Mesures de la Biodiversité

Eric Marcon

▶ **To cite this version:**

Eric Marcon. Mesures de la Biodiversité. Master. Kourou, France. 2015. cel-01205813v2

HAL Id: cel-01205813

https://agroparistech.hal.science/cel-01205813v2

Submitted on 19 Jan 2016 (v2), last revised 12 Sep 2018 (v5)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mesures de la Biodiversité

Eric Marcon

Version: 19 janvier 2016

Je remercie François Morneau pour sa relecture attentive de la première version de ce document et sa contribution décisive pour la bonne utilisation de R dans les exemples.

Bruno Hérault et Chris Baraloto ont contribué à la revue de la littérature de façon significative.

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/`) ;
- *knitr* (`http://yihui.name/knitr/`).

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

Les opinions émises par les auteurs sont personnelles et n'engagent ni l'UMR EcoFoG ni ses tutelles.

Photographie en couverture: Hadrien Lalagüe

Table des matières

Notations **vii**

1 Notions de Diversité **1**
1.1 Composantes . . . 2
1.1.1 Richesse . . . 3
1.1.2 Équitabilité . . . 3
1.1.3 Disparité . . . 4
1.1.4 Agrégation . . . 4
1.2 Niveaux de l'étude . . . 4
1.2.1 Diversité alpha, beta et gamma . . . 5
1.2.2 Décomposition . . . 5
1.3 SAD et SAR . . . 6

2 Mesures neutres de la diversité alpha ou gamma **9**
2.1 Mesures fondées sur les relations aire-espèces . . . 9
2.1.1 La relation d'Arrhenius . . . 10
2.1.2 Le modèle de Gleason . . . 11
2.1.3 Courbes d'accumulation . . . 11
2.1.4 Indice alpha de Fisher . . . 13
2.1.5 Le modèle de Plotkin . . . 18
2.1.6 Le modèle de placement aléatoire . . . 20
2.1.7 Influence de la structure spatiale . . . 21
2.1.8 Synthèse . . . 23
2.2 Taux de couverture . . . 26
2.2.1 Formule des fréquences de Good-Turing . . . 26
2.2.2 Taux de couverture et déficit de couverture . . . 26
2.3 Richesse spécifique . . . 28
2.3.1 Techniques d'estimation non paramétrique . . . 28
2.3.2 Inférence du nombre d'espèces à partir de la SAD . . . 37
2.3.3 Inférence du nombre d'espèces à partir de courbes d'accumulation . . . 38
2.3.4 Diversité générique . . . 42
2.3.5 Combien y a-t-il d'espèces différentes sur Terre ? . . . 43
2.3.6 Le problème de l'espèce . . . 44
2.4 Indice de Simpson . . . 45
2.4.1 Définition . . . 45
2.4.2 Biais d'estimation . . . 46
2.5 Indice de Shannon . . . 47
2.5.1 Définition . . . 47

2.5.2 Biais d'estimation . . . 48
2.6 Indice de Hurlbert . . . 53
2.6.1 Définition . . . 53
2.6.2 Biais d'estimation . . . 54

3 Entropie **57**
3.1 Définition de l'entropie . . . 57
3.2 Entropie relative . . . 59
3.3 Entropie et biodiversité : historique . . . 60
3.4 Entropie HCDT . . . 62
3.4.1 Logarithmes déformés . . . 62
3.4.2 Entropie et diversité . . . 64
3.4.3 Profils de diversité . . . 66
3.4.4 Biais d'estimation . . . 67
3.4.5 Raréfaction et extrapolation . . . 74
3.5 Autres approches . . . 74
3.5.1 L'entropie de Rényi . . . 74
3.5.2 L'entropie généralisée des mesures d'inégalité . . . 75
3.6 Équitabilité . . . 76
3.6.1 Approche axiomatique . . . 76
3.6.2 Indice classiques . . . 77
3.6.3 Indice alternatifs . . . 78
3.6.4 Indice de Pielou . . . 78
3.6.5 Généralisation de l'indice de Pielou . . . 79

