

[35] Chiu et al., "An Improved Nonparametric Lower Bound of Species Richness via a Modified Good-Turing Frequency Formula," see n. 21, p. 10.

inférieure du nombre d'espèces possible. Dans ce cas, les estimations du jackknife d'ordre 1 sont cohérentes avec celles de Chao. En revanche, quand l'échantillonnage est insuffisant, l'estimateur jackknife d'ordre supérieur à 1 permet de réduire le biais d'estimation, au prix d'une variance accrue.[36]

Enfin, Beguinot[37] suggère d'utiliser en règle générale le jackknife 2 (mais ne traite pas les cas dans lesquels l'échantillonnage est trop faible pour justifier un ordre supérieur) tant que le nombre de singletons est supérieur à $2 - \sqrt{(2)} \approx 0,6$ fois le nombre de doubletons. Le ratio des singletons sur les doubletons diminue quand l'échantillonnage approche de l'exhaustivité. Quand le seuil de 0,6 est dépassé, la valeur de l'estimateur de Chao devient supérieur au jackknife 2 et doit être utilisé. Ce seuil est cohérent avec les règles de Brose et al.[38]

[36] E. Marcon. "Practical Estimation of Diversity from Abundance Data." In: *HAL* 01212435 (version 2 2015).

[37] J. Béguinot. "Extrapolation of the Species Accumulation Curve for Incomplete Species Samplings: A New Nonparametric Approach to Estimate the Degree of Sample Completeness and Decide When to Stop Sampling." In: *Annual Research & Review in Biology* 8.5 (2015), pp. 1–9. DOI: 10.9734/ARRB/2015/22351; Béguinot, "Basic Theoretical Arguments Advocating Jackknife-2 as Usually Being the Most Appropriate Nonparametric Estimator of Total Species Richness," see n. 27, p. 34.

[38] Brose et al., "Estimating Species Richness: Sensitivity to Sample Coverage and Insensitivity to Spatial Patterns," see n. 32, p. 36.

Prédiction de la richesse d'un nouvel échantillon

La prédiction du nombre d'espèces $\hat{S}'$ découvert dans une nouvelle placette d'un habitat dans lequel on a déjà échantillonné est une question importante, par exemple pour évaluer le nombre d'espèces préservées dans le cadre d'une mise en réserve, ou évaluer le nombre d'espèces perdues en réduisant la surface d'une forêt.

Shen et al.[39] proposent un estimateur et le confrontent avec succès à des estimateurs antérieurs. On note $\hat{f}_0(n)$ l'estimateur du nombre d'espèces non observées dans le premier échantillon, et $\hat{C}$ l'estimateur de son taux de couverture. L'estimateur du nombre d'espèces du nouvel échantillon de n' individus est

$$\hat{S}' = \hat{f}_0(n) \left[1 - \left(1 - \frac{1 - \hat{C}}{\hat{f}_0(n)} \right)^{n'} \right]. \tag{3.23}$$

$\hat{f}_0(n)$ est obtenu par la différence entre les nombres d'espèces estimé et observé: $\hat{f}_0(n) = \hat{S} - f_{>0}(n)$.

Exemple de BCI, suite: combien de nouvelles espèces seront découvertes en échantillonnant plus?

[39] T.-J. Shen et al. "Predicting the Number of New Species in a Further Taxonomic Sampling." In: *Ecology* 84.3 (2003), pp. 798–804. DOI: 10.1890/0012-9658(2003)084[0798:PTNONS]2.0.CO;2.

```
# Espèces non observées
bci_abd %>%
  div_richness() %>%
  pull(diversity) %>%
  `-`(length(BCI_abd)) %>%
  print() ->
  bci_f_0
```

```
## [1] 19
```

```
# Taux de couverture
bci_abd %>%
  coverage %>%
  pull(coverage) %>%
  print ->
  bci_C
```

```
## [1] 0.9991146
```

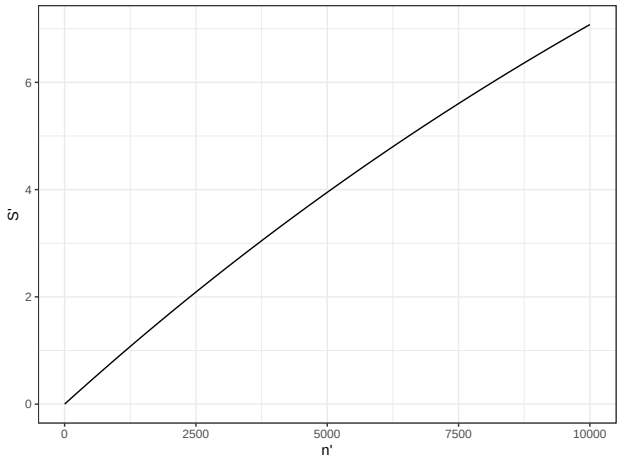


Figure 3.4: Nombre de nouvelles espèces découvertes en fonction de l'effort d'échantillonnage supplémentaire (données de BCI). Seulement 7 nouvelles espèces seront observées en échantillonnant 10000 arbres supplémentaires (environ 25 ha en plus des 50 ha de la parcelle qui contiennent 225 espèces).

Le taux de couverture de l'inventaire de BCI est très proche de 100%, donc peu de nouvelles espèces seront découvertes en augmentant l'effort d'échantillonnage. La courbe obtenue est en figure 3.4.

Le code R nécessaire pour réaliser la figure est:

```
# Nouvelles espèces en fonction du nombre de nouveaux individus
S_prime <- function(n_prime, f_0, C) {
  f_0 * (1 - (1 - (1 - C) / f_0)^n_prime)
}
# Graphique
tibble(x = 1:10000) %>%
  ggplot(aes(x)) +
  stat_function(
    fun = S_prime,
    args = list(f_0 = bci_f_0, C = bci_C)
  ) +
  labs(x = "n'", y = "S'")
```

La question de l'extrapolation de la richesse est traitée plus en détail dans les sections **??** et **??**.

Inférence du nombre d'espèces à partir de la SAD

Distribution de Preston

Preston[40] fournit dès l'introduction de son modèle log-normal une technique d'estimation du nombre total d'espèces par la célèbre méthode des octaves. Elle est disponible dans le package *vegan*:

[40] Preston, "The Commonness, and Rarity, of Species," see n. 1, p. 15.

```
veiledspec(bci_abd)
```

```
## Extrapolated     Observed       Veiled
##    235.40577    225.00000     10.40577
```

L'ajustement direct du modèle aux données, sans regroupement par octaves,[41] est également possible (figure 3.5):

[41] M. Williamson and K. J. Gaston. "The Lognormal Distribution Is Not an Appropriate Null Hypothesis for the Species-Abundance Distribution." In: *Journal of Animal Ecology* 74.2001 (2005), pp. 409–422. DOI: 10.1111/j.1365-2656.2005.00936.x.

```
bci_preston <- prestondistr(bci_abd)
veiledspec(bci_preston)
```

```
## Extrapolated     Observed       Veiled
##   230.931018   225.000000     5.931018
```

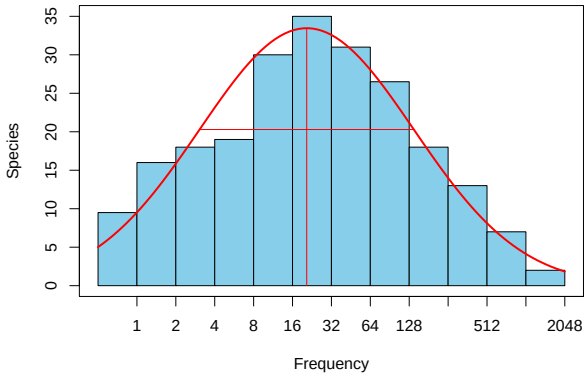


Figure 3.5: Ajustement du modèle de Preston aux données de BCI.

Le code R nécessaire pour réaliser la figure est:

```
plot(bci_preston)
```

Maximum de vraisemblance d'une distribution de Fisher

Norris and Pollock[42] supposent que la distribution des espèces suit le modèle de Fisher (voir chapitre **??**) et infèrent le nombre d'espèces par maximum de vraisemblance non paramétrique (ils ne cherchent pas à inférer les paramètres de la loi de probabilité de p_s mais seulement à ajuster au mieux le modèle de Poisson aux valeurs de $\hat{p}_s$ observées).

[42] J. L. Norris and K. H. Pollock. "Non-Parametric MLE for Poisson Species Abundance Models Allowing for Heterogeneity between Species." In: *Environmental and Ecological Statistics* 5.4 (1998), pp. 391–402. DOI: 10.1023/A:1009659922745.

Le calcul est possible avec la librairie *SPECIES* de R:

```
# Mise au format requis (matrice:
# colonne 1 = nombre d'individus
# colonne 2 = nombres d'espèces ayant ce nombre d'individus)
bci_abd_freq_count <- as.matrix(abd_freq_count(bci_abd))
# Regroupement de la queue de distribution: la longueur du vecteur est limitée à 25 pour al
bci_abd_freq_count[25, 2] <- sum(
  bci_abd_freq_count[25:nrow(bci_abd_freq_count), 2]
)
bci_abd_freq_count <- bci_abd_freq_count[1:25, ]
unpmle(bci_abd_freq_count)
```

```
## Method: Unconditional NPMLE method by Norris and Pollock 1996, 1998,
##         using algorithm by Wang and Lindsay 2005:
##
##         MLE=                                239
##         Estimated Poisson mixture components:
##         p=                                  1.10372 3.595437 10.60832
##         pi=                                 0.402579 0.2525368 0.3448842

## $Nhat
## [1] 239
```

Le problème de cette méthode d'estimation est qu'elle diverge fréquemment. Les calculs n'aboutissent pas si la queue de distribution n'est pas regroupée (il existe 108 valeurs différentes de n_s dans l'exemple de BCI: aucune des fonctions de *SPECIES* ne fonctionne en l'état).

Wang et al.[43] ont amélioré sa stabilité en pénalisant le calcul de la vraisemblance:

[43] J.-P. Wang et al. "Gene Capture Prediction and Overlap Estimation in EST Sequencing from One or Multiple Libraries." In: *BMC Bioinformatics* 6.1 (2005), p. 300. DOI: 10.1186/1471-2105-6-300.

```
pnpmle(bci_abd_freq_count)
```

```
## Method: Penalized NPMLE method by  Wang and Lindsay 2005.
##
##         MLE=                                134
##         Estimated zero-truncated Poisson mixture components:
##         p=                                  4.802198
##         pi=                                 NaN

## $Nhat
## [1] 134
```

Enfin, Wang[44] perfectionne la technique d'estimation en supposant que les p_s suivent une loi gamma et en estimant aussi ses paramètres. La souplesse de la loi gamma permet d'ajuster le modèle à des lois diverses et l'estimateur de Wang est très performant.

[44] J.-P. Wang. "Estimating Species Richness by a Poisson-compound Gamma Model." In: *Biometrika* 97.3 (2010), pp. 727–740. DOI: 10.1093/biomet/asq026.

Il est disponible dans *SPECIES*: fonction `pcg`. Son défaut est qu'il nécessite un très long temps de calcul (plusieurs heures selon les données).

```
# Calcul long
pcg(bci_abd_freq_count)
```

Inférence du nombre d'espèces à partir de courbes d'accumulation

Cette approche consiste à extrapoler la courbe d'accumulation observée.

Le modèle le plus connu est celui de Michaelis-Menten[45] proposé par Clench.[46] En fonction de l'effort d'échantillonnage n, évalué en temps (il s'agit de la collecte de papillons), le nombre d'espèces découvert augmente jusqu'à une asymptote égale au nombre d'espèces total:

[45] L. Michaelis and M. L. Menten. "Die Kinetik Der Invertinwirkung." In: *Biochemische Zeitschrift* 49 (1913), pp. 333–369. PMID: 21888353.

[46] H. K. Clench. "How to Make Regional Lists of Butterflies: Some Thoughts." In: *Journal of the Lepidopterists' Society* 33.4 (1979), pp. 216–231.

$$S(n) = S\frac{n}{K+n}. \tag{3.24}$$

K est une constante, que Clench relie à la difficulté de collecte.

L'estimation empirique du modèle de Michaelis-Menten peut être faite avec R[47]. Les 50 carrés de BCI sont utilisés pour fabriquer une courbe d'accumulation:

[47] Fiche TD de J.R. Lobry: http://pbil.univ-lyon1.fr/R/pdf/tdr47.pdf

```
# Cumul de l'inventaire
# Nombre d'arbres par espèce, cumulé par carré
bci_cumul_n_s <- apply(BCI, 2, cumsum)
# Nombre d'arbres cumulé par carré
bci_cumul_n <- cumsum(rowSums(BCI))
# Nombre total d'arbres
bci_n <- sum(BCI)
# Nombre d'espèces cumulées par carré
bci_cumul_S <- apply(bci_cumul_n_s, 1, function(x) sum(x>0))
```

Le modèle est ajusté par `nlsfit`. Des valeurs de départ doivent être fournies pour K et $\hat{S}$. K est la valeur de n correspondant à $\hat{S}^n = \hat{S}/2$. Une approximation suffisante est $n/4$. Pour $\hat{S}$, le nombre total d'espèces inventoriées est un bon choix. Le résultat se trouve en figure 3.7.

```
# Ajustement du modèle
(nlsfit <- nls(
  bci_cumul_S ~ S * bci_cumul_n / (K + bci_cumul_n),
  data = list(bci_cumul_n, bci_cumul_S),
  start = list(K = max(bci_cumul_n) / 4, S = max(bci_cumul_S))
))
```

```
## Nonlinear regression model
##   model: bci_cumul_S ~ S * bci_cumul_n/(K + bci_cumul_n)
##    data: list(bci_cumul_n, bci_cumul_S)
##      K      S
## 1251.0  232.8
##  residual sum-of-squares: 3066
##
## Number of iterations to convergence: 6
## Achieved convergence tolerance: 7.866e-06
```

L'estimation précédente utilise la méthode des moindres carrés, qui suppose l'indépendance des résidus, hypothèses évidemment violée par une courbe d'accumulation.[48] L'estimation par le maximum de vraisemblance est plus convenable.[49] Elle utilise la totalité des points de la courbe d'accumulation. La courbe d'accumulation de BCI est présentée en figure **??**. Ses données sont utilisées ici:

[48] Colwell and Coddington, "Estimating Terrestrial Biodiversity through Extrapolation," see n. 29, p. 34.

[49] J. G. W. Raaijmakers. "Statistical Analysis of the Michaelis-Menten Equation." In: *Biometrics* 43.4 (1987), pp. 793–803. DOI: 10.2307/2531533.

```
bci_sac <- specaccum(BCI, "random")
# Calculs intermédiaires
y_i <- bci_sac$richness
n <- length(y_i)
x_i <- y_i / (1:n)
x_bar <- mean(x_i)
y_bar <- mean(y_i)
S_yy <- sum((y_i - y_bar)^2)
```

```
S_xx <- sum((x_i - x_bar)^2)
S_xy <- sum((x_i - x_bar) * (y_i - y_bar))
# Estimations
(K_hat <- (x_bar * S_yy - y_bar * S_xy) / (y_bar * S_xx - x_bar * S_xy))
```

```
## [1] 1.712278
```

```
(S_hat <- y_bar + K_hat * x_bar)
```

```
## [1] 223.6211
```

L'estimation précédente repose sur une approximation numérique. Le paramètre K peut être estimé plus précisément par résolution numérique de l'équation exacte du maximum de vraisemblance:

```
# Equation que K_hat doit annuler
f <- function(K_hat) {
  S_xy +
    K_hat * S_xx -
    (S_yy + 2 * K_hat * S_xy + K_hat^2 * S_xx) * sum(x_i / (y_i + K_hat * x_i) / n)
}
# Résolution numérique, l'intervalle de recherche doit être fourni
solution <- uniroot(f, c(0, 1E+7))
(K_hat <- solution$root)
```

```
## [1] 1.712293
```

```
(S_hat <- y_bar + K_hat * x_bar)
```

```
## [1] 223.6213
```

Le nombre d'espèces estimé est 224, inférieur au nombre d'espèces observé.

Pour calculer l'intervalle de confiance, il est plus simple de passer par une transformation linéaire du modèle:[50]

$$\left[\frac{1}{\hat{S}(n)}\right] = \frac{K}{\hat{S}}\left[\frac{1}{n}\right] + \frac{1}{\hat{S}} \tag{3.25}$$

Le nombre d'espèces est estimé par l'inverse de l'ordonnée à l'origine du modèle.

[50] H. Lineweaver and D. Burk. "The Determination of Enzyme Dissociation Constants." In: *Journal of the American Chemical Society* 56.3 (1934), pp. 658–666. DOI: 10.1021/ja01318a036.

```
y <- 1 / bci_cumul_S
x <- 1 / bci_cumul_n
lm1 <- lm(y ~ x)
(S <- 1/lm1$coef[1])
```

```
## (Intercept)
##     217.002
```

On voit assez clairement que le modèle (figure 3.6) s'ajuste mal quand il est représenté sous cette forme.[51]

Le code R nécessaire pour réaliser la figure est:

[51] Raaijmakers, see n. 49.

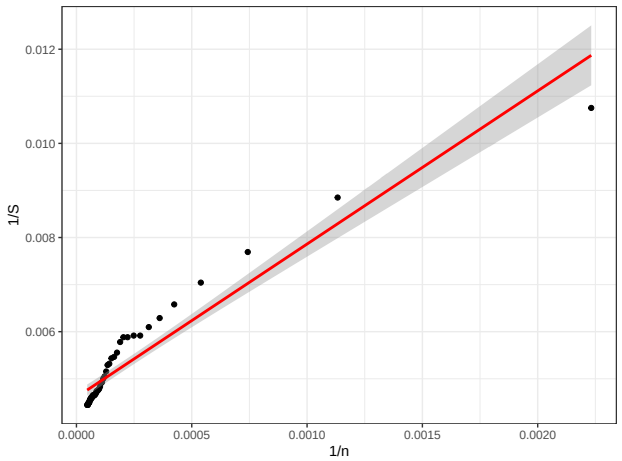


Figure 3.6: Ajustement du même modèle de Michaelis-Menten transformé selon Lineweaver et Burk.

```
tibble(x, y) %>%
  ggplot(aes(x, y)) +
    geom_point() +
    stat_smooth(method = "lm", col = "red") +
    labs(x = "1/n", y="1/S")
```

Le nombre d'espèces estimé est inférieur au nombre observé, qui ne se trouve même pas dans l'intervalle de confiance à 95%. Le modèle de Michaelis-Menten ne convient pas.

Soberón M. and Llorente B.[52] développent un cadre théorique plus vaste qui permet d'ajuster la courbe d'accumulation à plusieurs modèles. Ces modèles sont efficaces empiriquement mais manquent de support théorique pour justifier leur forme. Le modèle le plus simple est exponentiel négatif. Si la probabilité de trouver une nouvelle espèce est proportionnelle au nombre d'espèces non encore découvertes, la courbe d'accumulation suit la relation

$$S() = S\left(1 - e^{kn}\right). \tag{3.26}$$

Les paramètres peuvent être estimés par la méthode des moindres carrés:

```
(nlsexp <- nls(
  bci_cumul_S ~ S * (1 - exp(k * bci_cumul_n)),
  data = list(bci_cumul_n, bci_cumul_S),
  start = list(S = max(bci_cumul_S), k = -1 / 1000)
))
```

```
## Nonlinear regression model
##   model: bci_cumul_S ~ S * (1 - exp(k * bci_cumul_n))
##    data: list(bci_cumul_n, bci_cumul_S)
##          S          k
## 212.198081  -0.000494
##  residual sum-of-squares: 10664
##
## Number of iterations to convergence: 13
## Achieved convergence tolerance: 8.057e-06
```

[52] J. Soberón M. and J. Llorente B. "The Use of Species Accumulation Functions for the Prediction of Species Richness." In: *Conservation Biology* 7.3 (1993), pp. 480–488. DOI: 10.1046/j.1523-1739.1993.07030480.x.

[53] L. R. Holdridge et al. *Forest Environments in Tropical Life Zones*. Oxford: Pergamon Press, 1971.

Ce modèle, proposé par Holdridge et al.,[53] sous-estime la richesse parce que la probabilité de découvrir une nouvelle espèce diminue plus vite que le nombre d'espèces restant à découvrir: les dernières espèces sont plus rares et donc plus difficiles à détecter.