4 Diversité fonctionnelle et phylogénétique **83**
4.1 Cadre . . . 84
4.1.1 Dissimilarité et distance . . . 84
4.1.2 Distance phylogénétique . . . 84
4.1.3 Distance fonctionnelle . . . 85
4.1.4 Équivalence des deux diversités . . . 86
4.1.5 Typologie des mesures . . . 87
4.2 Mesures particulières . . . 88
4.2.1 Richesse, équitabilité et divergence fonctionnelle . . . 88
4.2.2 Originalité, richesse et équitabilité phylogénétique . . . 92
4.2.3 FD et PD . . . 96
4.2.4 Indice de Rao . . . 97
4.2.5 Diversité et moyenne . . . 101
4.2.6 H_p et I_1 . . . 101
4.3 Entropie phylogénétique . . . 102
4.3.1 Généralisation de l'entropie HCDT . . . 102
4.3.2 Biais d'estimation . . . 102
4.3.3 Entropie et diversité . . . 103
4.3.4 Diversité individuelle . . . 105
4.3.5 Arbres non ultramétriques . . . 105
4.4 Diversité de Leinster et Cobbold . . . 105
4.4.1 Définitions . . . 105
4.4.2 Diversité neutre . . . 106
4.4.3 Diversité non neutre . . . 107

4.4.4 Entropie de Ricotta et Szeidl . . . 108
4.4.5 Définition de la similarité . . . 109
4.4.6 Biais d'estimation . . . 110
4.4.7 Diversité phylogénétique . . . 112
4.5 Synthèse . . . 112
4.6 Diversité individuelle . . . 113
4.7 Diversité des valeurs propres . . . 115
4.8 Diversité de Scheiner . . . 118

5 Diversité beta et décomposition 123
5.1 Cadre . . . 124
5.1.1 Définition de la diversité beta par objectifs . . . 124
5.1.2 Définitions de la diversité beta, mesure dérivée . . . 127
5.1.3 Le débat sur la décomposition . . . 127
5.1.4 Décomposition multiplicative de la diversité . . . 128
5.1.5 Définitions de la diversité alpha . . . 129
5.2 Décomposition de l'entropie HCDT . . . 130
5.2.1 Décomposition de l'indice de Shannon . . . 130
5.2.2 Décomposition de l'entropie généralisée . . . 133
5.2.3 Décomposition du nombre d'espèces . . . 135
5.2.4 Décomposition de l'indice de Gini-Simpson . . . 136
5.2.5 Décomposition de l'indice de Rao . . . 138
5.2.6 Synthèse . . . 140
5.3 Normalisation . . . 141
5.3.1 Nécessité de normaliser la diversité beta . . . 141
5.3.2 L'indice de chevauchement C_{qN} . . . 143
5.4 Décomposition de la diversité phylogénétique . . . 144
5.5 Partitionnement de la diversité de Leinster et Cobbold . . . 146
5.6 Autres approches . . . 147
5.6.1 L'indice de Simpson spatialement explicite . . . 147
5.6.2 Le variogramme de complémentarité . . . 148
5.6.3 Le partitionnement de la diversité selon Pélissier et Couteron . . . 150
5.7 Tests de significativité . . . 151
5.7.1 Test de non-nullité . . . 151
5.7.2 Analyse de la variabilité . . . 153

6 Diversité de paires 155
6.1 Variations sur l'entropie quadratique . . . 155
6.2 Diversité fonctionnelle . . . 156
6.3 Diversité beta . . . 156
6.3.1 Données de présence-absence . . . 156
6.3.2 Données d'abondance . . . 161
6.3.3 Données d'abondance et similarité des espèces . . . 162
6.3.4 Distance entre paires de communautés et diversité β . . . 163
6.4 Diversité de Kosman . . . 165
6.5 Tests de significativité . . . 166
6.5.1 Test de non-nullité . . . 166
6.5.2 Analyse de la variabilité . . . 167

7 Diversité jointe, diversité structurelle **169**
7.1 Récursivité de l'entropie 170
7.1.1 Entropie de Shannon 170
7.1.2 Entropie HCDT 171
7.1.3 Intérêt de l'information mutuelle 171
7.2 Décomposition de la diversité jointe 172

8 Conclusion **177**

Bibliographie **181**

Notations

Les notations peuvent différer de celles de la littérature citée pour l'homogénéité de ce document. Souvent, la littérature écologique n'est pas très rigoureuse sur la distinction de ce qui est un paramètre, une variable aléatoire ou son estimation. Par exemple, le nombre d'individus échantillonnés (choisi) devrait être noté n, un nombre d'individus aléatoire (résultat d'un processus stochastique) devrait être noté N et son estimateur $\hat{N}$. La confusion est fréquente dans la formulation des mesures de diversité et leurs estimateurs. Dans la mesure du possible, des notations plus rigoureuses que les originales sont utilisées ici.

α_ν : la probabilité moyenne des espèces représentées par ν individus.

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 $d_{s,t}$, la dissimilarité entre l'espèce s et l'espèce t.

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

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

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. Son estimateur, $\hat{p}_s$ est la fréquence observée. $p_{s,i}$ est la même probabilité dans la communauté i.

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}.

$\mathcal{P} = \{p_1, p_2, \ldots, p_s, \ldots, p_S\}$: la distribution des probabilités p_s.

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

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

S_ν^n : le nombre d'espèces, considéré comme une variable aléatoire, observées ν fois dans l'échantillonnage. L'indice est le nombre de fois où l'espèce est détectée : par exemple S_1 ou $S_{\neq 0}$. L'exposant est la taille de l'échantillon : S^A pour la surface A ou $\hat{S}^n$ pour un échantillon de n individus. S_0^A est le nombre d'espèces non rencontrées dans la surface A. Pour alléger les notations, s'il n'y a pas d'ambiguïté, l'indice est omis pour les espèces présentes : $S_{\neq 0}^A$ est noté S^A. Si l'exposant n'est pas noté, l'échantillon n'est pas précisé et peut être aussi bien un nombre d'individus qu'une surface.

s_ν^n : le nombre d'espèces observées, avec les mêmes notations que ci-dessus. s_ν^n peut être considéré comme une réalisation de S_ν^n.

$t_{1-\alpha/2}^n$: le quantile d'une loi de Student à n degrés de liberté au seuil de risque α, classiquement 1,96 pour n grand et $\alpha = 5\%$.

Chapitre 1

Notions de Diversité

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

[1] C. Meine et al. (2006). « "A mission-driven discipline" : The growth of conservation biology ». In : *Conservation Biology* 20.3, p. 631–651.

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

[3] P. Blandin (2014). « La diversité du vivant avant (et après) la biodiversité : repères historiques et épistémologiques ». In : *La biodiversité en question. Enjeux philosophiques, éthiques et scientifiques*. Sous la dir. d'E. Casetta et J. Delord. Paris : Editions Matériologiques. Chap. 1, p. 31–68.

[4] J. Delord (2014). « La biodiversité : imposture scientifique ou ruse épistémologique ? » In : *La biodiversité en question. Enjeux philosophiques, éthiques et scientifiques*. Sous la dir. d'E. Casetta et J. Delord. Paris : Editions Matériologiques. Chap. 3, p. 83–118.

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

[6] J. Speth et al. (1992). « Foreword ». In : *Global Biodiversity Strategy*. Sous la dir. de K. Courrier. Washington, D.C. : WRI, IUCN, UNEP, p. v–vi.

[7] M. Loreau (2005). « Discours de clôture ». In : *Actes de la Conférence internationale Biodiversité Science et Gouvernance*. Sous la dir. de R. Barbault et J.-P. Le Duc. Paris, France. : IRD Editions, p. 254–256.

> et avec leur environnement physique, et la variété des écosystèmes complexes 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 (autre concept flou[8]). L'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. 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.*[9] pour une revue sur cette question ou Cardinale *et al.*[10] pour les conséquences de l'érosion de la biodiversité sur les services écosystémiques). La mesure de la diversité est un sujet important en tant que tel,[11] 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.[12]

Les calculs nécessaires sont réalisés dans R,[13] essentiellement avec le package *entropart*.[14] L'ensemble du code est disponible sur le site web[15] où se trouvent les mises à jour de ce document.

1.1 Composantes

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.[16]

[8] J. Hey (2001). « The mind of the species problem ». In : *Trends in Ecology & Evolution* 16.7, p. 326–329.

[9] F. S. I. Chapin et al. (2000). « Consequences of changing biodiversity ». In : *Nature* 405.6783, p. 234–242.

[10] B. J. Cardinale et al. (2012). « Biodiversity loss and its impact on humanity ». In : *Nature* 486.7401, p. 59–67.

[11] A. Purvis et A. Hector (2000). « Getting the measure of biodiversity. » In : *Nature* 405.6783, p. 212–9.

[12] C. Ricotta (2005e). « Through the jungle of biological diversity ». In : *Acta Biotheoretica* 53.1, p. 29–38.

[13] R Core Team (2015). *R : A Language and Environment for Statistical Computing*. Vienna, Austria : R Foundation for Statistical Computing.

[14] É. Marcon et B. Hérault (2015b). « entropart, an R Package to Measure and Partition Diversity ». In : *Journal of Statistical Software* 67.8, p. 1–26.

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

[16] R. H. Whittaker (1965). « Dominance and diversity in land plant communities ». In : *Science* 147.3655, p. 250–260.