Un modèle plus réaliste définit cette probabilité comme une fonction décroissante du nombre d'espèces manquantes. La fonction la plus simple est une exponentielle négative mais elle ne s'annule jamais et le nombre d'espèces n'a pas d'asymptote. Un paramètre supplémentaire pour obtenir l'asymptote est nécessaire et aboutir à la relation

$$f = \frac{1}{z} \ln \left[\frac{a}{c} - \frac{a-c}{c} e^{-czn} \right]. \tag{3.27}$$

Les paramètres à estimer sont z, a et c.

```
(nlslog <- nls(
  bci_cumul_S ~ 1 / z * log(a / c - (a - c) / c * exp(-c * z * bci_cumul_n)),
  data = list(bci_cumul_n, bci_cumul_S),
  start = list(z = .05, a = 1, c = .001)
))
```

```
## Nonlinear regression model
##   model: bci_cumul_S ~ 1/z * log(a/c - (a - c)/c * exp(-c * z * bci_cumul_n))
##    data: list(bci_cumul_n, bci_cumul_S)
##         z         a         c
## 0.025139 0.755365 0.001114
##  residual sum-of-squares: 446.1
##
## Number of iterations to convergence: 4
## Achieved convergence tolerance: 4.754e-06
```

```
# Nombre d'espèces
coefs <- coef(nlslog)
log(coefs["a"] / coefs["c"]) / coefs["z"]
```

```
##        a
## 259.3127
```

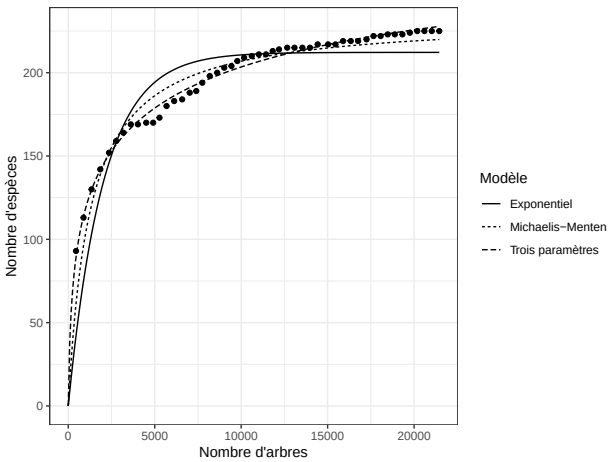


Figure 3.7: Ajustement des modèles de Michaelis-Menten et de de Soberón et Llorente (modèle exponentiel négatif et modèle à trois paramètres) aux données de BCI. Les points représentent le nombre d'espèces cumulées en fonction du nombre d'arbres. Le modèle exponentiel négatif (Holdridge) sous-estime la richesse, plus que celui de Michaelis-Menten (Clench). Le modèle à trois paramètres s'ajuste mieux aux données, mais il surestime probablement la richesse.

L'estimation est cette fois supérieure à celle du jackknife (244 espèces).

La figure 3.7 présente les deux ajustements de modèle de Soberón et Llorente avec celui de Clench. L'estimation de la richesse par extrapolation est plus incertaine que par les méthodes non paramétriques. Elle est très peu utilisée.

Le code R nécessaire pour réaliser la figure est:

```
x <- seq(from = 0, to = max(bci_cumul_n), length = 255)
x_new <- list(bci_cumul_n = x)
tibble(
  x,
  `Michaelis-Menten` = predict(nlsfit, newdata = x_new),
  `Exponentiel` = predict(nlsexp, newdata = x_new),
  `Trois paramètres` = predict(nlslog, newdata = x_new)
) %>%
  pivot_longer(cols = -x) %>%
  ggplot() +
    geom_line(aes(x = x, y = value, lty = name)) +
    geom_point(aes(bci_cumul_n, bci_cumul_S), data.frame(bci_cumul_n, bci_cumul_S)) +
    labs(
      x = "Nombre d'arbres",
      y = "Nombre d'espèces",
      lty = "Modèle"
    )
```

Diversité générique

La détermination des genres est plus facile et fiable que celle des espèces, le biais d'échantillonnage moins sensible (le nombre de singletons diminue rapidement en regroupant les données), et les coûts d'inventaire sont généralement largement réduits.[54] Le choix d'estimer la diversité de taxons de rang supérieur (genres ou même familles au lieu des espèces) est envisageable.[55]

Empiriquement, la corrélation entre la richesse générique et la richesse spécifique (des angiospermes, des oiseaux et des mammifères) est bonne en forêt tropicale,[56] suffisante pour comparer les communautés, même si la prédiction de la richesse spécifique à partir de la richesse générique est très imprécise.

Cartozo et al.[57] ont montré que le nombre de taxons de niveau supérieur (genre par rapport aux espèces, ordres par rapport aux sous-ordres) est universellement proportionnel au nombre de taxons du niveau immédiatement inférieur à la puissance 0,61. Cette relation est validée à l'échelle mondiale pour les systèmes végétaux. La loi de puissance reste valide pour des assemblages aléatoires, c'est donc la conséquence de propriétés mathématiques,[58] mais la puissance de la relation est plus élevée (les communautés réelles sont plus agrégées du point de vue phylogénétique que sous l'hypothèse nulle d'un assemblage aléatoire) et varie entre les niveaux.

[54] A. Balmford et al. "Using Higher-Taxon Richness as a Surrogate for Species Richness: II. Local Applications." In: *Proceedings of the Royal Society of London, Series B: Biological Sciences* 263 (1996), pp. 1571–1575. DOI: 10.1098/rspb.1996.0230.

[55] P. H. Williams and K. J. Gaston. "Measuring More of Biodiversity: Can Higher-Taxon Richness Predict Wholesale Species Richness?" In: *Biological Conservation* 67 (1994), pp. 211–217. DOI: 10.1016/0006-3207(94)90612-2.

[56] A. Balmford et al. "Using Higher-Taxon Richness as a Surrogate for Species Richness: I. Regional Tests." In: *Proceedings of the Royal Society of London, Series B: Biological Sciences* 263 (1996), pp. 1267–1274. DOI: 10.1098/rspb.1996.0186.

[57] C. C. Cartozo et al. "Quantifying the Taxonomic Diversity in Real Species Communities." In: *Journal of Physics A: Mathematical and Theoretical* 41 (2008), p. 224012. DOI: 10.1088/1751-8113/41/22/224012.

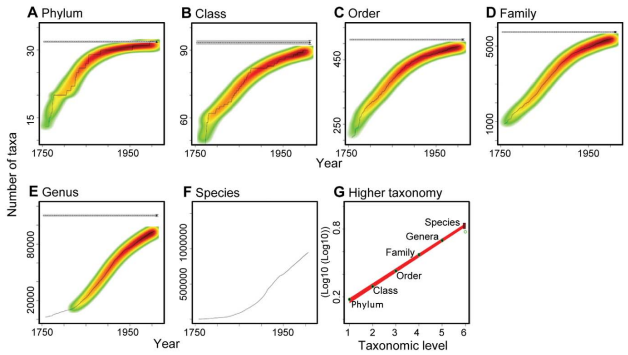
[58] G. Caldarelli et al. "Scale-Free Networks from Varying Vertex Intrinsic Fitness." In: *Physical Review Letters* 89.25 (2002), p. 258702. DOI: 10.1103/PhysRevLett.89.258702.

Combien y a-t-il d'espèces différentes sur Terre?

La question de l'estimation du nombre total d'espèces génère une abondante littérature. Mora et al.[59] en font une revue et proposent une méthode nouvelle.

Dans chaque règne, le nombre de taxons de niveaux supérieurs (phylums, classes, ordres, familles et même genres) est estimé par des modèles prolongeant jusqu'à leur asymptote les valeurs connues en fonction du temps. Cette méthode est applicable jusqu'au niveau du genre (figure **??**, A à E). Le nombre de taxon de chaque niveau est lié à celui du niveau précédent, ce qui est représenté par la figure **??**, G[60] sous la forme du relation linéaire entre le logarithme du logarithme du nombre de taxons et le rang (1 pour les phylums, 5 pour les genres). La droite est prolongée jusqu'au rang 6 pour obtenir le nombre d'espèces. Une façon alternative de décrire la méthode est de dire que le nombre de taxons du niveau $n+1$ est égal à celui du niveau n à la puissance k. La pente de la droite de la figure est $\ln k$. Aucune justification de ce résultat majeur n'est donnée par les auteurs, si ce n'est leur vérification empirique.

Le nombre total d'espèces estimé est 8,7 millions, tous règnes confondus, dans la fourchette des estimations précédentes (de 3 à 100 millions), et nécessitant près de 500 ans d'inventaires au rythme actuel des découvertes.[61]



En se limitant aux arbres, l'estimation se monte à 16000 espèces pour l'Amazonie,[62] de l'ordre de 5000 pour l'Afrique et entre 40000 et 53000 pour l'ensemble des tropiques[63] (donc pour l'ensemble de la planète, le nombre d'espèces non-tropicales étant négligeable). Ces estimations sont obtenues par extrapolation du modèle en log-séries (chapitre **??**) et sont sujettes au paradoxe de Fisher: les espèces représentées par un très petit nombre d'individu dans le modèle, notamment les singletons, sont les plus nombreuses. Une discussion approfondie est donnée par Hubbell:[64] les espèces récemment apparues au sens du modèle ne sont pas détectables avant plusieurs générations, créant un décalage

[59] C. Mora et al. "How Many Species Are There on Earth and in the Ocean?" In: *PLoS Biology* 9.8 (2011), e1001127. DOI: 10.1371/journal.pbio.1001127.

[60] Ibid., figure 1.

[61] R. M. May. "Why Worry about How Many Species and Their Loss?" In: *PLoS Biology* 9.8 (2011), e1001130. DOI: 10.1371%2Fjournal.pbio.1001130.

[62] H. ter Steege et al. "Hyperdominance in the Amazonian Tree Flora." In: *Science* 342.6156 (2013), p. 1243092. DOI: 10.1126/science.1243092.

[63] J. W. F. Slik et al. "An Estimate of the Number of Tropical Tree Species." In: *Proceedings of the National Academy of Sciences of the United States of America* 112.24 (2015), pp. 7472–7477. DOI: 10.1073/pnas.1423147112.

[64] S. P. Hubbell. "Estimating the Global Number of Tropical Tree Species, and Fisher's Paradox." In: *Proceedings of the National Academy of Sciences* 112.24 (2015), pp. 7343–7344. DOI: 10.1073/pnas.1507730112.

entre le nombre d'espèces reconnues par la taxonomie et le modèle.

J. B. Wilson et al.[65] compilent les relevés du nombre d'espèces de plantes vasculaires en fonction de la surface et retiennent uniquement les plus riches à chaque échelle spatiale (du millimètre carré à l'hectare). Ces relevés sont tous situés en forêt tropicale ou en prairie tempérée gérée (les perturbations régulières et modérées y favorisent la diversité, conformément à la théorie de la perturbation intermédiaire[66]). La relation entre le nombre d'espèces et la surface est celle d'Arrhenius. Son extrapolation à la surface terrestre donne environ 220000 espèces, comparables à l'estimation de 275000 espèces rapportée par Mora et al.

[65] J. B. Wilson et al. "Plant Species Richness: The World Records." In: *Journal of Vegetation Science* 23.4 (2012), pp. 796–802. DOI: 10.1111/j.1654-1103.2012.01400.x.

[66] J. H. Connell. "Diversity in Tropical Rain Forests and Coral Reefs." In: *Science* 199.4335 (1978), pp. 1302–1310. DOI: 10.1126/science.199.4335.1302.

3.2 Indice de Simpson

Définition

On note p_s la probabilité qu'un individu tiré au hasard appartienne à l'espèce s. L'indice de Simpson,[67] ou Gini-Simpson, est

$$E = 1 - \sum_{s=1}^{S} p_s^2. \tag{3.28}$$

Il peut être interprété comme la probabilité que deux individus tirés au hasard soient d'espèces différentes. Il est compris dans l'intervalle $[0; 1[$. Sa valeur diminue avec la régularité de la distribution: $E = 0$ si une seule espèce a une probabilité de 1, $E = 1 - 1/S$ si les S espèces ont la même probabilité $p_s = 1/S$. La valeur 1 est atteinte pour un nombre infini d'espèces, de probabilités nulles.

Deux autres formes de l'indice sont utilisées. Tout d'abord, la probabilité que deux individus soient de la même espèce, souvent appelée *indice de concentration de Simpson*, qui est celui défini dans l'article original de Simpson:

$$D = \sum_{s=1}^{S} p_s^2. \tag{3.29}$$

L'indice de Simpson est parfois considéré comme une mesure d'équitabilité[68] mais il varie avec la richesse: cette approche est donc erronée. S. H. Hurlbert[69] l'a divisé par sa valeur maximale $1 - 1/S$ pour obtenir une mesure d'équitabilité valide généralisée plus tard par Mendes et al.,[70] voir section **??**. Le nombre d'espèces doit être estimé par les méthodes présentées plus haut, pour ne pas dépendre de la taille de l'échantillon.

[67] E. H. Simpson. "Measurement of Diversity." In: *Nature* 163.4148 (1949), p. 688. DOI: 10.1038/163688a0.

[68] T. D. Olszewski. "A Unified Mathematical Framework for the Measurement of Richness and Evenness within and among Multiple Communities." In: *Oikos* 104.2 (2004), pp. 377–387. DOI: 10.1111/j.0030-1299.2004.12519.x.

[69] S. H. Hurlbert. "The Nonconcept of Species Diversity: A Critique and Alternative Parameters." In: *Ecology* 52.4 (1971), pp. 577–586. DOI: 10.2307/1934145.

[70] R. S. Mendes et al. "A Unified Index to Measure Ecological Diversity and Species Rarity." In: *Ecography* 31.4 (2008), pp. 450–456. DOI: 10.1111/j.0906-7590.2008.05469.x.

L'estimateur du maximum de vraisemblance de l'indice est

$$\hat{E} = 1 - \sum_{s=1}^{f_{>0}} \hat{p}_s^2. \tag{3.30}$$

Le calcul de l'indice de Simpson peut se faire avec la fonction `diversity` disponible dans le package *vegan* de R ou avec la fonction `ent_simpson` du package *divent*, qui peut traiter plusieurs sites en même temps:

```
paracou_6_abd %>%
  # Transformation des abondances en probabilités
  as_probabilities() %>%
  ent_simpson()
```

```
## # A tibble: 4 x 5
##   site      weight estimator order entropy
##   <chr>      <dbl> <chr>     <dbl>   <dbl>
## 1 subplot_1   1.56 naive         2   0.975
## 2 subplot_2   1.56 naive         2   0.976
## 3 subplot_3   1.56 naive         2   0.978
## 4 subplot_4   1.56 naive         2   0.971
```

Un historique de la définition de l'indice, de Gini[71] à Simpson, inspiré par Turing, est fourni par Ellerman.[72]

[71] C. Gini. *Variabilità e Mutabilità*. Bologna: C. Cuppini, 1912.

[72] D. Ellerman. "An Introduction to Logical Entropy and Its Relation to Shannon Entropy." In: *International Journal of Semantic Computing* 7.02 (2013), pp. 121–145. DOI: 10.1142/S1793351X13400059.

Estimation

Définissons l'indicatrice $\mathbf{1}_{sh}$ valant 1 si l'individu h appartient à l'espèce s, 0 sinon. $\mathbf{1}_{sh}$ suit une loi de Bernoulli d'espérance p_s et de variance $p_s(1-p_s)$. E est la somme sur toutes les espèces de cette variance. Un estimateur non biaisé d'une variance à partir d'un échantillon est la somme des écarts quadratiques divisée par le nombre d'observation moins une. L'estimateur $\hat{E}$ est légèrement biaisé parce qu'il est calculé à partir des $\hat{p}_s$, ce qui revient à diviser la somme des écarts par n, et non $n-1$. Un estimateur non biaisé est[73]

$$\tilde{E} = \left(\frac{n}{n-1}\right)\left(1 - \sum_{s=1}^{f_{>0}} \hat{p}_s^2\right). \tag{3.31}$$

[73] Good, "The Population Frequency of Species and the Estimation of Population Parameters," see n. 16, p. 9; R. Lande. "Statistics and Partitioning of Species Diversity, and Similarity among Multiple Communities." In: *Oikos* 76.1 (1996), pp. 5–13. DOI: 10.2307/3545743.

La correction par $n/(n-1)$ tend rapidement vers 1 quand la taille de l'échantillon augmente: l'estimateur est très peu biaisé.

Le non-échantillonnage des espèces rares est pris en compte dans cette correction parce qu'elle considère que $\tilde{E}$ est l'estimateur de variance d'un échantillon et non d'une population complètement connue. Il est négligeable: si p_s est petit, p_s^2 est négligeable dans la somme.

Simpson a fourni un estimateur non biaisé de D, à partir du calcul du nombre de paires d'individus tirés sans remise:

$$\tilde{D} = \frac{\sum_{s=1}^{S} n_s (n_s - 1)}{n (n - 1)}. \tag{3.32}$$

L'argumentation est totalement différente, mais le résultat est le même: $\tilde{E} = 1 - \tilde{D}$.

La fonction `ent_simpson` de *divent* accepte comme argument un vecteur d'abondances ou un dataframe contenant les données et propose par défaut la correction de Lande:

```
paracou_6_abd %>%
  ent_simpson()
```

```
## # A tibble: 4 x 5
##   site      weight estimator order entropy
##   <chr>      <dbl> <chr>     <dbl>   <dbl>
## 1 subplot_1   1.56 Lande         2   0.976
## 2 subplot_2   1.56 Lande         2   0.978
## 3 subplot_3   1.56 Lande         2   0.980
## 4 subplot_4   1.56 Lande         2   0.972
```

3.3 Indice de Shannon

Définition

L'indice de Shannon,[74] aussi appelé indice de Shannon-Weaver ou Shannon-Wiener,[75] ou simplement *entropie* est dérivé de la théorie de l'information:

$$H = -\sum_{s=1}^{S} p_s \ln p_s. \tag{3.33}$$

Considérons une placette forestière contenant S espèces végétales différentes. La probabilité qu'une plante choisie au hasard appartienne à l'espèce s est notée p_s. On prélève n plantes, et on enregistre la liste ordonnée des espèces des n plantes. Si n est suffisamment grand, le nombre de plantes de l'espèce s est np_s. On note L le nombre de listes respectant ces conditions:

$$L = \frac{n!}{\prod_{i=1}^{S} (np_s)!}. \tag{3.34}$$

Ce résultat est obtenu en calculant le nombre de positions possibles dans la liste pour les individus de la première espèce: $\binom{n}{np_1}$. Le nombre de positions pour la deuxième espèce est $\binom{n-np_1}{np_2}$. Pour la S-ième espèce, le nombre est $\binom{n-np_1-\cdots-np_{s-1}}{np_i}$. Les produits de combinaisons se simplifient pour donner l'équation (3.34).

[74] C. E. Shannon. "A Mathematical Theory of Communication." In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423, 623–656. DOI: 10.1002/j.1538-7305.1948.tb01338.x; C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1963.

[75] I. F. Spellerberg and P. J. Feder. "A Tribute to Claude Shannon (1916–2001) and a Plea for More Rigorous Use of Species Richness, Species Diversity and the 'Shannon–Wiener' Index." In: *Global Ecology and Biogeography* 12.3 (2003), pp. 177–179. DOI: 10.1046/j.1466-822X.2003.00015.x.

On peut maintenant écrire le logarithme de L:

$$\ln L = \ln n! - \sum_{s=1}^{S} \ln np_s!.$$

On utilise l'approximation de Stirling,

$$\ln n! \approx n \ln n - n,$$

pour obtenir après simplifications:

$$\ln L = -n \sum_{s=1}^{S} p_s \ln p_s. \tag{3.35}$$

$H = (\ln L)/n$ est l'indice de Shannon. Ce résultat est connu sous le nom de formule de Brillouin.[76] À l'origine, Shannon a utilisé un logarithme de base 2 pour que H soit le nombre moyen de questions binaires (réponse oui ou non) nécessaire pour identifier l'espèce d'une plante (un caractère utilisé dans une chaîne dans le contexte du travail de Shannon). Les logarithmes naturels, de base 2 ou 10 ont été utilisés par la suite.[77]

[76] L. Brillouin. *Science and Information Theory*. 2nd ed. Oxford: Academic Press, 1962.

La formule (3.35) est celle de l'indice de Theil,[78] présenté en détail par Conceição and Ferreira,[79] à l'origine utilisé pour mesurer les inégalités de revenu puis pour caractériser les structures spatiales en économie. L'indice est proportionnel au nombre de plantes choisies, on peut donc le diviser par n et on obtient l'indice de biodiversité de Shannon. Ces indices ont été définis en choisissant des lettres au hasard pour former des chaînes de caractères. Leur valeur est le nombre de chaînes de caractères différentes que l'on peut obtenir avec l'ensemble des lettres disponibles, c'est-à-dire la quantité d'information contenue dans l'ensemble des lettres. L'indice de Shannon donne une mesure de la biodiversité en tant que quantité d'information.

[77] E. C. Pielou. "The Measurement of Diversity in Different Types of Biological Collections." In: *Journal of Theoretical Biology* 13.C (1966), pp. 131–144. DOI: 10.1016/0022-5193(66)90013-0.

[78] H. Theil. *Economics and Information Theory*. Chicago: Rand McNally & Company, 1967.

[79] P. Conceição and P. Ferreira. *The Young Person's Guide to the Theil Index: Suggesting Intuitive Interpretations and Exploring Analytical Applications*. Austin, Texas, 2000, p. 54.

L'estimateur du maximum de vraisemblance de l'indice est

$$\hat{H} = -\sum_{s=1}^{f_{>0}} \hat{p}_s \ln \hat{p}_s. \tag{3.36}$$

Le calcul de l'indice de Shannon peut se faire avec la fonction `diversity` disponible dans le package *vegan* de R ou avec la fonction `ent_shannon` de *divent*:

```
bci_abd %>%
  as_probabilities() %>%
  ent_shannon()
```

```
## # A tibble: 1 x 5
##   site   weight estimator order entropy
##   <chr>   <dbl> <chr>     <dbl>   <dbl>
## 1 site_1      1 naive         1    4.27
```

La distribution de l'estimateur est connue[80] mais elle est inutile en pratique à cause du biais d'estimation.

Bulmer[81] établit une relation entre l'indice de Shannon et l'indice α de Fisher, à condition que la distribution de l'abondance des espèces soit log-normale:

$$\hat{H} = \quad (\hat{\alpha} + 1) - \quad (1)\,. \tag{3.37}$$

$(\cdot)$ est la fonction digamma, et $\hat{\alpha}$ est l'estimateur de l'indice de Fisher (**??**):

```
digamma(fisher.alpha(colSums(BCI)) + 1) - digamma(1)
```

```
## [1] 4.148323
```

La sous-estimation est assez sévère sur cet exemple.

Estimation

Basharin[82] a montré que l'estimateur naïf de l'indice de Shannon était biaisé parce que des espèces ne sont pas échantillonnées. Si S est le nombre d'espèces réel et n le nombre d'individus échantillonnés, le biais est

$$\mathbb{E}\left(\hat{H}\right) - H = -\frac{S-1}{2n} + O\left(n^{-2}\right). \tag{3.38}$$

$O(n^{-2})$ est un terme négligeable. La valeur estimée à partir des données est donc trop faible, d'autant plus que le nombre d'espèces total est grand mais d'autant moins que l'échantillonnage est important. Comme le nombre d'espèces S n'est pas observable, le biais réel est inconnu.

L'estimateur de Miller-Madow[83] utilise l'information disponible, en sous-estimant le nombre d'espèces et donc l'entropie:

$$\tilde{H} = -\sum_{s=1}^{f_{>0}} \hat{p}_s \ln \hat{p}_s + \frac{f_{>0} - 1}{2n}. \tag{3.39}$$

Chao and Shen[84] établissent un estimateur moins biaisé à partir du taux de couverture de l'échantillonnage $\hat{C}$:

$$\tilde{H} = -\sum_{s=1}^{f_{>0}} \frac{\hat{C}\hat{p}_s \ln\left(\hat{C}\hat{p}_s\right)}{1 - \left(1 - \hat{C}\hat{p}_s\right)^n}. \tag{3.40}$$

Multiplier les fréquences observées par le taux de couverture permet d'obtenir un estimateur non biaisé des probabilités conditionnellement aux espèces non observées.[85]

Le terme au dénominateur est la correction de Horvitz and Thompson:[86] chaque terme de la somme est divisé par la prob-

[80] K. Hutcheson. "A Test for Comparing Diversities Based on the Shannon Formula." In: *Journal of Theoretical Biology* 29 (1970), pp. 151–154. DOI: 10.1016/0022-5193(70)90124-4.

[81] Bulmer, "On Fitting the Poisson Lognormal Distribution to Species-Abundance Data," see n. 13, p. 17.

[82] G. P. Basharin. "On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables." In: *Theory of Probability and its Applications* 4.3 (1959), pp. 333–336. DOI: 10.1137/1104033.

[83] G. A. Miller. "Note on the Bias of Information Estimates." In: *Information Theory in Psychology: Problems and Methods*. Ed. by H. Quastler. Glencoe, Ill.: Free Press, 1955, pp. 95–100.

[84] A. Chao and T.-J. Shen. "Nonparametric Estimation of Shannon's Index of Diversity When There Are Unseen Species in Sample." In: *Environmental and Ecological Statistics* 10.4 (2003), pp. 429–443. DOI: 10.1023/A:1026096204727.

[85] J. Ashbridge and I. B. J. Goudie. "Coverage-Adjusted Estimators for Mark-Recapture in Heterogeneous Populations." In: *Communications in Statistics - Simulation and Computation* 29.4 (2000), pp. 1215–1237. DOI: 10.1080/03610910008813661.

[86] D. G. Horvitz and D. J. Thompson. "A Generalization of Sampling without Replacement from a Finite Universe." In: *Journal of the American Statistical Association* 47.260 (1952), pp. 663–685. DOI: 10.1080/01621459.1952.10483446.

abilité d'observer au moins une fois l'espèce correspondante. Il tend vers 1 quand la taille de l'échantillon augmente.

Beck and Schwanghart[87] montrent que la correction du biais est efficace, même à des niveaux de complétude de l'échantillonnage (voir section 3.1) très faibles. Vu et al.[88] étudient la vitesse de convergence de l'estimateur.

Z. Zhang[89] définit l'indice de Simpson généralisé:

$$\zeta_{u,v} = \sum_{s=1}^{S} p_s^u (1 - p_s)^v, \tag{3.41}$$

où $\zeta_{u,v}$ est la somme sur toutes les espèces de la probabilité de rencontrer u fois l'espèce dans un échantillon de taille $u+v$. L'indice de Shannon peut s'exprimer en fonction de $\zeta_{1,v}$:

$$H = \sum_{v=1}^{\infty} \frac{1}{v} \zeta_{1,v}. \tag{3.42}$$

Les premiers termes de la somme, jusqu'à $v = n-1$ peuvent être estimés à partir des données, les suivants constituent le biais de l'estimateur, qui est calculé en pratique par

$$H_z = \sum_{v=1}^{n-1} \frac{1}{v} \left\{ \frac{n^{v+1} [n - (v+1)]!}{n!} \sum_{s=1}^{f_{>0}} p_s \prod_{j=0}^{v-1} \left(1 - \hat{p}_s - \frac{j}{n}\right) \right\}. \tag{3.43}$$

Z. Zhang[90] montre que le biais de l'estimateur H_z est asymptotiquement normal et calcule sa variance. Z. Zhang and Grabchak[91] améliorent l'estimateur en le complétant par un estimateur de son biais, mais les calculs deviennent excessivement complexes. Vinck et al.[92] appliquent la même démarche avec un estimateur bayésien du biais, utilisant un prior aussi plat que possible pour la valeur de l'entropie (et non un prior plat sur les probabilités, qui tire l'estimateur vers l'entropie maximale). Cet estimateur nécessite de connaître le nombre d'espèces, ce qui empêche son utilisation sur des données d'écologie.

Pielou[93] a développé une autre méthode de correction de biais lorsque de nombreux relevés de petite taille sont disponibles. $\ln L$ est calculé pour un relevé choisi aléatoirement puis les données du premier relevé sont ajoutées à celles d'un autre, puis un autre jusqu'à ce que $H = (\ln L)/n$ n'augmente plus: la diversité augmente dans un premier temps mais se stabilise quand l'effet des espèces ajoutées est compensé par celui de la diminution de l'équitabilité due aux espèces présentes dans tous les relevés. À partir de ce seuil, l'augmentation de $\ln L$ par individu ajouté est calculée pour chaque relevé supplémentaire. Son espérance, estimée par sa

[87] Beck and Schwanghart, "Comparing Measures of Species Diversity from Incomplete Inventories: An Update," see n. 33, p. 36.

[88] V. Q. Vu et al. "Coverage-Adjusted Entropy Estimation." In: *Statistics in Medicine* 26.21 (2007), pp. 4039–4060. DOI: 10.1002/sim.2942.

[89] Z. Zhang. "Entropy Estimation in Turing's Perspective." In: *Neural Computation* 24.5 (2012), pp. 1368–1389. DOI: 10.1162/NECO_a_00266.

[90] Z. Zhang. "Asymptotic Normality of an Entropy Estimator with Exponentially Decaying Bias." In: *IEEE Transactions on Information Theory* 59.1 (2013), pp. 504–508. DOI: 10.1109/TIT.2012.2217393.

[91] Z. Zhang and M. Grabchak. "Bias Adjustment for a Nonparametric Entropy Estimator." In: *Entropy* 15.6 (2013), pp. 1999–2011. DOI: 10.3390/e15061999.

[92] M. Vinck et al. "Estimation of the Entropy Based on Its Polynomial Representation." In: *Physical Review E* 85.5 (2012). DOI: 10.1103/PhysRevE.85.051139.

[93] E. C. Pielou. "Species-Diversity and Pattern-Diversity in the Study of Ecological Succession." In: *Journal of Theoretical Biology* 10.2 (1966), pp. 370–383. DOI: 10.1016/0022-5193(66)90133-0.

moyenne calculée en ajoutant tous les relevés disponibles, est $\tilde{H}$.

Chao et al.[94] utilisent l'estimateur de la pente de la courbe de raréfaction, calculé précédemment[95] pour estimer la richesse spécifique, pour fournir un estimateur extrêmement performant:

$$\tilde{H} = -\sum_{s=1}^{f_{>0}} \frac{n_s}{n} \left(\quad (n) - \quad (n_s) \right) \tag{3.44}$$

$$- \frac{s_1}{n}(1-A)^{1-n} \left(-\ln(A) - \sum_{r=1}^{n-1} \frac{1}{r}(1-A)^r \right), \tag{3.45}$$

où $(\cdot)$ est la fonction digamma et A vaut:

- $2s_2/[(n-1)s_1 + 2s_2]$ en présence de singletons et doubletons;
- $2/[(n-1)(s_1-1)+2]$ en présence de singletons seulement;
- 1 en absence de singletons et doubletons.

Enfin, la littérature de physique statistique s'est abondamment intéressée à cette question (Bonachela et al.[96] en font une revue). Le problème traité est la non-linéarité de l'indice de Shannon par rapport aux probabilités qui entraîne un biais d'estimation. La fonction logarithme fournit un exemple simple: l'espérance de $\ln(p_s)$ n'est pas le logarithme de l'espérance de p_s parce que la fonction ln est concave. Chaque estimateur $\hat{p}_s$ fluctue autour de p_s mais vaut p_s en moyenne. À cause de la concavité, $\ln(\hat{p}_s)$ est en moyenne inférieur à $\ln(p_s)$: cette relation est connue sous le nom d'inégalité de Jensen.[97] L'indice de Shannon est concave (figure 3.8) donc son estimateur (3.33) est biaisé négativement, même sans prendre en considération les espèces non observées.

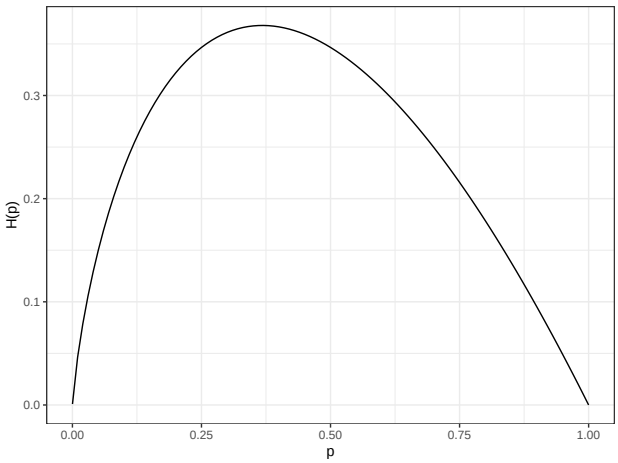


Figure 3.8: Courbe de $x \ln x$ entre 0 et 1.

Code de la figure 3.8:

[94] A. Chao et al. "Entropy and the Species Accumulation Curve: A Novel Entropy Estimator via Discovery Rates of New Species." In: *Methods in Ecology and Evolution* 4.11 (2013), pp. 1091–1100. DOI: 10.1111/2041-210x.12108.

[95] Chao and Jost, "Coverage-Based Rarefaction and Extrapolation: Standardizing Samples by Completeness Rather than Size," see n. 17, p. 9.

[96] J. A. Bonachela et al. "Entropy Estimates of Small Data Sets." In: *Journal of Physics A: Mathematical and Theoretical* 41.202001 (2008), pp. 1–9. DOI: 10.1088/1751-8113/41/20/202001.

[97] J. L. W. V. Jensen. "Sur les fonctions convexes et les inégalités entre les valeurs moyennes." In: *Acta Mathematica* 30.1 (1906), pp. 175–193. DOI: 10.1007/bf02418571.

```
tibble(x = c(0.0001, 1)) %>%
  ggplot(aes(x)) +
    stat_function(fun = function(x) -x * log(x)) +
    labs(x = "p", y = "H(p)")
```

Le biais peut être évalué par simulation: 10000 tirages sont réalisés dans une loi normale d'espérance p_s choisie et d'écart-type 0.01. Le biais est la différence entre $-p_s \ln p_s$ (connu) et la moyenne des 1000 valeurs de $-\hat{p}_s \ln \hat{p}_s$ (la probabilité est estimée par sa réalisation à chaque tirage). La valeur du biais en fonction de p_s est en figure 3.9. Le biais de l'indice de Shannon est la somme des biais pour toutes les probabilités spécifiques de la communauté étudiée, et son calcul est toujours l'objet de recherches.

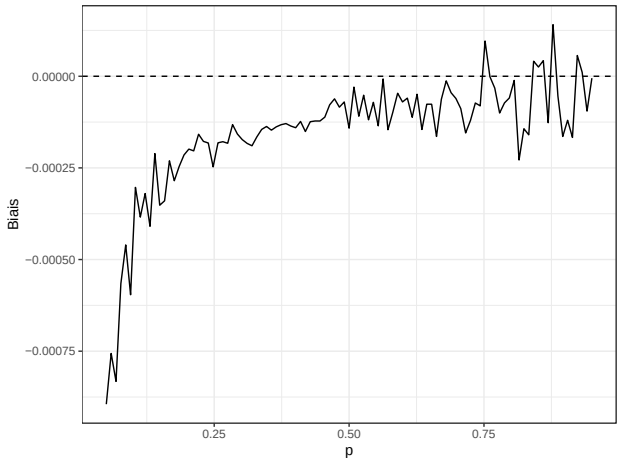


Figure 3.9: Biais de $\hat{p}_s \ln \hat{p}_s$ entre 0 et 1.

Code de la figure 3.9:

```
bias_p <- function (p) {
  p_s <-rnorm(10000, p, 0.01)
  p * log(p) - mean(p_s * log(p_s))
}

bias <- function (p_s) {
  # Applique bias_p à chaque valeur de p_s
  sapply(p_s, bias_p)
}

tibble(x = c(0.05, 0.95)) %>%
  ggplot(aes(x)) +
    stat_function(fun = bias) +
    geom_hline(yintercept = 0, lty = 2) +
    labs(x = "p", y = "Biais")
```

Grassberger[98] a fourni la correction de référence:

$$\tilde{H} = -\sum_{s=1}^{f_{>0}} \frac{n_s}{n} \left(\ln(n) - \quad (n_s) - \frac{(-1)^{n_s}}{n_s + 1} \right). \tag{3.46}$$

Grassberger[99] l'a perfectionnée:

[98] P. Grassberger. "Finite Sample Corrections to Entropy and Dimension Estimates." In: *Physics Letters A* 128.6-7 (1988), pp. 369–373. DOI: 10.1016/0375-9601(88)90193-4.

[99] P. Grassberger. "Entropy Estimates from Insufficient Samplings." In: *arXiv Physics e-prints* 0307138.v2 (2003).

$$\tilde{H} = -\sum_{s=1}^{f_{>0}} \frac{n_s}{n} \left((n) - (n_s) - (-1)^{n_s} \int_0^1 \frac{t^{n_s-1}}{1+t}\, dt \right). \quad (3.47)$$

Enfin, Schürmann[100] l'a généralisée pour définir une famille de corrections dépendant d'un paramètre ξ:

$$\tilde{H} = -\sum_{s=1}^{f_{>0}} \frac{n_s}{n} \left((n) - (n_s) - (-1)^{n_s} \int_0^{\frac{1}{\xi}-1} \frac{t^{n_s-1}}{1+t}\, dt \right). \quad (3.48)$$

Le biais d'estimation diminue avec ξ mais l'erreur quadratique augmente. Schürmann suggère d'utiliser $\xi = e^{-1/2}$ comme meilleur compromis.

La fonction `ent_shannon` permet toutes ces corrections.

```
ent_shannon(colSums(BCI), estimator = "ChaoJost")
```

```
## # A tibble: 1 x 3
##   estimator order entropy
##   <chr>     <dbl>   <dbl>
## 1 ChaoJost      1    4.28
```

```
ent_shannon(colSums(BCI), estimator = "Grassberger")
```

```
## # A tibble: 1 x 3
##   estimator   order entropy
##   <chr>       <dbl>   <dbl>
## 1 Grassberger     1    4.28
```

```
ent_shannon(colSums(BCI), estimator = "Grassberger2003")
```

```
## # A tibble: 1 x 3
##   estimator       order entropy
##   <chr>           <dbl>   <dbl>
## 1 Grassberger2003     1    4.28
```

```
ent_shannon(colSums(BCI), estimator = "Schurmann")
```

```
## # A tibble: 1 x 3
##   estimator order entropy
##   <chr>     <dbl>   <dbl>
## 1 Schurmann     1    4.28
```

```
ent_shannon(colSums(BCI), estimator = "ZhangHz")
```

```
## # A tibble: 1 x 3
##   estimator order entropy
##   <chr>     <dbl>   <dbl>
## 1 ZhangHz       1    4.27
```

[100] T. Schürmann. "Bias Analysis in Entropy Estimation." In: *Journal of Physics A: Mathematical and General* 37.27 (2004), pp. L295–L301. DOI: 10.1088/0305-4470/37/27/L02.

[101] J. Hausser and K. Strimmer. "Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks." In: *Journal of Machine Learning Research* 10 (2009), pp. 1469–1484.

[102] W. James and C. Stein. "Estimation with Quadratic Loss." In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Ed. by J. Neyman. Vol. 1. Berkeley, California: University of California Press, 1961, pp. 361–379.

D'autres estimateurs peu utilisés en écologie sont disponibles dans le package *entropy*.[101] La contraction de Stein[102] consiste à estimer la distribution des probabilités d'occurrence des espèces par la pondération optimale entre un estimateur à faible biais et un estimateur à faible variance. L'estimateur $\hat{p}_s$ est sans biais mais a une variance importante. L'estimateur $1/S$, si S est connu, est de variance nulle mais est très biaisé. Comme le nombre d'espèces est en général inconnu, il doit être estimé par une méthode quelconque, mais de préférence le surestimant plutôt que le sous-estimant. L'estimateur de James-Stein (*shrinkage estimator*) optimal est

$$\tilde{p}_s = \hat{\lambda}\frac{1}{\hat{S}} + \left(1-\hat{\lambda}\right)\hat{p}_s, \tag{3.49}$$

où

$$\hat{\lambda} = \frac{1-\sum_{s=1}^{f_{>0}}\left(\hat{p}_s\right)^2}{(n-1)\sum_{s=1}^{f_{>0}}\left(\frac{1}{\hat{S}}-\hat{p}_s\right)^2}. \tag{3.50}$$

L'entropie est ensuite simplement estimée par l'estimateur plug-in: $\tilde{H} = -\sum_{s=1}^{f_{>0}}\tilde{p}_s \ln \tilde{p}_s$. Le calcul sous R est le suivant:

```
library("entropy")
entropy.shrink(colSums(BCI))
```

```
## Estimating optimal shrinkage intensity lambda.freq (frequencies): 0.0021
```

```
## [1] 4.275689
## attr(,"lambda.freqs")
## [1] 0.002074043
```

Le principe même de l'estimation rapproche la distribution de l'équiprobabilité des espèces et donc augmente l'entropie. L'estimation précédente ignore les espèces non observées. Pour les inclure, le vecteur des abondance doit être allongé par autant de zéros que d'espèces estimées:

```
entropy.shrink(c(colSums(BCI), rep(0, bci_f_0)))
```

```
## Estimating optimal shrinkage intensity lambda.freq (frequencies): 0.002
```

```
## [1] 4.276543
## attr(,"lambda.freqs")
## [1] 0.002041748
```

Les fréquences sont estimées par la fonction `freqs.shrink`. Leur utilisation dans l'estimateur plug-in donne le même résultat:

```
ent_shannon(freqs.shrink(c(colSums(BCI), rep(0, bci_f_0))))
```

```
## Estimating optimal shrinkage intensity lambda.freq (frequencies): 0.002
```

```
## # A tibble: 1 x 3
##   estimator order entropy
##   <chr>     <dbl>   <dbl>
## 1 naive         1    4.28
```

Appliqué à des données de biodiversité aquatique, l'estimateur de James-Stein obtient de meilleurs résultats que celui de Chao et Shen[103] quand l'échantillonnage est réduit.

[103] D. Liu et al. "Entropy of Hydrological Systems under Small Samples: Uncertainty and Variability." In: *Journal of Hydrology* 532 (2016), pp. 163–176. DOI: 10.1016/j.jhydrol.2015.11.019.

3.4 Indice de Hurlbert

Définition

L'indice de S. H. Hurlbert[104] est l'espérance du nombre d'espèces observées dans un échantillon de taille k choisie:

[104] S. H. Hurlbert, "The Nonconcept of Species Diversity: A Critique and Alternative Parameters," see n. 69, p. 48.

$$ {}_kS = \sum_{s=1}^{S} \left[1 - (1 - p_s)^k \right]. \tag{3.51} $$

Chaque terme de la somme est la probabilité d'observer au moins une fois l'espèce correspondante.

L'augmentation de la valeur de k permet de donner plus d'importance aux espèces rares.

L'indice peut être converti en nombre équivalent d'espèces,[105] c'est-à-dire le nombre d'espèces équiprobables nécessaire pour obtenir la même diversité, notés ${}_kD$ à partir de la relation

[105] Dauby and Hardy, "Sampled-Based Estimation of Diversity Sensu Stricto by Transforming Hurlbert Diversities into Effective Number of Species," see n. 18, p. 9.

$$ {}_kS = {}_kD \left[1 - \left(1 - \frac{1}{{}_kD} \right)^k \right]. \tag{3.52} $$

L'équation doit être résolue pour obtenir ${}_kD$ à partir de la valeur de ${}_kS$ estimée, numériquement pour $k > 3$.

Dans deux cas particuliers, les nombres équivalents d'espèces de Hurlbert sont identiques aux nombres de Hill: ${}_2D = {}^2D$ et ${}_\infty D = {}^0D$.

Pour les autres ordres entiers de diversité, il existe une correspondance parfaite entre les deux mesures:[106] la connaissance de l'une permet d'obtenir l'autre. Pour $k > 1$, on a[107]

[106] A. Chao et al. "Rarefaction and Extrapolation with Hill Numbers: A Framework for Sampling and Estimation in Species Diversity Studies." In: *Ecological Monographs* 84.1 (2014), pp. 45–67. DOI: 10.1890/13-0133.1, Annexe S2.

[107] T. Leinster and C. Cobbold. "Measuring Diversity: The Importance of Species Similarity." In: *Ecology* 93.3 (2012), pp. 477–489. DOI: 10.1890/10-2402.1.

$$ {}_kS = k + \sum_{q=2}^{k} \binom{k}{q} (-1)^{q+1} \left({}^qD\right)^{1-q}. \tag{3.53} $$

On a vu que ${}^0D$ égale ${}_\infty D$, donc ${}_\infty S = {}^0D$ (3.52).