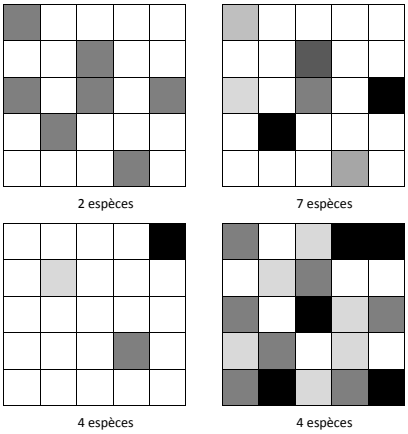


FIGURE 1.1 – Importance de la richesse (en haut) et de l'équitabilité (en bas) pour la définition de la diversité. Ligne du haut : toutes choses égales par ailleurs, une communauté de 7 espèces semble plus diverse qu'une communauté de 2 espèces. Ligne du bas : à richesse égale, une communauté moins équitable (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.

1.1.1 Richesse

La richesse (terme introduit par McIntosh[17]) 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.

[17] R. P. MCINTOSH (1967). « An Index of Diversity and the Relation of Certain Concepts to Diversity ». In : *Ecology* 48.3, p. 392–404.

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 microorganismes ;
- 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.

1.1.2 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 indices courants, comme ceux de Simpson ou de Shannon, évaluent à la fois la richesse et l'équitabilité.

1.1.3 Disparité

Les mesures classiques de la diversité, dites mesures de diversité neutre (*species-neutral diversity*) 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 phylogénétique et de diversité fonctionnelle prennent en compte cette notion, qui nécessite quelques définitions supplémentaires.[18]

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*[19] sont construites à partir de la dissimilarité entre les classes, avec ou sans pondération par la fréquence.

La disparité,[20] 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.

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).

[18] D. Mouillot et al. (2005b). « Niche overlap estimates based on quantitative functional traits : a new family of non-parametric indices ». In : *Oecologia* 145.3, p. 345–353 ; C. Ricotta (2007). « A semantic taxonomy for diversity measures ». In : *Acta Biotheoretica* 55.1, p. 23–33.

[19] S. Pavoine et M. B. Bonsall (2011). « Measuring biodiversity to explain community assembly : a unified approach ». In : *Biological Reviews* 86.4, p. 792–812.

[20] B. Runnegar (1987). « Rates and Modes of Evolution in the Mollusca ». In : *Rates of Evolution*. Sous la dir. de M. Campbell et M. F. Day. London : Allen & Unwin, p. 39–60.

1.1.4 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[21] estime que les trois composantes, qu'il nomme *variété* (richesse), *équilibre* (équitabilité) et *disparité*, recouvrent tous les aspects de la diversité.

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.[22]

[21] A. Stirling (2007). « A general framework for analysing diversity in science, technology and society. » In : *Journal of the Royal Society, Interface / the Royal Society* 4.15, p. 707–719.

[22] L. Jost (2010b). « The Relation between Evenness and Diversity ». In : *Diversity* 2.2, p. 207–232.

1.2 Niveaux de l'étude

La diversité est classiquement estimée à plusieurs niveaux emboîtés, nommés α, β et γ par Whittaker,[23] qui a nommé α la diversité locale qu'il mesurait avec l'indice α de Fisher (voir page 14) et a utilisé les lettres suivantes selon ses besoins.

[23] R. H. Whittaker (1960). « Vegetation of the Siskiyou Mountains, Oregon and California ». In : *Ecological Monographs* 30.3, p. 279–338, page 320.

1.2.1 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.