Inversement, pour $q > 1$:

$$ \left({}^qD\right)^{1-q} = q + \sum_{k=2}^{q} \binom{q}{k} (-1)^{k+1} \, {}_kS. \tag{3.54} $$

Pour $q = 1$, la relation est[108]

$$ {}^{1}H = -1 + \sum_{k=2}^{\infty} \frac{{}_{k}S}{k\,(k-1)}. \tag{3.55} $$

[108] C. X. Mao. "Estimating Species Accumulation Curves and Diversity Indices." In: *Statistica Sinica* 17 (2007), pp. 761–774.

Estimation

Hurlbert fournit un estimateur non biaisé de son indice (n est la taille de l'échantillon, n_s le nombre d'individus de l'espèce s):

$$ {}_{k}\tilde{S} = \sum_{s=1}^{f_{>0}} \left[1 - \binom{n - n_s}{k} \Big/ \binom{n}{k} \right]. \tag{3.56} $$

Dauby and Hardy[109] montrent que cet estimateur est très peu sensible à la taille de l'échantillon, et obtient de meilleurs résultats sur ce point que les estimateurs de Chao et Shen pour l'indice de Shannon ou du nombre d'espèces. W. Smith and Grassle[110] ont calculé sa variance.

[109] Dauby and Hardy, "Sampled-Based Estimation of Diversity Sensu Stricto by Transforming Hurlbert Diversities into Effective Number of Species," see n. 18, p. 9.

[110] W. Smith and J. F. Grassle. "Sampling Properties of a Family of Diversity Measures." In: *Biometrics* 33.2 (1977), pp. 283–292. DOI: 10.2307/2529778. JSTOR: 2529778.

Le calcul de la diversité de Hurlbert est possible dans *divent*:

```
# Indice de Hurlbert (probabilités)
ent_hurlbert(as_probabilities(colSums(BCI)), k = 2)
```

```
## # A tibble: 1 x 5
##   site   weight estimator order entropy
##   <chr>   <dbl> <chr>     <dbl>   <dbl>
## 1 site_1      1 naive         2    1.97
```

```
# Estimateur sans biais (abondances)
ent_hurlbert(colSums(BCI), k = 2)
```

```
## # A tibble: 1 x 3
##   estimator order entropy
##   <chr>     <dbl>   <dbl>
## 1 Hurlbert      2    1.97
```

```
# Nombre effectif d'espèces
div_hurlbert(colSums(BCI), k = 2)
```

```
## # A tibble: 1 x 3
##   estimator order diversity
##   <chr>     <dbl>     <dbl>
## 1 Hurlbert      2      38.1
```

Chapter 4

Entropie

L'essentiel

L'entropie est la surprise moyenne apportée par l'observation des individus d'une communauté, d'autant plus grande qu'un individu appartient à une espèce plus rare. L'entropie HCDT permet d'unifier les indices classiques de diversité: son paramètre, appelé ordre, fixe l'importance donnée aux espèces rares. L'entropie d'ordre 0 est la richesse; celle d'ordre 1, l'indice de Shannon; celle d'ordre 2, celui de Simpson. L'entropie est la moyenne du logarithme déformé de la rareté des espèces, définie comme l'inverse de leur probabilité.

L'entropie va de pair avec la diversité au sens strict (Nombres de Hill): le nombre d'espèces équiprobables dont l'entropie est la même que celle de la communauté réelle. La diversité est l'exponentielle déformée de l'entropie. Les profils de diversité représentent la diversité en fonction de son ordre et permettent la comparaison de communautés.

L'estimation de la diversité est difficile pour des ordres inférieurs à $0,5$ dans des taxocènes très divers comme les arbres des forêts tropicales.

L'entropie peut être entendue comme la surprise moyenne fournie par l'observation d'un échantillon. C'est intuitivement une bonne mesure de diversité.[1] Ses propriétés mathématiques permettent d'unifier les mesures de diversité dans un cadre général.

4.1 Définition de l'entropie

Les textes fondateurs sont Davis[2] et surtout Theil[3] en économétrie, et Shannon[4] pour la mesure de la diversité. Une revue est fournie par Maasoumi.[5]

[1] E. C. Pielou. *Ecological Diversity*. New York: Wiley, 1975.

[2] H. T. Davis. *The Theory of Econometrics*. Bloomington, Indiana: The Principia Press, 1941.

[3] Theil, *Economics and Information Theory*, see n. 78, p. 51.

[4] Shannon, "A Mathematical Theory of Communication," see n. 74, p. 50; Shannon and Weaver, *The Mathematical Theory of Communication*, see n. 74, p. 50.

[5] E. Maasoumi. "A Compendium to Information Theory in Economics and Econometrics." In: *Econometric Reviews* 12.2 (1993), pp. 137–181. DOI: 10.1080/07474939308800260.

Considérons une expérience dont les résultats possibles sont $\{r_1, r_2, \dots, r_S\}$. La probabilité d'obtenir r_s est p_s, et $\mathbf{p} = (p_1, p_2, \dots, p_S)$ est le vecteur composé des probabilités d'obtenir chaque résultat. Les probabilités sont connues *a priori*. Tout ce qui suit est vrai aussi pour des valeurs de r continues, dont on connaîtrait la densité de probabilité.

On considère maintenant un échantillon de valeurs de r. La présence de r_s dans l'échantillon est peu étonnante si p_s est grande: elle apporte peu d'information supplémentaire par rapport à la simple connaissance des probabilités. En revanche, si p_s est petite, la présence de r_s est surprenante. On définit donc une fonction d'information, $I(p_s)$, décroissante quand la probabilité augmente, de $I(0) > 0$ (éventuellement $+\infty$) à $I(1) = 0$, parce qu'observer un résultat certain n'apporte aucune information. Chaque valeur observée dans l'échantillon apporte une certaine quantité d'information, dont la somme est l'information de l'échantillon. Patil and Taillie[6] appellent l'information "rareté".

[6] G. P. Patil and C. Taillie. "Diversity as a Concept and Its Measurement." In: *Journal of the American Statistical Association* 77.379 (1982), pp. 548–561. DOI: 10.2307/2287709. JSTOR: 2287709.

La quantité d'information attendue de l'expérience est $\sum_{s=1}^{S} p_s I(p_s) = H(\mathbf{p})$. Si on choisit $I(p_s) = -\ln(p_s)$, $H(\mathbf{p})$ est l'indice de Shannon, mais bien d'autres formes de $I(p_s)$ sont possibles. $H(\mathbf{p})$ est appelée *entropie*. C'est une mesure de l'incertitude (de la volatilité) du résultat de l'expérience. Si le résultat est certain (une seule valeur p_S vaut 1), l'entropie est nulle. L'entropie est maximale quand les résultats sont équiprobables.

Si $\mathbf{p}$ est la distribution des probabilité des espèces dans une communauté, Patil and Taillie[7] montrent que:

[7] Ibid.

- Si $I(p_s) = (1-p_s)/p_s$, alors $H(\mathbf{p})$ est le nombre d'espèces S moins 1;
- Si $I(p_s) = -\ln(p_s)$, alors $H(\mathbf{p})$ est l'indice de Shannon;
- Si $I(p_s) = 1 - p_s$, alors $H(\mathbf{p})$ est l'indice de Simpson.

Ces trois fonctions d'information sont représentées en figure **??**.

Le code R nécessaire pour réaliser la figure est:

```
I0 <- function(p) (1- p) / p
I1 <- function(p) -log(p)
I2 <- function(p) 1 - p
tibble(x = c(0, 1)) %>%
  ggplot(aes(x)) +
    stat_function(fun = I0) +
    stat_function(fun = I1, lty = 2) +
    stat_function(fun = I2, lty = 3) +
    coord_cartesian(ylim = c(0, 10)) +
    labs(x = "p", y = "I(p)")
```

La contribution de chaque espèce à la valeur totale de l'entropie est représentée figure 4.1.

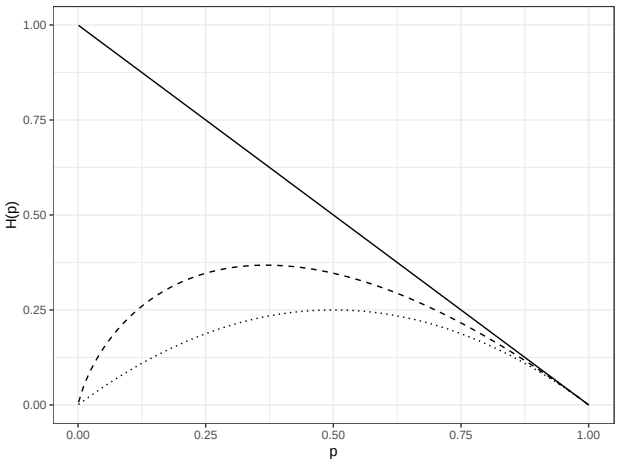


Figure 4.1: Valeur de $p_s I(p_s)$ dans le nombre d'espèces (trait plein), l'indice de Shannon (pointillés longs) et l'indice de Simpson (pointillés). Les espèces rares contribuent peu, sauf pour le nombre d'espèces.

Code R:

```
H0 <- function(p) 1 - p
H1 <- function(p) -p * log(p)
H2 <- function(p) p * (1 - p)
tibble(x = c(0.001, 1)) %>%
  ggplot(aes(x)) +
    stat_function(fun = H0) +
    stat_function(fun = H1, lty = 2) +
    stat_function(fun = H2, lty = 3) +
    labs(x = "p", y = "H(p)")
```

4.2 Entropie relative

Considérons maintenant les probabilités q_s formant l'ensemble $\mathbf{q}$ obtenues par la réalisation de l'expérience. Elles sont différentes des probabilités p_s, par exemple parce que l'expérience ne s'est pas déroulée exactement comme prévu. On définit le gain d'information $I(q_s, p_s)$ comme la quantité d'information supplémentaire fournie par l'observation d'un résultat de l'expérience, connaissant les probabilités *a priori*. La quantité totale d'information fournie par l'expérience, $\sum_{s=1}^{S} q_s I(q_s, p_s) = H(\mathbf{q}, \mathbf{p})$, est souvent appelée entropie relative. Elle peut être vue comme une distance entre la distribution *a priori* et la distribution *a posteriori*. Il est possible que les distributions $\mathbf{p}$ et $\mathbf{q}$ soit identiques, que le gain d'information soit donc nul, mais les estimateurs empiriques n'étant pas exactement égaux entre eux, des tests de significativité de la valeur de $\hat{H}(\mathbf{q}, \mathbf{p})$ seront nécessaires.

Quelques formes possibles de $H(\mathbf{q}, \mathbf{p})$ sont:

- La divergence de Kullback-Leibler[8] connue par les économistes comme l'indice de dissimilarité de Theil:[9]

$$T = \sum_{s=1}^{S} q_s \ln \frac{q_s}{p_s}; \tag{4.1}$$

[8] S. Kullback and R. A. Leibler. "On Information and Sufficiency." In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86. JSTOR: 2236703.

[9] Theil, *Economics and Information Theory*, see n. 78, p. 51.

- Sa proche parente, appelée parfois deuxième mesure de Theil,[10] qui inverse simplement les rôles de p et q:

$$L = \sum_{s=1}^{S} p_s \ln \frac{p_s}{q_s}. \tag{4.2}$$

[10] Conceição and Ferreira, *The Young Person's Guide to the Theil Index: Suggesting Intuitive Interpretations and Exploring Analytical Applications*, see n. 79, p. 51.

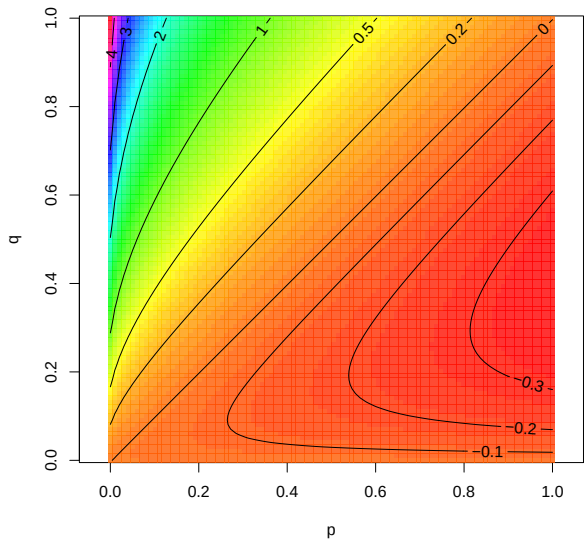


Figure 4.2: Valeur de $q \ln(q/p)$ en fonction de p et q. La divergence de Kullback-Leibler est la somme de cette valeur pour toutes les espèces

L'entropie relative est essentielle pour la définition de la diversité β présentée dans le chapitre **??**. En se limitant à la diversité α, on peut remarquer que l'indice de Shannon est la divergence de Kullback-Leibler entre la distribution observée et l'équiprobabilité des espèces.[11]

[11] E. Marcon et al. "The Decomposition of Shannon's Entropy and a Confidence Interval for *Beta* Diversity." In: *Oikos* 121.4 (2012), pp. 516–522. DOI: 10.1111/j.1600-0706.2011.19267.x.

La contribution à la divergence de chaque couple (p_s, q_s) est représentées en figure 4.2. Elle est positive quand $q_s > p_s$ (la valeur observée est plus grande que la valeur attendue), et croît avec q_s pour p_s fixé. Elle tend vers l'infini quand $p_s \to 0$. Les valeurs les plus négatives ne sont pas obtenues pour les valeurs minimales de q parce que les évènements très rarement observés influent peu sur la quantité d'information totale. Le minimum $-e^{-1} \approx 0.37$ est atteint pour $p_s = 1$ et $q_s = e^{-1}$, la valeur qui annule la dérivée de $q \ln(q)$. En somme, la divergence de Kullback-Leibler est surtout influencée par les évènements beaucoup plus observés qu'attendus.

Le code R nécessaire pour réaliser la figure est:

```
p <- q <- seq(0.01, 1, .01)
KB <- function(p, q) q * log(q / p)
xyz <- outer(p, q, FUN = "KB")
library("sp")
image(
  xyz,
  col = rainbow(n = 100, alpha = 0.8),
  xlab = "p",
```

```
  ylab = "q",
  asp = 1
)
contour(
  xyz,
  levels = c(seq(-.3, 0, .1), c(.2, .5), seq(1, 4, 1)),
  labcex = 1,
  add = T
)
```

4.3 L'appropriation de l'entropie par la biodiversité

MacArthur[12] est le premier à avoir introduit la théorie de l'information en écologie.[13] MacArthur s'intéressait aux réseaux trophiques et cherchait à mesurer leur stabilité: l'indice de Shannon qui comptabilise le nombre de relations possibles lui paraissait une bonne façon de l'évaluer. Mais l'efficacité implique la spécialisation, ignorée dans H qui est une mesure neutre (toutes les espèces y jouent le même rôle). MacArthur a abandonné cette voie.

Les premiers travaux consistant à généraliser l'indice de Shannon sont dus à Rényi.[14] L'entropie d'ordre q de Rényi est

$$ {}^{q}R = \frac{1}{1 - q \ln \sum_{q=1}^{S} p_s^q}. \tag{4.3} $$

Rényi pose également les axiomes pour une mesure d'entropie $R(\mathbf{p})$, où $\mathbf{p} = (p_1, p_2, \dots, p_S)$:

- La symétrie: les espèces doivent être interchangeables, aucune n'a de rôle particulier et leur ordre est indifférent;
- La mesure doit être continue par rapport aux probabilités;
- La valeur maximale est atteinte si toutes les probabilités sont égales.

Il montre que ${}^{q}R$ respecte les 3 axiomes.

Patil and Taillie[15] ont montré de plus que:

- L'introduction d'une espèce dans une communauté augmente sa diversité (conséquence de la décroissance de $g(p_s)$);
- Le remplacement d'un individu d'une espèce fréquente par un individu d'une espèce plus rare augmente l'entropie à condition que $R(\mathbf{p})$ soit concave. Dans la littérature économique sur les inégalités, cette propriété est connue sous le nom de Pigou-Dalton.[16]

[12] R. H. MacArthur. "Fluctuations of Animal Populations and a Measure of Community Stability." In: *Ecology* 36.3 (1955), pp. 533–536. DOI: doi:10.2307/1929601.

[13] R. E. Ulanowicz. "Information Theory in Ecology." In: *Computers & Chemistry* 25.4 (2001), pp. 393–399. DOI: 10.1016/S0097-8485(01)00073-0.

[14] A. Rényi. "On Measures of Entropy and Information." In: *4th Berkeley Symposium on Mathematical Statistics and Probability*. Ed. by J. Neyman. Vol. 1. Berkeley, USA: University of California Press, 1961, pp. 547–561.

[15] Patil and Taillie, "Diversity as a Concept and Its Measurement," see n. 6, p. 62.

[16] H. Dalton. "The Measurement of the Inequality of Incomes." In: *The Economic Journal* 30.119 (1920), pp. 348–361. DOI: 10.2307/2223525.

[17] M. O. Hill. "Diversity and Evenness: A Unifying Notation and Its Consequences." In: *Ecology* 54.2 (1973), pp. 427–432. DOI: 10.2307/1934352.

Hill[17] transforme l'entropie de Rényi en *nombres de Hill*, qui en sont simplement l'exponentielle:

$$ {}^qD = \left(\sum_{s=1}^{S} p_s^q \right)^{\frac{1}{1-q}}. \tag{4.4} $$

Le souci de Hill était de rendre les indices de diversité intelligibles après l'article remarqué de S. H. Hurlbert[18] intitulé "le non-concept de diversité spécifique". Hurlbert reprochait à la littérature sur la diversité sa trop grande abstraction et son éloignement des réalités biologiques, notamment en fournissant des exemples dans lesquels l'ordre des communautés n'est pas le même selon l'indice de diversité choisi. Les nombres de Hill sont le nombre d'espèces équiprobables donnant la même valeur de diversité que la distribution observée. Ils sont des transformations simples des indices classiques:

- ${}^0D$ est le nombre d'espèces;
- ${}^1D = e^H$, l'exponentielle de l'indice de Shannon;
- ${}^2D = 1/(1-E)$, l'inverse de l'indice de concentration de Simpson, connu sous le nom d'indice de Stoddart.[19]

Ces résultats avaient déjà été obtenus avec une autre approche par MacArthur[20] et repris par Adelman[21] dans la littérature économique.

Les nombres de Hill sont des "nombres effectifs" ou "nombres équivalents". Le concept a été défini rigoureusement par Gregorius,[22] d'après Wright[23] (qui avait le premier défini la taille effective d'une population): étant donné une variable caractéristique (ici, l'entropie) fonction seulement d'une variable numérique (ici, le nombre d'espèces), dans un cas idéal (ici, l'équiprobabilité des espèces), le nombre effectif est la valeur de la variable numérique pour laquelle la variable caractéristique est celle du jeu de données.

Gregorius[24] montre que de nombreux autres indices de diversité sont acceptables dans le sens où ils vérifient les axiomes précédents et, de plus, que la diversité d'un assemblage de communautés est obligatoirement supérieure à la diversité moyenne de ces communautés (l'égalité n'étant possible que si les communautés sont toutes identiques). Cette dernière propriété sera traitée en détail dans la partie consacrée à la décomposition de la diversité. Ces indices doivent vérifier deux propriétés: leur fonction d'information doit être décroissante, et ils doivent être une fonction strictement concave de p_s. Parmi les possibilités, $I(p_s) = \cos(p_s \pi/2)$ est envisageable par exemple: le choix de la fonction d'information est virtuellement illimité, mais seules quelques unes seront interprétables clairement.

[18] S. H. Hurlbert, "The Nonconcept of Species Diversity: A Critique and Alternative Parameters," see n. 69, p. 48.

[19] J. A. Stoddart. "A Genotypic Diversity Measure." In: *Journal of Heredity* 74 (1983), pp. 489–490. DOI: 10.1093/oxfordjournals.jhered.a109852.

[20] R. H. MacArthur. "Patterns of Species Diversity." In: *Biological Reviews* 40.4 (1965), pp. 510–533. DOI: 10.1111/j.1469-185X.1965.tb00815.x.

[21] M. A. Adelman. "Comment on the "H" Concentration Measure as a Numbers-Equivalent." In: *The Review of Economics and Statistics* 51.1 (1969), pp. 99–101. DOI: 10.2307/1926955.

[22] H.-R. Gregorius. "On the Concept of Effective Number." In: *Theoretical population biology* 40.2 (1991), pp. 269–83. DOI: 10.1016/0040-5809(91)90056-L. PMID: 1788824.

[23] S. Wright. "Evolution in Mendelian Populations." In: *Genetics* 16.2 (1931), pp. 97–159.

[24] H.-R. Gregorius. "Partitioning of Diversity : The "within Communities" Component." In: *Web Ecology* 14 (2014), pp. 51–60. DOI: 10.5194/we-14-51-2014.

Un nombre équivalent d'espèces existe pour tous ces indices, il est toujours égal à l'inverse de l'image de l'indice par la réciproque de la fonction d'information:

$$D = \frac{1}{I^{-1}\left(\sum_{s=1}^{S} p_s I(p_s)\right)}. \tag{4.5}$$

D'autres entropies ont été utilisées, avec plus ou moins de succès. Par exemple, Ricotta and Avena[25] proposent d'utiliser la fonction d'information $I(p_s) = -\ln(k_s)$ où k_s est la dissimilarité totale de l'espèce s avec les autres (par exemple, la somme des distances aux autres espèces dans un arbre phylogénétique, voir section **??**), normalisée pour que $\sum_s k_s = 1$. Ainsi, les espèces les plus originales apportent peu d'information, ce qui n'est pas très intuitif. Les auteurs montrent que leur mesure est la somme de l'entropie de Shannon et de la divergence de Kullback-Leibler entre les probabilités et les dissimilarités des espèces. Ricotta and Szeidl[26] ont défini plus tard une entropie augmentant avec l'originalité de chaque espèce, présentée au chapitre **??**.

[25] Ricotta and Avena, "An Information-Theoretical Measure of Taxonomic Diversity," see n. 3, p. 25.

[26] C. Ricotta and L. Szeidl. "Towards a Unifying Approach to Diversity Measures: Bridging the Gap between the Shannon Entropy and Rao's Quadratic Index." In: *Theoretical Population Biology* 70.3 (2006), pp. 237–243. DOI: 10.1016/j.tpb.2006.06.003.

4.4 Entropie HCDT

Tsallis[27] propose une classe de mesures appelée entropie généralisée, définie par Havrda and Charvát[28] pour la première fois et redécouverte plusieurs fois, notamment par Daróczy,[29] d'où son nom *entropie HCDT* (voir Mendes et al.,[30] page 451, pour un historique complet):

$${}^{q}H = \frac{1}{q-1}\left(1 - \sum_{s=1}^{S} p_s^q\right). \tag{4.6}$$

Tsallis a montré que les indices de Simpson et de Shannon étaient des cas particuliers d'entropie généralisée, retrouvant, sans faire le rapprochement,[31] la définition d'un indice de diversité de Patil and Taillie.[32]

Ces résultats ont été complétés par d'autres et repris en écologie par Keylock[33] et Jost.[34] Là encore:

- Le nombre d'espèces moins 1 est ${}^{0}H$;
- L'indice de Shannon est ${}^{1}H$;
- L'indice de Gini-Simpson est ${}^{2}H$.

L'entropie HCDT est particulièrement attractive parce que sa relation avec la diversité au sens strict est simple, après introduction du formalisme adapté (les logarithmes déformés). Son biais d'estimation peut être corrigé globalement, et non

[27] C. Tsallis. "Possible Generalization of Boltzmann-Gibbs Statistics." In: *Journal of Statistical Physics* 52.1 (1988), pp. 479–487. DOI: 10.1007/BF01016429.

[28] J. Havrda and F. Charvát. "Quantification Method of Classification Processes. Concept of Structural Alpha-Entropy." In: *Kybernetika* 3.1 (1967), pp. 30–35.

[29] Z. Daróczy. "Generalized Information Functions." In: *Information and Control* 16.1 (1970), pp. 36–51. DOI: 10.1016/s0019-9958(70)80040-7.

[30] Mendes et al., "A Unified Index to Measure Ecological Diversity and Species Rarity," see n. 70, p. 48.

[31] C. Ricotta. "On Parametric Diversity Indices in Ecology: A Historical Note." In: *Community Ecology* 6.2 (2005), pp. 241–244. DOI: 10.1556/ComEc.6.2005.2.12.

[32] Patil and Taillie, "Diversity as a Concept and Its Measurement," see n. 6, p. 62.

[33] C. J. Keylock. "Simpson Diversity and the Shannon-Wiener Index as Special Cases of a Generalized Entropy." In: *Oikos* 109.1 (2005), pp. 203–207. DOI: 10.1111/j.0030-1299.2005.13735.x.

[34] L. Jost. "Entropy and Diversity." In: *Oikos* 113.2 (2006), pp. 363–375. DOI: 10.1111/j.2006.0030-1299.14714.x; L. Jost. "Partitioning Diversity into Independent Alpha and Beta Components." In: *Ecology* 88.10 (2007), pp. 2427–2439. DOI: 10.1890/06-1736.1.

seulement pour les cas particuliers (nombre d'espèces, Shannon, Simpson). Enfin, sa décomposition sera présentée en détail dans le chapitre **??**.

4.5 Logarithmes déformés

L'écriture de l'entropie HCDT est largement simplifiée en introduisant le formalisme des logarithmes déformés.[35] Le logarithme d'ordre q est défini par

$$\ln_q x = \frac{x^{1-q} - 1}{1 - q}, \tag{4.7}$$

dont la forme est identique à la transformation de Box and Cox[36] utilisée en statistiques pour normaliser une variable.

Le logarithme déformé converge vers le logarithme naturel quand $q \to 1$ (figure 4.3).

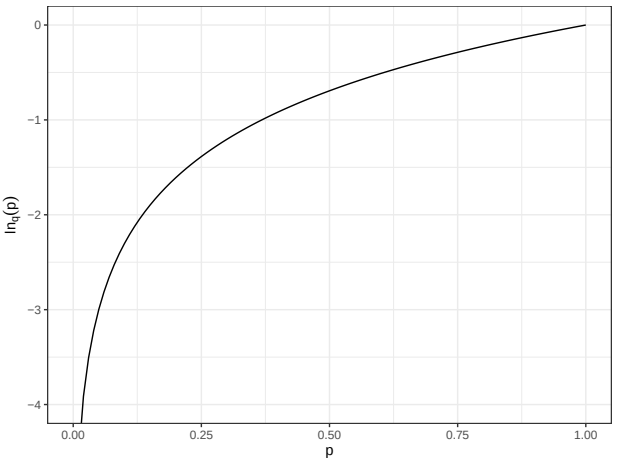


Figure 4.3: Valeur du logarithme d'ordre q de probabilités entre 0 et 1 pour différentes valeurs de q: $q = 0$ (pointillés longs rouges), la courbe est une droite; $q = 1$ (trait plein): logarithme naturel; $q = 2$ (pointillés courts bleus): la courbe a la même forme que le logarithme naturel pour les valeurs positives de q; $q = -1$ (pointillés alternés verts): la courbe est convexe pour les valeurs négatives de q.

Le code R nécessaire pour réaliser la figure est:

```
ln0 <- function(p) lnq(p, 0)
ln2 <- function(p) lnq(p, 2)
lnm1 <- function(p) lnq(p, -1)
tibble(x = c(0, 1)) %>%
  ggplot(aes(x)) +
    stat_function(fun = log) +
    stat_function(fun = ln0, lty = 2, col = "red") +
    stat_function(fun = ln2, lty = 3, col = "blue") +
    stat_function(fun = lnm1, lty = 4, col = "green") +
    coord_cartesian(ylim = c(-4, 0)) +
    labs(x = "p", y = expression(ln[q](p)))
```

Sa fonction inverse est l'exponentielle d'ordre q:

$$e_q^x = \left[1 + (1 - q)\, x\right]^{\frac{1}{1-q}}. \tag{4.8}$$

Enfin, le logarithme déformé est subadditif:

$$\ln_q (xy) = \ln_q x + \ln_q y - (q - 1) \left(\ln_q x\right) \left(\ln_q y\right). \tag{4.9}$$

[35] C. Tsallis. "What Are the Numbers That Experiments Provide?" In: *Química Nova* 17.6 (1994), pp. 468–471.

[36] G. E. P. Box and D. R. Cox. "An Analysis of Transformations." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 26.2 (1964), pp. 211–252. DOI: 10.1111/j.2517-6161.1964.tb00553.x. JSTOR: 2984418.

Ses propriétés sont les suivantes:

$$\ln_q \frac{1}{x} = -x^{q-1} \ln_q x; \tag{4.10}$$

$$\ln_q (xy) = \ln_q x + x^{1-q} \ln_q y; \tag{4.11}$$

$$\ln_q \left(\frac{x}{y}\right) = \ln_q x - \left(\frac{x}{y}\right)^{1-q} \ln_q y; \tag{4.12}$$

et

$$e_q^{x+y} = e_q^x e_q^{\frac{y}{1+(1-q)x}}. \tag{4.13}$$

Si $q > 1$, $\lim_{x \to +\infty} (\ln_q x) = 1/(q-1)$, donc e_q^x n'est pas définie pour $x > 1/(q-1)$.

La dérivée du logarithme déformé est, quel que soit q,

$$\ln_q' (x) = x^{-q}. \tag{4.14}$$

Les dérivées première et seconde de l'exponentielle déformée sont, quel que soit q:

$$\exp_q' (x) = \left(e_q^x\right)^q; \tag{4.15}$$

$$\exp_q'' (x) = \left(e_q^x\right)^{2q-1}. \tag{4.16}$$

Ces fonctions sont implémentées dans le package *divent*: `ln_q(x, q)` et `exp_q(x, q)`.

L'entropie d'ordre q s'écrit

$${}^qH = \frac{1}{q-1}\left(1 - \sum_{s=1}^{S} p_s^q\right) = -\sum_s p_s^q \ln_q p_s = \sum_s p_s \ln_q \frac{1}{p_s}. \tag{4.17}$$

Ces trois formes sont équivalentes mais les deux dernières s'interprètent comme une généralisation de l'entropie de Shannon.[37] La dernière est la plus intéressante parce qu'elle permet de définir l'entropie en général comme la moyenne du logarithme de l'inverse des probabilités, que nous appellerons la rareté des espèces.

[37] E. Marcon et al. "Generalization of the Partitioning of Shannon Diversity." In: *Plos One* 9.3 (2014), e90289. DOI: 10.1371/journal.pone.0090289.

Le calcul de ${}^qH$ peut se faire avec la fonction `ent_tsallis` de la librairie *divent*:

```
ent_tsallis(bci_prob, q= 1.5)
```

```
## # A tibble: 1 x 3
##   estimator order entropy
##   <chr>     <dbl>   <dbl>
## 1 naive       1.5    1.72
```

4.6 Entropie et diversité

On voit immédiatement que l'entropie de Tsallis est le logarithme d'ordre q du nombre de Hill correspondant, comme l'entropie de Rényi en est le logarithme naturel:

$$ {}^{q}H = \ln_q {}^{q}D; \tag{4.18} $$

$$ {}^{q}D = e_q^{{}^{q}H}. \tag{4.19} $$

L'entropie est utile pour les calculs: la correction des biais d'estimation notamment. Les nombres de Hill, ou *nombres équivalents d'espèces* ou *nombres effectif d'espèces* permettent une appréhension plus intuitive de la notion de biodiversité.[38] En raison de leurs propriétés, notamment de décomposition (voir le chapitre **??**), Jost[39] les appelle "vraie diversité". S. Hoffmann and A. Hoffmann[40] critiquent cette définition totalitaire et fournissent une revue historique plus lointaine sur les origines de ces mesures. Jost[41] reconnaît qu'un autre terme aurait pu être choisi ("diversité neutre" ou "diversité mathématique" par exemple).

Dauby and Hardy[42] écrivent "diversité au sens strict"; Gregorius[43] "diversité explicite".

Quoi qu'il en soit, les nombres de Hill respectent le principe de réplication (voir Chao et al.,[44] section 3 pour une discussion et un historique): si I communautés de même taille, de même niveau de diversité D, mais sans espèces en commun sont regroupées dans une méta-communauté, la diversité de la méta-communauté doit être $I \times D$.

L'intérêt de ces approches est de fournir une définition paramétrique de la diversité, qui donne plus ou moins d'importance aux espèces rares:

- ${}^{-\infty}D = 1/\min(p_S)$ est l'inverse de la proportion de la communauté représentée par l'espèce la plus rare (toutes les autres espèces sont ignorées). Le biais d'estimation est incontrôlable: l'espèce la plus rare n'est pas dans l'échantillon tant que l'inventaire n'est pas exhaustif;
- ${}^{0}D$ est le nombre d'espèces (alors que ${}^{0}H$ est le nombre d'espèces moins 1). C'est la mesure classique qui donne le plus d'importance aux espèces rares: toutes les espèces ont la même importance, quel que soit leur effectif en termes d'individus. Il est bien adapté à une approche patrimoniale, celle du collectionneur qui considère que l'existence d'une espèce supplémentaire a un intérêt en soi, par exemple parce qu'elle peut contenir une molécule valorisable. Comme les espèces rares sont difficiles à échantillonner, le biais d'estimation est très

[38] Jost, "Entropy and Diversity," see n. 34, p. 67.

[39] Jost, "Partitioning Diversity into Independent Alpha and Beta Components," see n. 34, p. 67.

[40] S. Hoffmann and A. Hoffmann. "Is There a "True" Diversity?" In: *Ecological Economics* 65.2 (2008), pp. 213–215. DOI: 10.1016/j.ecolecon.2008.01.009.

[41] L. Jost. "Mismeasuring Biological Diversity: Response to Hoffmann and Hoffmann (2008)." In: *Ecological Economics* 68 (2009), pp. 925–928. DOI: 10.1016/j.ecolecon.2008.10.015.

[42] Dauby and Hardy, "Sampled-Based Estimation of Diversity Sensu Stricto by Transforming Hurlbert Diversities into Effective Number of Species," see n. 18, p. 9.

[43] H.-R. Gregorius. "Linking Diversity and Differentiation." In: *Diversity* 2.3 (2010), pp. 370–394. DOI: 10.3390/d2030370.

[44] A. Chao et al. "Phylogenetic Diversity Measures Based on Hill Numbers." In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 365.1558 (2010), pp. 3599–3609. DOI: 10.1098/rstb.2010.0272Supplementary.

important, et sa résolution a généré une littérature en soi (section 3.1);
- ${}^1D$ est l'exponentielle de l'indice de Shannon donne la même importance à tous les individus. Il est adapté à une approche d'écologue, intéressé par les interactions possibles: le nombre de combinaisons d'espèces en est une approche satisfaisante. Le biais d'estimation est sensible;
- ${}^2D$ est l'inverse de l'indice de concentration de Gini-Simpson donne moins d'importance aux espèces rares. Hill[45] l'appelle "le nombre d'espèces très abondantes". Il comptabilise les interactions possibles entre paires d'individus: les espèces rares interviennent dans peu de paires, et influent peu sur l'indice. En conséquence, le biais d'estimation est très petit; de plus, un estimateur non biaisé existe;
- ${}^\infty D = 1/d$ est l'inverse de l'indice de Berger-Parker[46] qui est la proportion de la communauté représentée par l'espèce la plus abondante: $d = \max(\mathbf{p})$. Toutes les autres espèces sont ignorées.

[45] Hill, "Diversity and Evenness: A Unifying Notation and Its Consequences," see n. 17, p. 65.

[46] W. H. Berger and F. L. Parker. "Diversity of Planktonic Foraminifera in Deep-Sea Sediments." In: *Science* 168.3937 (1970), pp. 1345–1347. DOI: 10.1126/science.168.3937.1345.

Le calcul de ${}^qD$ peut se faire avec la fonction `div_hill` de la librairie *divent*:

```
div_hill(bci_prob, q = 1.5)
```

```
## # A tibble: 1 x 3
##   estimator order diversity
##   <chr>     <dbl>     <dbl>
## 1 naive       1.5      49.6
```

Les propriétés mathématiques de la diversité ne sont pas celles de l'entropie. L'entropie doit être une fonction concave des probabilités comme on l'a vu plus haut, mais pas la diversité (un exemple de confusion est fourni par Gadagkar,[47] qui reproche à ${}^2D$ de ne pas être concave). L'entropie est une moyenne pondérée par les probabilités de la fonction d'information, c'est donc une fonction linéaire des probabilités, propriété importante pour définir l'entropie α (section **??**) comme la moyenne des entropies de plusieurs communautés, ou l'entropie phylogénétique (chapitre **??**) comme la moyenne de l'entropie sur les périodes d'un arbre. La diversité n'est pas une fonction linéaire des probabilités: la diversité moyenne n'est en général pas la moyenne des diversités.

[47] R. Gadagkar. "An Undesirable Property of Hill's Diversity Index N2." In: *Oecologia* 80 (1989), pp. 140–141. DOI: 10.1007/BF00789944.

4.7 Synthèse

L'inverse de la probabilité d'une espèce, $1/p_s$, définit sa rareté. L'entropie est la moyenne du logarithme de la rareté:

$$ {}^{q}H = \frac{1}{q-1}\left(1 - \sum_{s=1}^{S} p_s^q\right) = \sum_s p_s \ln_q \frac{1}{p_s}. \tag{4.20} $$

La diversité est son exponentielle:

$$ {}^{q}D = e_q^{{}^{q}H}. \tag{4.21} $$

4.8 Estimation

Les estimateurs peuvent être classés dans quatre méthodes principales. La plus simple consiste simplement à insérer l'estimateur de p_s dans la définition de la diversité à évaluer pour obtenir ce que l'on appelle l'estimateur *plug-in*. Quand l'estimateur des probabilités est $\hat{p}_s = n_s/n$, l'estimateur est dit *naïf*. L'estimateur naïf de l'entropie HCDT d'ordre q est:

$$ {}^{q}\hat{H} = \sum_s \hat{p}_s \ln_q \frac{1}{\hat{p}_s} \tag{4.22} $$

L'estimateur naïf est inutile dans les communautés hyperdiverses car il sous-estime fortement la diversité en raison des espèces non observées et de la non-linéarité de la l'entropie en fonction des probabilités.

Des progrès ont été réalisés dans l'estimation de la distribution réelle de la probabilité des espèces en ajustant un modèle de leur distribution aux données. La distribution des espèces non observées peut être ajoutée si leur nombre est estimé et qu'une forme de distribution est choisie. La première méthode d'estimation de l'entropie consiste donc à estimer précisément la distribution réelle des probabilités puis à lui appliquer l'estimateur plug-in.

Chao et al.[48] ont utilisé un modèle à deux paramètres basé sur l'estimation de la couverture généralisée de l'échantillon (non détaillé ici), estimé la richesse totale avec l'estimateur Chao1 et modélisé les espèces non observées comme une distribution géométrique pour dévoiler la distribution complète rang-abondance d'une communauté observée. Ils ont appliqué l'estimateur plug-in à cette distribution pour obtenir l'estimateur appelé "Chao-unveiled" dans *divent*. L'estimateur "iChao-unveiled" est une variation de l'estimateur "Chao-unveiled", où la richesse est estimée par l'estimateur iChao1.

L'estimateur du jackknife[49] a montré de bonnes performances pour estimer la richesse lorsque l'effort d'échantillonnage est trop faible pour que l'estimateur Chao1 soit performant.[50]

[48] A. Chao et al. "Unveiling the Species-Rank Abundance Distribution by Generalizing Good-Turing Sample Coverage Theory." In: *Ecology* 96.5 (2015), pp. 1189–1201. DOI: 10.1890/14-0550.1.

[49] Burnham and Overton, "Robust Estimation of Population Size When Capture Probabilities Vary among Animals," see n. 9, p. 26.

[50] Brose et al., "Estimating Species Richness: Sensitivity to Sample Coverage and Insensitivity to Spatial Patterns," see n. 32, p. 36.

L'estimation de la richesse à l'aide de l'estimateur du jackknife, dont l'ordre est choisi en fonction des données, définit l'estimateur "jackknife-unveiled". L'utilisation de l'estimateur jackknife pour estimer la queue de la distribution d'abondance n'était pas l'intention de Chao et al.[51] car elle n'est pas cohérente avec leur cadre théorique.

[51] Chao et al., see n. 48.

La seconde méthode repose sur l'estimateur Horvitz and Thompson[52] de la somme pondérée d'une fonction de ses éléments $x_1, x_2, ..., x_S$, soit $\sum_s p_s f(x_s)$ lorsque certains d'entre eux ne sont pas observés. Un estimateur sans biais de la somme est obtenu lorsque chaque terme est divisé par sa probabilité d'être observé $1-(1-p_s)^n$. Chao and Shen[53] ont proposé de le combiner avec l'estimateur de la couverture de l'échantillon: conditionnellement à l'ensemble des espèces observées, un estimateur sans biais[54] de p_s est $\tilde{p}_s = \hat{C}\hat{p}_s$. Chao et Shen ont estimé l'entropie de Shannon; la méthode a ensuite été étendue à l'entropie HCDT.[55]

[52] Horvitz and Thompson, "A Generalization of Sampling without Replacement from a Finite Universe," see n. 86, p. 52.

[53] Chao and Shen, "Nonparametric Estimation of Shannon's Index of Diversity When There Are Unseen Species in Sample," see n. 84, p. 52.

[54] Ashbridge and Goudie, "Coverage-Adjusted Estimators for Mark-Recapture in Heterogeneous Populations," see n. 85, p. 52.

[55] Marcon et al., "Generalization of the Partitioning of Shannon Diversity," see n. 37, p. 69.

$$ {}^q\tilde{H} = \sum_s \frac{\hat{C}\hat{p}_s \ln_q \frac{1}{\hat{C}\hat{p}_s}}{1-\left(1-\hat{C}\hat{p}_s\right)^n} \tag{4.23} $$

Un progrès supplémentaire peut être fait en remplaçant l'estimateur conditionnel des probabilités $\tilde{p}_s = \hat{C}\hat{p}_s$ par celui de Chao et al.[56] Étant donné que l'estimateur de probabilité amélioré dépend de la couverture généralisée de l'échantillon, l'estimateur amélioré de Chao-Shen est appelé "generalized coverage".

[56] Chao et al., see n. 48.

La troisième méthode a été établie par Grassberger[57] qui a donné un estimateur à biais réduit de la valeur d'un entier à la puissance q. p_s^q s'écrit n_s^q/n^q et n_s^q est estimé[58] par:

[57] Grassberger, "Finite Sample Corrections to Entropy and Dimension Estimates," see n. 98, p. 55.

[58] Marcon et al., "Generalization of the Partitioning of Shannon Diversity," see n. 37, p. 69.

$$ \tilde{n}_s^q = \frac{(n_s+1)}{(n_s-q+1)} + \frac{(-1)^n\ (1+q)\sin\pi q}{\pi\,(n+1)} \tag{4.24} $$

L'estimateur de p_s^q est simplement $\tilde{p}_s^q = \tilde{n}_s^q/n^q$. On l'introduit dans la formule de l'entropie pour obtenir l'estimateur de Grassberger:

$$ {}^q\tilde{H} = \frac{1-\sum_s \tilde{p}_s^q}{q-1} \tag{4.25} $$

La dernière méthode a fait l'objet d'une importante littérature. Une revue peut être trouvée dans Chao et al.,[59] Appendix A. Elle repose sur l'estimation de $h_q = \sum_s p_s^q$. h_q peut être écrit comme la somme suivante:

[59] Chao et al., "Entropy and the Species Accumulation Curve: A Novel Entropy Estimator via Discovery Rates of New Species," see n. 94, p. 54.

$$ h_q = \sum_{r=0}^{\infty} \binom{q-1}{r} (-1)^r\, \zeta_r \tag{4.26} $$

ζ_r est l'entropie de Simpson généralisée $\sum_s p_s(1-p_s)^r$ définie par Z. Zhang and Zhou.[60] Les n premiers éléments de la somme, notés $\tilde{h}_q$, peuvent être estimés sans biais:[61]

$$\tilde{h}_q = \sum_{s=1}^{S} \hat{p}_s \sum_{v=1}^{n-n_s} \left[\prod_{i=1}^{v} \frac{i-q}{i} \prod_{j=1}^{v} \left(1 - \frac{n_s - 1}{n-j}\right) \right] \tag{4.27}$$

Z. Zhang[62] montre que le biais dû à l'ignorance des termes restants est asymptotiquement normal et décroît exponentiellement vite. L'estimateur de Z. Zhang and Grabchak[63] est celui basé sur $\tilde{h}_q$:

$${}^q\tilde{H} = \frac{1-\tilde{h}_q}{q-1} \tag{4.28}$$

Des tentatives ont été faites pour estimer le biais restant.[64] La plus aboutie est celle de Chao and Jost,[65] qui complète Chao et al.[66] Elle repose sur l'estimation du nombre total d'espèces par l'estimateur Chao1 et quelques approximations, notamment que les probabilités réelles des espèces non observées peuvent être supposées presque égales. Il en résulte que l'estimateur de la probabilité moyenne des espèces échantillonnées une fois est également égal à l'estimateur de la probabilité des espèces non observées. Sa valeur est notée A. Elle vaut $2f_2/[(n-1)f_1 + 2f_2]$ si les singletons et les doubletons sont présents ou $2/[(n-1)(f_1-1)+2]$ si les doubletons sont absents. L'estimateur Chao-Jost de l'entropie HCDT est:

$${}^q\tilde{H} = \frac{1}{q-1}\left[1 - \tilde{h}_q - \frac{f_1}{n}(1-A)^{1-n}\left(A^{q-1} - \sum_{r=0}^{n-1}\binom{q-1}{r}(A-1)^r\right)\right] \tag{4.29}$$

En l'absence de singletons et de doubletons, A est fixé à 1 et l'estimateur est identique à celui de Zhang et Grabchak.

4.9 Profils de diversité

Leinster and Cobbold,[67] après Hill,[68] Patil and Taillie,[69] Tothmeresz[70] et Kindt et al.,[71] recommandent de tracer des profils de diversité, c'est-à-dire la valeur de la diversité ${}^qD$ en fonction de l'ordre q (figure 4.4) pour comparer plusieurs communautés. Une communauté peut être déclarée plus diverse qu'une autre si son profil de diversité est au-dessus de l'autre pour toutes les valeurs de q. Si les courbes se croisent, il n'y a pas de relation d'ordre.[72]

Lande et al.[73] montrent que si la diversité de Simpson et la richesse de deux communautés n'ont pas le même ordre, alors

[60] Z. Zhang and J. Zhou. "Re-Parameterization of Multinomial Distributions and Diversity Indices." In: *Journal of Statistical Planning and Inference* 140.7 (2010), pp. 1731–1738. DOI: 10.1016/j.jspi.2009.12.023.

[61] Z. Zhang and M. Grabchak. "Entropic Representation and Estimation of Diversity Indices." In: *Journal of Nonparametric Statistics* 28.3 (2016), pp. 563–575. DOI: 10.1080/10485252.2016.1190357.

[62] Z. Zhang, "Asymptotic Normality of an Entropy Estimator with Exponentially Decaying Bias," see n. 90, p. 53.

[63] Z. Zhang and Grabchak, see n. 61.

[64] Z. Zhang and Grabchak, "Bias Adjustment for a Nonparametric Entropy Estimator," see n. 91, p. 53.

[65] A. Chao and L. Jost. "Estimating Diversity and Entropy Profiles via Discovery Rates of New Species." In: *Methods in Ecology and Evolution* 6.8 (2015), pp. 873–882. DOI: 10.1111/2041-210X.12349.

[66] Chao et al., "Entropy and the Species Accumulation Curve: A Novel Entropy Estimator via Discovery Rates of New Species," see n. 94, p. 54.

[67] Leinster and Cobbold, "Measuring Diversity: The Importance of Species Similarity," see n. 107, p. 58.

[68] Hill, "Diversity and Evenness: A Unifying Notation and Its Consequences," see n. 17, p. 65.

[69] Patil and Taillie, "Diversity as a Concept and Its Measurement," see n. 6, p. 62.

[70] B. Tothmeresz. "Comparison of Different Methods for Diversity Ordering." In: *Journal of Vegetation Science* 6.2 (1995), pp. 283–290. DOI: 10.2307/3236223.

[71] R. Kindt et al. "Tree Diversity in Western Kenya: Using Profiles to Characterise Richness and Evenness." In: *Biodiversity and Conservation* 15.4 (2006), pp. 1253–1270. DOI: 10.1007/s10531-005-0772-x.

[72] Tothmeresz, see n. 70.

[73] R. Lande et al. "When Species Accumulation Curves Intersect: Implications for Ranking Diversity Using Small Samples." In: *Oikos* 89.3 (2000), pp. 601–605. DOI: 10.1034/j.1600-0706.2000.890320.x.

les courbes d'accumulation du nombre d'espèces en fonction du nombre d'individus échantillonnés se croisent aussi.

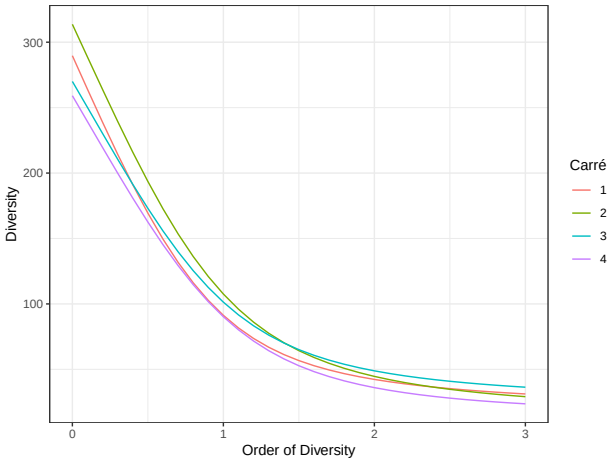


Figure 4.4: Profils de diversité des quatre carrés de la parcelle 6 de Paracou. La correction du biais d'estimation est celle de Chao et Jost. Le carré 2 est plus divers que le carré 4, mais pas que les carrés 1 et 3, qui sont plus divers dans les grands ordres de diversité. Le profil de diversité est tracé ici jusqu'à l'ordre $q = 3$.

Code R pour réaliser la figure 4.4:

```
paracou_6_abd %>%
  # Suppression de "subplot_" dans les noms des carrés
  mutate(site = str_replace(.$site, "subplot_", "")) %>%
  # Profil
  profile_hill(orders = seq(0, 3, .1), estimator = "ChaoJost") %>%
  autoplot() +
  labs(color = "Carré")
```

Pallmann et al.[74] ont développé un test statistique pour comparer la diversité de deux communautés pour plusieurs valeurs de q simultanément.

C. Liu et al.[75] nomment *séparables* des communautés dont les profils ne se croisent pas. Ils montrent que des communautés peuvent être séparables en selon un profil de diversité de ${}^qD$ sans l'être forcément selon un profil de diversité de Hurlbert (section 3.4), et inversement. Ils montrent que les communautés séparables selon un troisième type de profil, celui de la queue de distribution,[76] le sont dans tous les cas. Le profil de la queue de distribution est construit en classant les espèces de la plus fréquente à la plus rare et en traçant la probabilité qu'un individu appartienne à une espèce plus rare que l'espèce en abscisse (figure 4.5).

Code R:

```
paracou_6_abd %>%
  # Transformation en probabilités
  as_probabilities() %>%
  # Elimination des colonnes de description
  as.matrix() %>%
  # Tri des espèces par ordre croissant de probabilité
  apply(MARGIN = 1, FUN = sort) %>%
  # Cumul des probabilités. Attention: apply a transposé la matrice
  apply(MARGIN = 2, FUN = cumsum) %>%
  # Classement par valeurs décroissantes
  apply(MARGIN = 2, FUN = rev) %>%
```

[74] P. Pallmann et al. "Assessing Group Differences in Biodiversity by Simultaneously Testing a User-Defined Selection of Diversity Indices." In: *Molecular Ecology Resources* 12.6 (2012), pp. 1068–1078. DOI: 10.1111/1755-0998.12004.

[75] C. Liu et al. "Unifying and Distinguishing Diversity Ordering Methods for Comparing Communities." In: *Population Ecology* 49.2 (2006), pp. 89–100. DOI: 10.1007/s10144-006-0026-0.

[76] Patil and Taillie, "Diversity as a Concept and Its Measurement," see n. 6, p. 62.

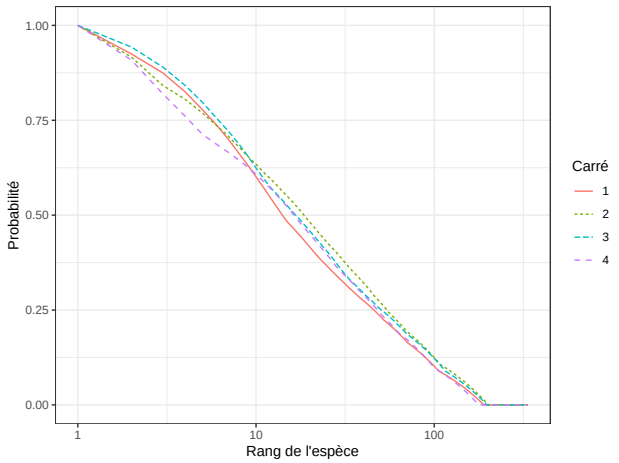


Figure 4.5: Profil de queue de distribution calculé pour les carrés de la parcelle 6 de Paracou. En abscisse: rang de l'espèce dans le classement de la plus fréquente à la plus rare; en ordonnée: probabilité qu'un individu de la communauté appartienne à une espèce plus rare. Les profils de queue de distribution se croisent d'autant plus tôt qu'ils se croisent pour de grands ordres de diversité dans la figure 4.4.

```
# Création d'un tibble pour le graphique
as_tibble() %>%
# Nom des colonnes: 1 à 4
rename_with(~ str_replace(., "V", "")) %>%
# Ajout d'une colonne pour l'ordre
mutate(rank = seq_len(nrow(.))) %>%
# Création du graphique
pivot_longer(cols = -rank) %>%
ggplot() +
  geom_line(aes(x = rank, y = value, color = name, lty = name)) +
  scale_x_log10() +
  labs(
    x = "Rang de l'espèce",
    y = "Probabilité",
    color = "Carré",
    lty = "Carré"
  )
```

Les coordonnées des points du profil sont définies par

$$y(x) = \sum_{s=x+1}^{S} p_{[s]},\ x \in \{0, 1, \dots, S\}. \tag{4.30}$$

$p_{[s]}$ est la probabilité de l'espèce s; les espèces sont classées par probabilité décroissante.

Ce profil est exhaustif (toutes les espèces sont représentées) alors que les autres profils de diversité ne sont représentés que pour un intervalle restreint du paramètre et qu'un croisement de courbes peut se produire au-delà. En revanche, il ne prend pas en compte les espèces non observées.

Fattorini and Marcheselli[77] proposent un test pour comparer deux profils de queue de distribution à partir d'échantillonnages multiples (nécessaires pour évaluer la variance de chacune des probabilités) mais qui néglige les espèces non observées.

[77] L. Fattorini and M. Marcheselli. "Inference on Intrinsic Diversity Profiles of Biological Populations." In: *Environmetrics* 10.5 (1999), pp. 589–599. DOI: 10.1002/(SICI)1099-095X(199909/10)10:5<589::AID-ENV374>3.0.CO;2-0.

Bibliography

Adelman, M. A. "Comment on the "H" Concentration Measure as a Numbers-Equivalent." In: *The Review of Economics and Statistics* 51.1 (1969), pp. 99–101. DOI: 10.2307/1926955 (cit. on p. 66).

Agapow, P. M., O. R. P. Binindal-Emonds, K. A. Crandall, J. L. Gittleman, G. M. Mace, J. C. Marshall, and A. Purvis. "The Impact of Species Concept on Biodiversity Studies." In: *The Quarterly Review of Biology* 79.2 (2004), pp. 161–179. DOI: 10.1086/383542 (cit. on p. 14).

Ashbridge, J. and I. B. J. Goudie. "Coverage-Adjusted Estimators for Mark-Recapture in Heterogeneous Populations." In: *Communications in Statistics - Simulation and Computation* 29.4 (2000), pp. 1215–1237. DOI: 10.1080/03610910008813661 (cit. on pp. 52, 73).

Balmford, A., M. J. B. Green, and M. G. Murray. "Using Higher-Taxon Richness as a Surrogate for Species Richness: I. Regional Tests." In: *Proceedings of the Royal Society of London, Series B: Biological Sciences* 263 (1996), pp. 1267–1274. DOI: 10.1098/rspb.1996.0186 (cit. on p. 46).

Balmford, A., R. E. Green, and M. Jenkins. "Measuring the Changing State of Nature." In: *Trends in Ecology & Evolution* 18.7 (2003), pp. 326–330. DOI: 10.1016/S0169-5347(03)00067-3 (cit. on p. 25).

Balmford, A., A. H. M. Jayasuriya, and M. J. B. Green. "Using Higher-Taxon Richness as a Surrogate for Species Richness: II. Local Applications." In: *Proceedings of the Royal Society of London, Series B: Biological Sciences* 263 (1996), pp. 1571–1575. DOI: 10.1098/rspb.1996.0230 (cit. on p. 46).

Barberousse, A. and S. Samadi. "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, 2014, pp. 155–182 (cit. on p. 14).

Basharin, G. P. "On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables." In: *Theory of Probability and its Applications* 4.3 (1959), pp. 333–336. DOI: 10.1137/1104033 (cit. on p. 52).

Basset, Y., L. Cizek, P. Cuénoud, R. K. Didham, F. Guilhaumon, O. Missa, V. Novotny, F. Ødegaard, T. Roslin, J. Schmidl, A. K. Tishechkin, N. N. Winchester, D. W. Roubik, H.-P. Aberlenc, J. Bail, H. Barrios, J. R. Bridle, G. Castaño-Meneses, B. Corbara, G. Curletti, W. Duarte da Rocha, D. De Bakker, J. H. C. Delabie, A. Dejean, L. L. Fagan, A. Floren, R. L. Kitching, E. Medianero, S. E. Miller, E. Gama de Oliveira, J. Orivel, M. Pollet, M. Rapp, S. P. Ribeiro, Y. Roisin, J. B. Schmidt, L. Sørensen, and M. Leponce. "Arthropod Diversity in a Tropical Forest." In: *Science* 338.6113 (2012), pp. 1481–1484. DOI: 10.1126/science.1226727 (cit. on p. 26).

Beck, J. and W. Schwanghart. "Comparing Measures of Species Diversity from Incomplete Inventories: An Update." In: *Methods in Ecology and Evolution* 1.1 (2010), pp. 38–44. DOI: 10.1111/j.2041-210X.2009.00003.x (cit. on pp. 36, 53).

Béguinot, J. "An Algebraic Derivation of Chao's Estimator of the Number of Species in a Community Highlights the Condition Allowing Chao to Deliver Centered Estimates." In: *International Scholarly Research Notices* 2014 (Article ID 847328 2014). DOI: 10.1155/2014/847328 (cit. on p. 28).

— "Extrapolation of the Species Accumulation Curve for Incomplete Species Samplings: A New Nonparametric Approach to Estimate the Degree of Sample Completeness and Decide When to Stop Sampling." In: *Annual Research & Review in Biology* 8.5 (2015), pp. 1–9. DOI: 10.9734/ARRB/2015/22351 (cit. on p. 38).

— "Basic Theoretical Arguments Advocating Jackknife-2 as Usually Being the Most Appropriate Nonparametric Estimator of Total Species Richness." In: *Annual Research & Review in Biology* 10.1 (2016), pp. 1–12. DOI: 10.9734/ARRB/2016/25104 (cit. on pp. 34, 38).

Berger, W. H. and F. L. Parker. "Diversity of Planktonic Foraminifera in Deep-Sea Sediments." In: *Science* 168.3937 (1970), pp. 1345–1347. DOI: 10.1126/science.168.3937.1345 (cit. on p. 71).

Blandin, P. "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, 2014, pp. 31–68 (cit. on p. v).

Bonachela, J. A., H. Hinrichsen, and M. A. Muñoz. "Entropy Estimates of Small Data Sets." In: *Journal of Physics A: Mathematical and Theoretical* 41.202001 (2008), pp. 1–9. DOI: 10.1088/1751-8113/41/20/202001 (cit. on p. 54).

Box, G. E. P. and D. R. Cox. "An Analysis of Transformations." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 26.2 (1964), pp. 211–252. DOI: 10.1111/j.2517-6161.1964.tb00553.x. JSTOR: 2984418 (cit. on p. 68).

Brillouin, L. *Science and Information Theory*. 2nd ed. Oxford: Academic Press, 1962 (cit. on p. 51).

Brose, U., N. D. Martinez, and R. J. Williams. "Estimating Species Richness: Sensitivity to Sample Coverage and Insensitivity to Spatial Patterns." In: *Ecology* 84.9 (2003), pp. 2364–2377. DOI: 10.1890/02-0558 (cit. on pp. 36, 38, 72).

Bulmer, M. G. "On Fitting the Poisson Lognormal Distribution to Species-Abundance Data." In: *Biometrics* 30.1 (1974), pp. 101–110. DOI: 10.2307/1939021 (cit. on pp. 17, 52).

Burnham, K. P. and W. S. Overton. "Estimation of the Size of a Closed Population When Capture Probabilities Vary among Animals." In: *Biometrika* 65.3 (1978), pp. 625–633. DOI: 10.2307/2335915 (cit. on p. 30).

— "Robust Estimation of Population Size When Capture Probabilities Vary among Animals." In: *Ecology* 60.5 (1979), pp. 927–936. DOI: 10.2307/1936861 (cit. on pp. 26, 30, 72).

Caldarelli, G., A. Capocci, P. De Los Rios, and M. A. Muñoz. "Scale-Free Networks from Varying Vertex Intrinsic Fitness." In: *Physical Review Letters* 89.25 (2002), p. 258702. DOI: 10.1103/PhysRevLett.89.258702 (cit. on p. 46).

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. "Biodiversity Loss and Its Impact on Humanity." In: *Nature* 486.7401 (2012), pp. 59–67. DOI: 10.1038/nature11148 (cit. on p. vi).

Cartozo, C. C., D. Garlaschelli, C. Ricotta, M. Barthélemy, and G. Caldarelli. "Quantifying the Taxonomic Diversity in Real Species Communities." In: *Journal of Physics A: Mathematical and Theoretical* 41 (2008), p. 224012. DOI: 10.1088/1751-8113/41/22/224012 (cit. on p. 46).

Casetta, E. "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, 2014, pp. 139–154 (cit. on p. 13).

Ceballos, G., P. R. Ehrlich, and R. Dirzo. "Biological Annihilation via the Ongoing Sixth Mass Extinction Signaled by Vertebrate Population Losses and Declines." In: *Proceedings of the National Academy of Sciences* (2017), p. 201704949. DOI: 10.1073/pnas.1704949114 (cit. on p. vi).

Chao, A. "Nonparametric Estimation of the Number of Classes in a Population." In: *Scandinavian Journal of Statistics* 11.4 (1984), pp. 265–270. JSTOR: 4615964 (cit. on pp. 26, 27, 31).

— "Estimating the Population Size for Capture-Recapture Data with Unequal Catchability." In: *Biometrics* 43.4 (1987), pp. 783–791. DOI: 10.2307/2531532 (cit. on p. 29).

— "Species Richness Estimation." In: *Encyclopedia of Statistical Sciences*. Ed. by N. Balakrishnan, C. B. Read, and B. Vidakovic. 2nd ed. New York: Wiley, 2004 (cit. on p. 28).

Chao, A., C.-H. Chiu, and L. Jost. "Phylogenetic Diversity Measures Based on Hill Numbers." In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 365.1558 (2010), pp. 3599–3609. DOI: 10.1098/rstb.2010.0272Supplementary (cit. on p. 70).

Chao, A., R. K. Colwell, C.-H. Chiu, and D. Townsend. "Seen Once or More than Once: Applying Good-Turing Theory to Estimate Species Richness Using Only Unique Observations and a Species List." In: *Methods in Ecology and Evolution* 8.10 (2017), pp. 1221–1232. DOI: 10.1111/2041-210X.12768 (cit. on p. 29).

Chao, A., N. J. Gotelli, T. C. Hsieh, E. L. Sander, K. H. Ma, R. K. Colwell, and A. M. Ellison. "Rarefaction and Extrapolation with Hill Numbers: A Framework for Sampling and Estimation in Species Diversity Studies." In: *Ecological Monographs* 84.1 (2014), pp. 45–67. DOI: 10.1890/13-0133.1 (cit. on p. 58).

Chao, A., T. C. Hsieh, R. L. Chazdon, R. K. Colwell, and N. J. Gotelli. "Unveiling the Species-Rank Abundance Distribution by Generalizing Good-Turing Sample Coverage Theory." In: *Ecology* 96.5 (2015), pp. 1189–1201. DOI: 10.1890/14-0550.1 (cit. on pp. 72, 73).

Chao, A. and L. Jost. "Coverage-Based Rarefaction and Extrapolation: Standardizing Samples by Completeness Rather than Size." In: *Ecology* 93.12 (2012), pp. 2533–2547. DOI: 10.1890/11-1952.1 (cit. on pp. 9, 54).

— "Estimating Diversity and Entropy Profiles via Discovery Rates of New Species." In: *Methods in Ecology and Evolution* 6.8 (2015), pp. 873–882. DOI: 10.1111/2041-210X.12349 (cit. on p. 74).

Chao, A. and S.-M. Lee. "Estimating the Number of Classes via Sample Coverage." In: *Journal of the American Statistical Association* 87.417 (1992), pp. 210–217. DOI: 10.1080/01621459.1992.10475194 (cit. on p. 29).

Chao, A., S.-M. Lee, and T.-C. Chen. "A Generalized Good's Nonparametric Coverage Estimator." In: *Chinese Journal of Mathematics* 16 (1988), pp. 189–199. JSTOR: 43836340 (cit. on pp. 10, 11).

Chao, A. and C.-W. Lin. "Nonparametric Lower Bounds for Species Richness and Shared Species Richness under Sampling without Replacement." In: *Biometrics* 68.3 (2012), pp. 912–921. DOI: 10.1111/j.1541-0420.2011.01739.x (cit. on p. 12).

Chao, A., K. H. Ma, T. C. Hsieh, and C.-H. Chiu. "SpadeR: Species Prediction and Diversity Estimation with R." In: (2016) (cit. on pp. 11, 34).

Chao, A. and T.-J. Shen. "Nonparametric Estimation of Shannon's Index of Diversity When There Are Unseen Species in Sample." In: *Environmental and Ecological Statistics* 10.4 (2003), pp. 429–443. DOI: 10.1023/A:1026096204727 (cit. on pp. 52, 73).

— *Program SPADE: Species Prediction and Diversity Estimation. Program and User's Guide*. CARE, 2010 (cit. on pp. 11, 30).

Chao, A., Y.-T. Wang, and L. Jost. "Entropy and the Species Accumulation Curve: A Novel Entropy Estimator via Discovery Rates of New Species." In: *Methods in Ecology and Evolution* 4.11 (2013), pp. 1091–1100. DOI: 10.1111/2041-210x.12108 (cit. on pp. 54, 73, 74).

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. "Consequences of Changing Biodiversity." In: *Nature* 405.6783

(2000), pp. 234–242. DOI: 10.1038/35012241 (cit. on p. vi).

Chiu, C.-H., Y.-T. Wang, B. A. Walther, and A. Chao. "An Improved Nonparametric Lower Bound of Species Richness via a Modified Good-Turing Frequency Formula." In: *Biometrics* 70.3 (2014), pp. 671–682. DOI: 10.1111/biom.12200. PMID: 24945937 (cit. on pp. 10, 29, 37).

Clarke, K. R. and R. M. Warwick. "A Further Biodiversity Index Applicable to Species Lists: Variation in Taxonomic Distinctness." In: *Marine Ecology-Progress Series* 216 (2001), pp. 265–278. DOI: 10.3354/meps216265 (cit. on p. 25).

Clench, H. K. "How to Make Regional Lists of Butterflies: Some Thoughts." In: *Journal of the Lepidopterists' Society* 33.4 (1979), pp. 216–231 (cit. on p. 41).

Colwell, R. K. and J. A. Coddington. "Estimating Terrestrial Biodiversity through Extrapolation." In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 345.1311 (1994), pp. 101–118. DOI: 10.1098/rstb.1994.0091 (cit. on pp. 34, 42).

Conceição, P. and P. Ferreira. *The Young Person's Guide to the Theil Index: Suggesting Intuitive Interpretations and Exploring Analytical Applications*. Austin, Texas, 2000, p. 54 (cit. on pp. 51, 64).

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

Connell, J. H. "Diversity in Tropical Rain Forests and Coral Reefs." In: *Science* 199.4335 (1978), pp. 1302–1310. DOI: 10.1126/science.199.4335.1302 (cit. on p. 48).

Cormack, R. M. "Log-Linear Models for Capture-Recapture." In: *Biometrics* 45.2 (1989), pp. 395–413. DOI: 10.2307/2531485 (cit. on p. 31).

Cracraft, J. "Species Concepts and Speciation Analysis." In: *Current Ornithology Volume 1*. Ed. by R. F. Johnston. Vol. 1. Current Ornithology. Springer US, 1983, pp. 159–187. DOI: 10.1007/978-1-4615-6781-3_6 (cit. on p. 13).

Dalton, H. "The Measurement of the Inequality of Incomes." In: *The Economic Journal* 30.119 (1920), pp. 348–361. DOI: 10.2307/2223525 (cit. on p. 65).

Daróczy, Z. "Generalized Information Functions." In: *Information and Control* 16.1 (1970), pp. 36–51. DOI: 10.1016/s0019-9958(70)80040-7 (cit. on p. 67).

Dauby, G. and O. J. Hardy. "Sampled-Based Estimation of Diversity Sensu Stricto by Transforming Hurlbert Diversities into Effective Number of Species." In: *Ecography* 35.7 (2012), pp. 661–672. DOI: 10.1111/j.1600-0587.2011.06860.x (cit. on pp. 9, 58, 59, 70).

Davis, H. T. *The Theory of Econometrics*. Bloomington, Indiana: The Principia Press, 1941 (cit. on p. 61).

DeLong, D. C. J. "Defining Biodiversity." In: *Wildlife Society Bulletin* 24.4 (1996), pp. 738–749. JSTOR: 3783168 (cit. on p. v).

Delord, J. "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, 2014, pp. 83–118. DOI: 10.3917/edmat.delor.2014.01.0083 (cit. on p. v).

Dengler, J. "Which Function Describes the Species-Area Relationship Best? A Review and Empirical Evaluation." In: *Journal of Biogeography* 36.4 (2009), pp. 728–744. DOI: 10.1111/j.1365-2699.2008.02038.x (cit. on p. 8).

Devictor, V., D. Mouillot, C. Meynard, F. Jiguet, W. Thuiller, and N. Mouquet. "Spatial Mismatch and Congruence between Taxonomic, Phylogenetic and Functional Diversity: The Need for Integrative Conservation Strategies in a Changing World." In: *Ecology letters* 13.8 (2010), pp. 1030–40. DOI: 10.1111/j.1461-0248.2010.01493.x. PMID: 20545736 (cit. on p. 25).

Dobzhansky, T. *Genetics and the Origin of Species*. New York: Columbia University Press, 1937 (cit. on p. 13).

Ellerman, D. "An Introduction to Logical Entropy and Its Relation to Shannon Entropy." In: *International Journal of Semantic Computing* 7.02 (2013), pp. 121–145. DOI: 10.1142/S1793351X13400059 (cit. on p. 49).

Engen, S. and R. Lande. "Population Dynamic Models Generating the Lognormal Species Abundance Distribution." In: *Mathematical Biosciences* 132.2 (1996), pp. 169–183. DOI: 10.1016/0025-5564(95)00054-2 (cit. on p. 18).

Eren, M. I., A. Chao, W.-H. Hwang, and R. K. Colwell. "Estimating the Richness of a Population When the Maximum Number of Classes Is Fixed: A Nonparametric Solution to an Archaeological Problem." In: *Plos One* 7.5 (2012). DOI: 10.1371/journal.pone.0034179 (cit. on p. 29).

Esty, W. W. "A Normal Limit Law for a Nonparametric Estimator of the Coverage of a Random Sample." In: *The Annals of Statistics* 11.3 (1983), pp. 905–912. DOI: 10.2307/2240652. JSTOR: 2240652 (cit. on p. 11).

Fattorini, L. and M. Marcheselli. "Inference on Intrinsic Diversity Profiles of Biological Populations." In: *Environmetrics* 10.5 (1999), pp. 589–599. DOI: 10.1002/(SICI)1099-095X(199909/10)10:5<589::AID-ENV374>3.0.CO;2-0 (cit. on p. 76).

Fisher, R. A., A. S. Corbet, and C. B. Williams. "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 (1943), pp. 42–58. DOI: 10.2307/1411 (cit. on pp. 16, 26).

Gadagkar, R. "An Undesirable Property of Hill's Diversity Index N2." In: *Oecologia* 80 (1989), pp. 140–141. DOI: 10.1007/BF00789944 (cit. on p. 71).

Gaston, K. J. "Global Patterns in Biodiversity." In: *Nature* 405.6783 (2000), pp. 220–227. DOI: 10.1038/35012228 (cit. on pp. 5, 6).

Gini, C. *Variabilità e Mutabilità*. Bologna: C. Cuppini, 1912 (cit. on p. 49).

Good, I. J. "The Population Frequency of Species and the Estimation of Population Parameters." In: *Biometrika* 40.3/4 (1953), pp. 237–264. DOI: 10.1093/biomet/40.3-4.237 (cit. on pp. 9, 49).

Gourlet-Fleury, S., J. M. Guehl, and O. Laroussinie. *Ecology & Management of a Neotropical Rainforest. Lessons Drawn from Paracou, a Long-Term Experimental Research Site in French Guiana*. Paris: Elsevier, 2004 (cit. on p. vii).

Grassberger, P. "Finite Sample Corrections to Entropy and Dimension Estimates." In: *Physics Letters A* 128.6-7 (1988), pp. 369–373. DOI: 10.1016/0375-9601(88)90193-4 (cit. on pp. 55, 73).

— "Entropy Estimates from Insufficient Samplings." In: *arXiv Physics e-prints* 0307138.v2 (2003) (cit. on p. 55).

Gregorius, H.-R. "On the Concept of Effective Number." In: *Theoretical population biology* 40.2 (1991), pp. 269–83. DOI: 10.1016/0040-5809(91)90056-L. PMID: 1788824 (cit. on p. 66).

— "Linking Diversity and Differentiation." In: *Diversity* 2.3 (2010), pp. 370–394. DOI: 10.3390/d2030370 (cit. on p. 70).

— "Partitioning of Diversity : The "within Communities" Component." In: *Web Ecology* 14 (2014), pp. 51–60. DOI: 10.5194/we-14-51-2014 (cit. on p. 66).

Haegeman, B., J. Hamelin, J. Moriarty, P. Neal, J. Dushoff, and J. S. Weitz. "Robust Estimation of Microbial Diversity in Theory and in Practice." In: *The ISME journal* 7.6 (2013), pp. 1092–101. DOI: 10.1038/ismej.2013.10 (cit. on p. 18).

Hausser, J. and K. Strimmer. "Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks." In: *Journal of Machine Learning Research* 10 (2009), pp. 1469–1484 (cit. on p. 56).

Havrda, J. and F. Charvát. "Quantification Method of Classification Processes. Concept of Structural Alpha-Entropy." In: *Kybernetika* 3.1 (1967), pp. 30–35 (cit. on p. 67).

Heltshe, J. F. and N. E. Forrester. "Estimating Species Richness Using the Jackknife Procedure." In: *Biometrics* 39.1 (1983), pp. 1–11. DOI: 10.2307/2530802. JSTOR: 2530802 (cit. on p. 31).

Hey, J. "The Mind of the Species Problem." In: *Trends in Ecology & Evolution* 16.7 (2001), pp. 326–329. DOI: 10.1016/S0169-5347(01)02145-0 (cit. on pp. vi, 14).

Hill, M. O. "Diversity and Evenness: A Unifying Notation and Its Consequences." In: *Ecology* 54.2 (1973), pp. 427–432. DOI: 10.2307/1934352 (cit. on pp. 65, 66, 71, 74).

Hoffmann, S. and A. Hoffmann. "Is There a "True" Diversity?" In: *Ecological Economics* 65.2 (2008), pp. 213–215. DOI: 10.1016/j.ecolecon.2008.01.009 (cit. on p. 70).

Holdridge, L. R., W. C. Grenke, W. H. Hatheway, T. Liang, and J. A. Tosi. *Forest Environments in Tropical Life Zones*. Oxford: Pergamon Press, 1971 (cit. on pp. 44, 45).

Horvitz, D. G. and D. J. Thompson. "A Generalization of Sampling without Replacement from a Finite Universe." In: *Journal of the American Statistical Association* 47.260 (1952), pp. 663–685. DOI: 10.1080/01621459.1952.10483446 (cit. on pp. 52, 73).

Hubbell, S. P. "Estimating the Global Number of Tropical Tree Species, and Fisher's Paradox." In: *Proceedings of the National Academy of Sciences* 112.24 (2015), pp. 7343–7344. DOI: 10.1073/pnas.1507730112 (cit. on p. 47).

— *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, 2001 (cit. on pp. 17, 18).

Hurlbert, S. H. "The Nonconcept of Species Diversity: A Critique and Alternative Parameters." In: *Ecology* 52.4 (1971), pp. 577–586. DOI: 10.2307/1934145 (cit. on pp. 48, 58, 66).

Hutcheson, K. "A Test for Comparing Diversities Based on the Shannon Formula." In: *Journal of Theoretical Biology* 29 (1970), pp. 151–154. DOI: 10.1016/0022-5193(70)90124-4 (cit. on p. 52).

Hwang, W.-H., C.-W. Lin, and T.-J. Shen. "Good-Turing Frequency Estimation in a Finite Population." In: *Biometrical journal* 57.2 (2014), pp. 321–339. DOI: 10.1002/bimj.201300168 (cit. on p. 12).

Izsák, J. and S. Pavoine. "Links between the Species Abundance Distribution and the Shape of the Corresponding Rank Abundance Curve." In: *Ecological Indicators* 14.1 (2012), pp. 1–6. DOI: 10.1016/j.ecolind.2011.06.030 (cit. on p. 15).

James, W. and C. Stein. "Estimation with Quadratic Loss." In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Ed. by J. Neyman. Vol. 1. Berkeley, California: University of California Press, 1961, pp. 361–379 (cit. on p. 56).

Jensen, J. L. W. V. "Sur les fonctions convexes et les inégalités entre les valeurs moyennes." In: *Acta Mathematica* 30.1 (1906), pp. 175–193. DOI: 10.1007/bf02418571 (cit. on p. 54).

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

Jost, L. "Entropy and Diversity." In: *Oikos* 113.2 (2006), pp. 363–375. DOI: 10.1111/j.2006.0030-1299.14714.x (cit. on pp. 67, 70).

— "Partitioning Diversity into Independent Alpha and Beta Components." In: *Ecology* 88.10 (2007), pp. 2427–2439. DOI: 10.1890/06-1736.1 (cit. on pp. 67, 70).

— "Mismeasuring Biological Diversity: Response to Hoffmann and Hoffmann (2008)." In: *Ecological Economics* 68 (2009), pp. 925–928. DOI: 10.1016/j.ecolecon.2008.10.015 (cit. on p. 70).

— "The Relation between Evenness and Diversity." In: *Diversity* 2.2 (2010), pp. 207–232. DOI: 10.3390/d2020207 (cit. on p. 5).

Keylock, C. J. "Simpson Diversity and the Shannon-Wiener Index as Special Cases of a Generalized Entropy." In: *Oikos* 109.1 (2005), pp. 203–207. DOI: 10.1111/j.0030-1299.2005.13735.x (cit. on p. 67).

Kindt, R., P. Van Damme, and A. J. Simons. "Tree Diversity in Western Kenya: Using Profiles to Characterise Richness and Evenness." In: *Biodiversity and Conservation* 15.4 (2006), pp. 1253–1270. DOI: 10.1007/s10531-005-0772-x (cit. on p. 74).

Kullback, S. and R. A. Leibler. "On Information and Sufficiency." In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86. JSTOR: 2236703 (cit. on p. 63).

Lande, R. "Statistics and Partitioning of Species Diversity, and Similarity among Multiple Commu-

nities." In: *Oikos* 76.1 (1996), pp. 5–13. DOI: 10.2307/3545743 (cit. on p. 49).

Lande, R., P. J. DeVries, and T. R. Walla. "When Species Accumulation Curves Intersect: Implications for Ranking Diversity Using Small Samples." In: *Oikos* 89.3 (2000), pp. 601–605. DOI: 10.1034/j.1600-0706.2000.890320.x (cit. on p. 74).

Leinster, T. and C. Cobbold. "Measuring Diversity: The Importance of Species Similarity." In: *Ecology* 93.3 (2012), pp. 477–489. DOI: 10.1890/10-2402.1 (cit. on pp. 58, 74).

Lineweaver, H. and D. Burk. "The Determination of Enzyme Dissociation Constants." In: *Journal of the American Chemical Society* 56.3 (1934), pp. 658–666. DOI: 10.1021/ja01318a036 (cit. on p. 43).

Liu, C., R. J. Whittaker, K. Ma, and J. R. Malcolm. "Unifying and Distinguishing Diversity Ordering Methods for Comparing Communities." In: *Population Ecology* 49.2 (2006), pp. 89–100. DOI: 10.1007/s10144-006-0026-0 (cit. on p. 75).

Liu, D., D. Wang, Y. Wang, J. Wu, V. P. Singh, X. Zeng, L. Wang, Y. Chen, X. Chen, L. Zhang, and S. Gu. "Entropy of Hydrological Systems under Small Samples: Uncertainty and Variability." In: *Journal of Hydrology* 532 (2016), pp. 163–176. DOI: 10.1016/j.jhydrol.2015.11.019 (cit. on p. 58).

Loreau, M. "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, 2005, pp. 254–256 (cit. on p. v).

Maasoumi, E. "A Compendium to Information Theory in Economics and Econometrics." In: *Econometric Reviews* 12.2 (1993), pp. 137–181. DOI: 10.1080/07474939308800260 (cit. on p. 61).

MacArthur, R. H. "Fluctuations of Animal Populations and a Measure of Community Stability." In: *Ecology* 36.3 (1955), pp. 533–536. DOI: doi:10.2307/1929601 (cit. on p. 65).

— "On the Relative Abundance of Bird Species." In: *Proceedings of the National Academy of Sciences of the United States of America* 43.3 (1957), pp. 293–295. DOI: 10.1073/pnas.43.3.293. JSTOR: 89566 (cit. on pp. 16, 17).

— "Patterns of Species Diversity." In: *Biological Reviews* 40.4 (1965), pp. 510–533. DOI: 10.1111/j.1469-185X.1965.tb00815.x (cit. on p. 66).

Magurran, A. E. *Ecological Diversity and Its Measurement*. Princeton, NJ: Princeton University Press, 1988 (cit. on pp. 16, 18).

Mao, C. X. "Estimating Species Accumulation Curves and Diversity Indices." In: *Statistica Sinica* 17 (2007), pp. 761–774 (cit. on p. 59).

Mao, C. X. and R. K. Colwell. "Estimation of Species Richness: Mixture Models, the Role of Rare Species, and Inferential Challenges." In: *Ecology* 86.5 (2005), pp. 1143–1153. DOI: 10.1890/04-1078 (cit. on p. 27).

Marcon, E. "Practical Estimation of Diversity from Abundance Data." In: *HAL* 01212435 (version 2 2015) (cit. on p. 38).

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

Marcon, E., B. Hérault, C. Baraloto, and G. Lang. "The Decomposition of Shannon's Entropy and a Confidence Interval for *Beta* Diversity." In: *Oikos* 121.4 (2012), pp. 516–522. DOI: 10.1111/j.1600-0706.2011.19267.x (cit. on p. 64).

Marcon, E., I. Scotti, B. Hérault, V. Rossi, and G. Lang. "Generalization of the Partitioning of Shannon Diversity." In: *Plos One* 9.3 (2014), e90289. DOI: 10.1371/journal.pone.0090289 (cit. on pp. 69, 73).

May, R. M. "Patterns of Species Abundance and Diversity." In: *Ecology and Evolution of Communities*. Ed. by M. L. Cody and J. M. Diamond. Harvard University Press, 1975, pp. 81–120 (cit. on pp. 17, 18).

— "Why Worry about How Many Species and Their Loss?" In: *PLoS Biology* 9.8 (2011), e1001130. DOI: 10.1371%2Fjournal.pbio.1001130 (cit. on p. 47).

Mayden, R. L. "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, 1997, pp. 381–424 (cit. on p. 13).

Mayr, E. *Systematics and the Origin of Species from the Viewpoint of a Zoologist*. New York: Columbia University Press, 1942 (cit. on p. 13).

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. "Species Abundance Distributions: Moving beyond Single Prediction Theories to Integration within an Ecological Framework." In: *Ecology Letters* 10.10 (2007), pp. 995–1015. DOI: 10.1111/j.1461-0248.2007.01094.x (cit. on p. 16).

Mcintosh, R. P. "An Index of Diversity and the Relation of Certain Concepts to Diversity." In: *Ecology* 48.3 (1967), pp. 392–404. DOI: 10.2307/1932674 (cit. on p. 3).

Meine, C., M. E. Soulé, and R. F. Noss. ""A Mission-Driven Discipline": The Growth of Conservation Biology." In: *Conservation Biology* 20.3 (2006), pp. 631–651. DOI: 10.1111/j.1523-1739.2006.00449.x (cit. on p. v).

Mendes, R. S., L. R. Evangelista, S. M. Thomaz, A. A. Agostinho, and L. C. Gomes. "A Unified Index to Measure Ecological Diversity and Species Rarity." In: *Ecography* 31.4 (2008), pp. 450–456. DOI: 10.1111/j.0906-7590.2008.05469.x (cit. on pp. 48, 67).

Michaelis, L. and M. L. Menten. "Die Kinetik Der Invertinwirkung." In: *Biochemische Zeitschrift* 49 (1913), pp. 333–369. PMID: 21888353 (cit. on p. 41).

Miller, G. A. "Note on the Bias of Information Estimates." In: *Information Theory in Psychology: Problems and Methods*. Ed. by H. Quastler. Glencoe, Ill.: Free Press, 1955, pp. 95–100 (cit. on p. 52).

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. "An Anthropocene Map of Genetic Diversity." In: *Science* 353.6307 (2016), pp. 1532–1535. DOI: 10.1126/science.aaf4381 (cit. on p. 6).

Mora, C., D. P. Tittensor, S. Adl, A. G. B. Simpson, and B. Worm. "How Many Species Are There

on Earth and in the Ocean?" In: *PLoS Biology* 9.8 (2011), e1001127. DOI: 10.1371/journal.pbio.1001127 (cit. on p. 47).

Moreno, C. E. and P. Rodríguez. "A Consistent Terminology for Quantifying Species Diversity?" In: *Oecologia* 163.2 (2010), pp. 279–82. DOI: 10.1007/s00442-010-1591-7 (cit. on p. 6).

Motomura, I. "On the statistical treatment of communities." In: *Zoological Magazine* 44 (1932), pp. 379–383 (cit. on pp. 16, 18).

Mouillot, D. and A. Leprêtre. "A Comparison of Species Diversity Estimators." In: *Researches on Population Ecology* 41.2 (1999), pp. 203–215. DOI: 10.1007/s101440050024 (cit. on p. 26).

Mouillot, D., W. Stubbs, M. Faure, O. Dumay, J.-A. Tomasini, J. B. Wilson, and T. D. Chi. "Niche Overlap Estimates Based on Quantitative Functional Traits: A New Family of Non-Parametric Indices." In: *Oecologia* 145.3 (2005), pp. 345–353. DOI: 10.1007/s00442-005-0151-z (cit. on p. 4).

Norris, J. L. and K. H. Pollock. "Non-Parametric MLE for Poisson Species Abundance Models Allowing for Heterogeneity between Species." In: *Environmental and Ecological Statistics* 5.4 (1998), pp. 391–402. DOI: 10.1023/A:1009659922745 (cit. on p. 40).

O'Hara, R. B. "Species Richness Estimators: How Many Species Can Dance on the Head of a Pin?" In: *Journal of Animal Ecology* 74 (2005), pp. 375–386. DOI: 10.1111/j.1365-2656.2005.00940.x (cit. on p. 26).

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. "Vegan: Community Ecology Package." In: (2012) (cit. on p. vii).

Olszewski, T. D. "A Unified Mathematical Framework for the Measurement of Richness and Evenness within and among Multiple Communities." In: *Oikos* 104.2 (2004), pp. 377–387. DOI: 10.1111/j.0030-1299.2004.12519.x (cit. on p. 48).

Pallmann, P., F. Schaarschmidt, L. A. Hothorn, C. Fischer, H. Nacke, K. U. Priesnitz, and N. J. Schork. "Assessing Group Differences in Biodiversity by Simultaneously Testing a User-Defined Selection of Diversity Indices." In: *Molecular Ecology Resources* 12.6 (2012), pp. 1068–1078. DOI: 10.1111/1755-0998.12004 (cit. on p. 75).

Patil, G. P. and C. Taillie. "Diversity as a Concept and Its Measurement." In: *Journal of the American Statistical Association* 77.379 (1982), pp. 548–561. DOI: 10.2307/2287709. JSTOR: 2287709 (cit. on pp. 62, 65, 67, 74, 75).

Pavoine, S. and M. B. Bonsall. "Measuring Biodiversity to Explain Community Assembly: A Unified Approach." In: *Biological Reviews* 86.4 (2011), pp. 792–812. DOI: 10.1111/j.1469-185X.2010.00171.x (cit. on p. 4).

Peet, R. K. "The Measurement of Species Diversity." In: *Annual review of ecology and systematics* 5 (1974), pp. 285–307. DOI: 10.1146/annurev.es.05.110174.001441 (cit. on p. 25).

Pernès, J., ed. *Gestion Des Ressources Génétiques Des Plantes. Tome 2 : Manuel*. Paris: Agence de Coopération culturelle et technique, 1984 (cit. on p. 14).

Pielou, E. C. "Species-Diversity and Pattern-Diversity in the Study of Ecological Succession." In: *Journal of Theoretical Biology* 10.2 (1966), pp. 370–383. DOI: 10.1016/0022-5193(66)90133-0 (cit. on p. 53).

— "The Measurement of Diversity in Different Types of Biological Collections." In: *Journal of Theoretical Biology* 13.C (1966), pp. 131–144. DOI: 10.1016/0022-5193(66)90013-0 (cit. on p. 51).

— *Ecological Diversity*. New York: Wiley, 1975 (cit. on p. 61).

Preston, F. W. "The Commonness, and Rarity, of Species." In: *Ecology* 29.3 (1948), pp. 254–283. DOI: 10.2307/1930989 (cit. on pp. 15, 16, 39).

Pueyo, S., F. He, and T. Zillio. "The Maximum Entropy Formalism and the Idiosyncratic Theory of Biodiversity." In: *Ecology letters* 10.11 (2007), pp. 1017–28. DOI: 10.1111/j.1461-0248.2007.01096.x (cit. on p. 18).

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

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2024 (cit. on p. vii).

Raaijmakers, J. G. W. "Statistical Analysis of the Michaelis-Menten Equation." In: *Biometrics* 43.4 (1987), pp. 793–803. DOI: 10.2307/2531533 (cit. on pp. 42, 43).

Rényi, A. "On Measures of Entropy and Information." In: *4th Berkeley Symposium on Mathematical Statistics and Probability*. Ed. by J. Neyman. Vol. 1. Berkeley, USA: University of California Press, 1961, pp. 547–561 (cit. on p. 65).

Richards, R. A. *The Species Problem. A Philosophical Analysis*. Cambridge: Cambridge University Press, 2010 (cit. on p. 13).

Ricotta, C. "On Parametric Diversity Indices in Ecology: A Historical Note." In: *Community Ecology* 6.2 (2005), pp. 241–244. DOI: 10.1556/ComEc.6.2005.2.12 (cit. on p. 67).

— "Through the Jungle of Biological Diversity." In: *Acta Biotheoretica* 53.1 (2005), pp. 29–38. DOI: 10.1007/s10441-005-7001-6 (cit. on p. vi).

— "A Semantic Taxonomy for Diversity Measures." In: *Acta Biotheoretica* 55.1 (2007), pp. 23–33. DOI: 10.1007/s10441-007-9008-7 (cit. on p. 4).

Ricotta, C. and G. C. Avena. "An Information-Theoretical Measure of Taxonomic Diversity." In: *Acta biotheoretica* 25.51 (2003), pp. 35–41. DOI: 10.1023/A:1023000322071 (cit. on pp. 25, 67).

Ricotta, C. and L. Szeidl. "Towards a Unifying Approach to Diversity Measures: Bridging the Gap between the Shannon Entropy and Rao's Quadratic Index." In: *Theoretical Population Biology* 70.3 (2006), pp. 237–243. DOI: 10.1016/j.tpb.2006.06.003 (cit. on p. 67).

Runnegar, B. "Rates and Modes of Evolution in the Mollusca." In: *Rates of Evolution*. Ed. by M. Campbell and M. F. Day. London: Allen & Unwin, 1987, pp. 39–60 (cit. on p. 5).

Scheiner, S. M. "Six Types of Species-Area Curves." In: *Global Ecology and Biogeography* 12.6 (2003), pp. 441–447. DOI: 10.1046/j.1466-822X.2003.00061.x (cit. on p. 8).

Schürmann, T. "Bias Analysis in Entropy Estimation." In: *Journal of Physics A: Mathematical*

and General 37.27 (2004), pp. L295–L301. DOI: 10.1088/0305-4470/37/27/L02 (cit. on p. 56).

Shannon, C. E. "A Mathematical Theory of Communication." In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423, 623–656. DOI: 10.1002/j.1538-7305.1948.tb01338.x (cit. on pp. 50, 61).

Shannon, C. E. and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1963 (cit. on pp. 50, 61).

Shen, T.-J., A. Chao, and C.-F. Lin. "Predicting the Number of New Species in a Further Taxonomic Sampling." In: *Ecology* 84.3 (2003), pp. 798–804. DOI: 10.1890/0012-9658(2003)084[0798:PTNONS]2.0.CO;2 (cit. on p. 38).

Simpson, E. H. "Measurement of Diversity." In: *Nature* 163.4148 (1949), p. 688. DOI: 10.1038/163688a0 (cit. on p. 48).

Slik, J. W. F. et al. "An Estimate of the Number of Tropical Tree Species." In: *Proceedings of the National Academy of Sciences of the United States of America* 112.24 (2015), pp. 7472–7477. DOI: 10.1073/pnas.1423147112 (cit. on p. 47).

Smith, E. P. and G. V. Belle. "Nonparametric Estimation of Species Richness." In: *Biometrics* 40.1 (1984), pp. 119–129. DOI: 10.1002/9780470015902.a0026329 (cit. on p. 34).

Smith, W. and J. F. Grassle. "Sampling Properties of a Family of Diversity Measures." In: *Biometrics* 33.2 (1977), pp. 283–292. DOI: 10.2307/2529778. JSTOR: 2529778 (cit. on p. 59).

Soberón M., J. and J. Llorente B. "The Use of Species Accumulation Functions for the Prediction of Species Richness." In: *Conservation Biology* 7.3 (1993), pp. 480–488. DOI: 10.1046/j.1523-1739.1993.07030480.x (cit. on p. 44).

Spellerberg, I. F. and P. J. Feder. "A Tribute to Claude Shannon (1916–2001) and a Plea for More Rigorous Use of Species Richness, Species Diversity and the 'Shannon–Wiener' Index." In: *Global Ecology and Biogeography* 12.3 (2003), pp. 177–179. DOI: 10.1046/j.1466-822X.2003.00015.x (cit. on p. 50).

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

Stegen, J. C. and A. H. Hurlbert. "Inferring Ecological Processes from Taxonomic, Phylogenetic and Functional Trait -Diversity." In: *PloS one* 6.6 (2011), e20906. DOI: 10.1371/journal.pone.0020906 (cit. on p. 25).

Stirling, A. "A General Framework for Analysing Diversity in Science, Technology and Society." In: *Journal of the Royal Society, Interface* 4.15 (2007), pp. 707–719. DOI: 10.1098/rsif.2007.0213 (cit. on p. 5).

Stoddart, J. A. "A Genotypic Diversity Measure." In: *Journal of Heredity* 74 (1983), pp. 489–490. DOI: 10.1093/oxfordjournals.jhered.a109852 (cit. on p. 66).

Sukumaran, J. and L. L. Knowles. "Multispecies Coalescent Delimits Structure, Not Species." In: *Proceedings of the National Academy of Sciences of the United States of America* in press (2017). DOI: 10.1073/PNAS.1607921114 (cit. on p. 13).

Ter Steege, H., N. C. A. Pitman, D. Sabatier, C. Baraloto, R. P. Salomão, J. E. Guevara, O. L. Phillips, C. V. Castilho, W. E. Magnusson, J.-F. Molino, A. Monteagudo, P. Núñez Vargas, J. C. Montero, T. R. Feldpausch, E. N. H. Coronado, T. J. Killeen, B. Mostacedo, R. Vasquez, R. L. Assis, J. Terborgh, F. Wittmann, A. C. S. Andrade, W. F. Laurance, S. G. W. Laurance, B. S. Marimon, B.-H. Marimon, I. C. Guimarães Vieira, I. L. Amaral, R. Brienen, H. Castellanos, D. Cárdenas López, J. F. Duivenvoorden, H. F. Mogollón, F. D. de Almeida Matos, N. Dávila, R. García-Villacorta, P. R. Stevenson Diaz, F. Costa, T. Emilio, C. Levis, J. Schietti, P. Souza, A. Alonso, F. Dallmeier, A. J. D. Montoya, M. T. Fernandez Piedade, A. Araujo-Murakami, L. Arroyo, R. Gribel, P. V. A. Fine, C. A. Peres, M. Toledo, G. A. Aymard, T. Baker, C. Cerón, J. Engel, T. W. Henkel, P. Maas, P. Petronelli, J. Stropp, C. E. Zartman, D. Daly, D. Neill, M. Silveira, M. R. Paredes, J. Chave, D. de Andrade Lima Filho, P. M. Jørgensen, A. Fuentes, J. Schöngart, F. Cornejo Valverde, A. Di Fiore, E. M. Jimenez, M. C. Peñuela-Mora, J. F. Phillips, G. Rivas, T. R. van Andel, P. von Hildebrand, B. Hoffman, E. L. Zent, Y. Malhi, A. Prieto, A. Rudas, A. R. Ruschell, N. Silva, V. Vos, S. Zent, A. A. Oliveira, A. C. Schutz, T. Gonzales, M. Trindade Nascimento, H. Ramirez-Angulo, R. Sierra, M. Tirado, M. N. Umaña Medina, G. van der Heijden, C. I. A. Vela, E. Vilanova Torre, C. Vriesendorp, O. Wang, K. R. Young, C. Baider, H. Balslev, C. Ferreira, I. Mesones, A. Torres-Lezama, L. E. Urrego Giraldo, R. Zagt, M. N. Alexiades, L. Hernandez, I. Huamantupa-Chuquimaco, W. Milliken, W. Palacios Cuenca, D. Pauletto, E. Valderrama Sandoval, L. Valenzuela Gamarra, K. G. Dexter, K. J. Feeley, G. Lopez-Gonzalez, and M. R. Silman. "Hyperdominance in the Amazonian Tree Flora." In: *Science* 342.6156 (2013), p. 1243092. DOI: 10.1126/science.1243092 (cit. on p. 47).

Theil, H. *Economics and Information Theory*. Chicago: Rand McNally & Company, 1967 (cit. on pp. 51, 61, 63).

Tothmeresz, B. "Comparison of Different Methods for Diversity Ordering." In: *Journal of Vegetation Science* 6.2 (1995), pp. 283–290. DOI: 10.2307/3236223 (cit. on p. 74).

Tsallis, C. "Possible Generalization of Boltzmann-Gibbs Statistics." In: *Journal of Statistical Physics* 52.1 (1988), pp. 479–487. DOI: 10.1007/BF01016429 (cit. on p. 67).

— "What Are the Numbers That Experiments Provide?" In: *Química Nova* 17.6 (1994), pp. 468–471 (cit. on p. 68).

Ulanowicz, R. E. "Information Theory in Ecology." In: *Computers & Chemistry* 25.4 (2001), pp. 393–399. DOI: 10.1016/S0097-8485(01)00073-0 (cit. on p. 65).

Van Valen, L. "Ecological Species, Multispecies, and Oaks." In: *Taxon* 25.2/3 (1976), pp. 233–239. DOI: 10.2307/1219444 (cit. on p. 13).

Vinck, M., F. P. Battaglia, V. B. Balakirsky, A. J. H. Vinck, and C. M. A. Pennartz. "Estimation of the Entropy Based on Its Polynomial Representation." In: *Physical Review E* 85.5 (2012). DOI: 10.1103/PhysRevE.85.051139 (cit. on p. 53).

Volkov, I., J. R. Banavar, S. P. Hubbell, and A. Maritan. "Neutral Theory and Relative Species Abundance in Ecology." In: *Nature* 424.6952 (2003),

pp. 1035–1037. DOI: 10.1038/nature01883 (cit. on p. 18).

Vu, V. Q., B. Yu, and R. E. Kass. "Coverage-Adjusted Entropy Estimation." In: *Statistics in Medicine* 26.21 (2007), pp. 4039–4060. DOI: 10.1002/sim.2942 (cit. on p. 53).

Wang, J.-P. "Estimating Species Richness by a Poisson-compound Gamma Model." In: *Biometrika* 97.3 (2010), pp. 727–740. DOI: 10.1093/biomet/asq026 (cit. on p. 41).

— "SPECIES: An R Package for Species Richness Estimation." In: *Journal of Statistical Software* 40.9 (2011), pp. 1–15. DOI: 10.18637/jss.v040.i09 (cit. on p. 35).

Wang, J.-P., B. Lindsay, L. Cui, P. K. Wall, J. Marion, J. Zhang, and C. DePamphilis. "Gene Capture Prediction and Overlap Estimation in EST Sequencing from One or Multiple Libraries." In: *BMC Bioinformatics* 6.1 (2005), p. 300. DOI: 10.1186/1471-2105-6-300 (cit. on p. 41).

Whittaker, R. H. "Vegetation of the Siskiyou Mountains, Oregon and California." In: *Ecological Monographs* 30.3 (1960), pp. 279–338. DOI: 10.2307/1943563. JSTOR: 1943563 (cit. on p. 5).

— "Dominance and Diversity in Land Plant Communities." In: *Science* 147.3655 (1965), pp. 250–260. DOI: 10.1126/science.147.3655.250 (cit. on pp. 3, 15).

— "Evolution and Measurement of Species Diversity." In: *Taxon* 21.2/3 (1972), pp. 213–251. DOI: 10.2307/1218190 (cit. on pp. 16, 18).

— "Evolution of Species Diversity in Land Communities." In: *Evolutionary Biology* 10 (1977). Ed. by M. K. Hecht, W. C. Steere, and B. Wallace, pp. 1–67 (cit. on p. 6).

Williams, P. H. and K. J. Gaston. "Measuring More of Biodiversity: Can Higher-Taxon Richness Predict Wholesale Species Richness?" In: *Biological Conservation* 67 (1994), pp. 211–217. DOI: 10.1016/0006-3207(94)90612-2 (cit. on p. 46).

Williamson, M. and K. J. Gaston. "The Lognormal Distribution Is Not an Appropriate Null Hypothesis for the Species-Abundance Distribution." In: *Journal of Animal Ecology* 74.2001 (2005), pp. 409–422. DOI: 10.1111/j.1365-2656.2005.00936.x (cit. on p. 40).

Williamson, M., K. J. Gaston, and W. M. Lonsdale. "The Species-Area Relationship Does Not Have an Asymptote!" In: *Journal of Biogeography* 28.7 (2001), pp. 827–830. DOI: 10.1046/j.1365-2699.2001.00603.x (cit. on p. 25).

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

Wilson, J. B., R. K. Peet, J. Dengler, and M. Pärtel. "Plant Species Richness: The World Records." In: *Journal of Vegetation Science* 23.4 (2012), pp. 796–802. DOI: 10.1111/j.1654-1103.2012.01400.x (cit. on p. 48).

Wright, S. "Evolution in Mendelian Populations." In: *Genetics* 16.2 (1931), pp. 97–159 (cit. on p. 66).

Zhang, C.-H. and Z. Zhang. "Asymptotic Normality of a Nonparametric Estimator of Sample Coverage." In: *Annals of Statistics* 37 (5A 2009), pp. 2582–2595. DOI: 10.1214/08-aos658 (cit. on p. 11).

Zhang, Z. "Entropy Estimation in Turing's Perspective." In: *Neural Computation* 24.5 (2012), pp. 1368–1389. DOI: 10.1162/NECO_a_00266 (cit. on p. 53).

— "Asymptotic Normality of an Entropy Estimator with Exponentially Decaying Bias." In: *IEEE Transactions on Information Theory* 59.1 (2013), pp. 504–508. DOI: 10.1109/TIT.2012.2217393 (cit. on pp. 53, 74).

Zhang, Z. and M. Grabchak. "Bias Adjustment for a Nonparametric Entropy Estimator." In: *Entropy* 15.6 (2013), pp. 1999–2011. DOI: 10.3390/e15061999 (cit. on pp. 53, 74).

— "Entropic Representation and Estimation of Diversity Indices." In: *Journal of Nonparametric Statistics* 28.3 (2016), pp. 563–575. DOI: 10.1080/10485252.2016.1190357 (cit. on p. 74).

Zhang, Z. and H. Huang. "Turing's Formula Revisited." In: *Journal of Quantitative Linguistics* 14.2-3 (2007), pp. 222–241. DOI: 10.1080/09296170701514189 (cit. on pp. 10, 11).

Zhang, Z. and J. Zhou. "Re-Parameterization of Multinomial Distributions and Diversity Indices." In: *Journal of Statistical Planning and Inference* 140.7 (2010), pp. 1731–1738. DOI: 10.1016/j.jspi.2009.12.023 (cit. on p. 74).

Résumé La biodiversité peut être mesurée de nombreuses façons.

La dualité entropie-diversité fournit un cadre clair et rigoureux pour le faire. L'entropie est la surprise moyenne fournie par les individus d'une communauté. Le choix de la fonction d'information qui mesure cette surprise à partir des probabilités d'occurence des espèces (ou d'autres catégories) permet de définir les mesures de diversités neutres, fonctionnelles ou phylogénétique présentées ici. L'entropie est transformée en diversité au sens strict par une fonction croissante (l'exponentielle déformée), ce qui simplifie son interprétation en tant que nombre équivalent d'espèces.

L'entropie phylogénétique généralise les indices de diversité classique, intègre si nécessaire la distance entre espèces, peut être écomposée et corrigée des biais d'estimation. Sa transformation en diversité au sens strict permet d'interpréter les valeurs sous une forme unique : un nombre équivalent d'espèces et un nombre équivalent de communautés. La diversité de Leinster et Cobbold généralise à son tour la diversité phylogénétique et permet d'autres définitions de la distance entre espèces. Le paramétrage des mesures (l'ordre de la diversité) permet de donner plus ou moins d'importance aux espèces rares et de tracer des profils de diversité.

La construction de ce cadre méthodologique est présentée en détail ainsi que plusieurs approches différentes, qui constituent l'état de l'art de la mesure de la biodiversité.

.

Unité Mixte de Recherche
ECOlogie
des
FOrêts
de
Guyane
cirad
cnrs
INRAE
Université des Antilles
Université de Guyane
AgroParisTech