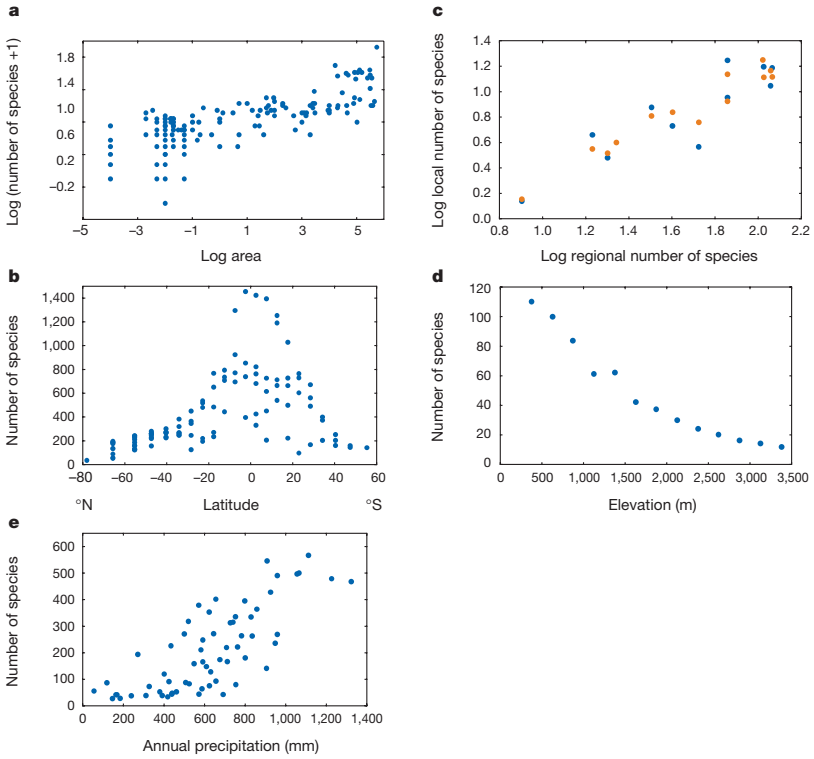


FIGURE 1.2 – Patrons de biodiversité.[a] (a) Le nombre d'espèces de vers de terre augmente en fonction de la surface échantillonnée, de 100 m^2 à plus de 500 000 km^2 selon la relation d'Arrhenius (voir page 10). (b) Nombre d'espèces d'oiseaux en fonction de la latitude. (c) Relation entre la richesse 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

[a] K. J. GASTON (2000). « Global patterns in biodiversity ». In : *Nature* 405.6783, p. 220–227, figure 1.

De façon générale,[24] 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), voir Figure 1.2. 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.

[24] K. J. GASTON (2000). « Global patterns in biodiversity ». In : *Nature* 405.6783, p. 220–227.

La diversité β mesure à quel points les systèmes locaux sont différents. Cette définition assez vague fait toujours l'objet de débats.[25]

[25] C. E. MORENO et P. RODRÍGUEZ (2010). « A consistent terminology for quantifying species diversity ? » In : *Oecologia* 163.2, p. 279–82.

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.

1.2.2 Décomposition

Whittaker[26] 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.

[26] R. H. WHITTAKER (1977). « Evolution of species diversity in land communities ». In : *Evolutionary Biology* 10, p. 1–67.

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 additive (ou multiplicative) décrivant la diversité totale, décomposable en la somme (ou le produit) convenablement pondérée 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 SAD et SAR

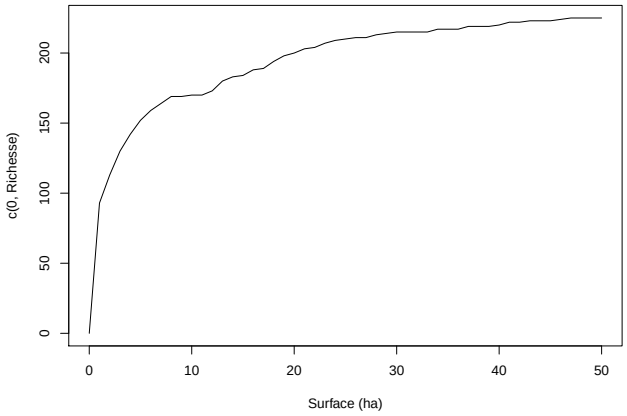


FIGURE 1.3 – Courbe d'accumulation des espèces d'arbres du dispositif de Barro Colorado Island. Le nombre d'espèces est cumulé dans l'ordre des carrés d'un hectare du dispositif.

La courbe aire-espèces (SAR : *Species Area Relationship*) représente le nombre d'espèces observées en fonction de la surface échantillonnée (Figure 2.1, page 10). Il existe plusieurs façons de prendre en compte cette relation,[27] classables en deux grandes familles[28] :

- 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 la communauté ;
- Si les points représentent un effort d'échantillonnage différent dans même communauté, on parle de 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 [29] montre l'accumulation des espèces pour les données de Barro Colorado Island,[30] 50 ha de forêt, inventaire

[27] S. M. SCHEINER (2003). « Six types of species-area curves ». In : *Global Ecology and Biogeography* 12.6, p. 441–447.

[28] J. DENGLER (2009). « Which function describes the species-area relationship best ? A review and empirical evaluation ». In : *Journal of Biogeography* 36.4, p. 728–744.

[30] S. P. HUBBELL et al. (2005). *Barro Colorado Forest Census Plot Data.*

[29] Le code R nécessaire pour réaliser les figures est :

des arbres de plus de 10 cm de diamètre à hauteur de poitrine. Une SAC peut être tracée en fonction de la surface, du nombre d'individus ou du nombres de placettes d'échantillonnage, selon les besoins.

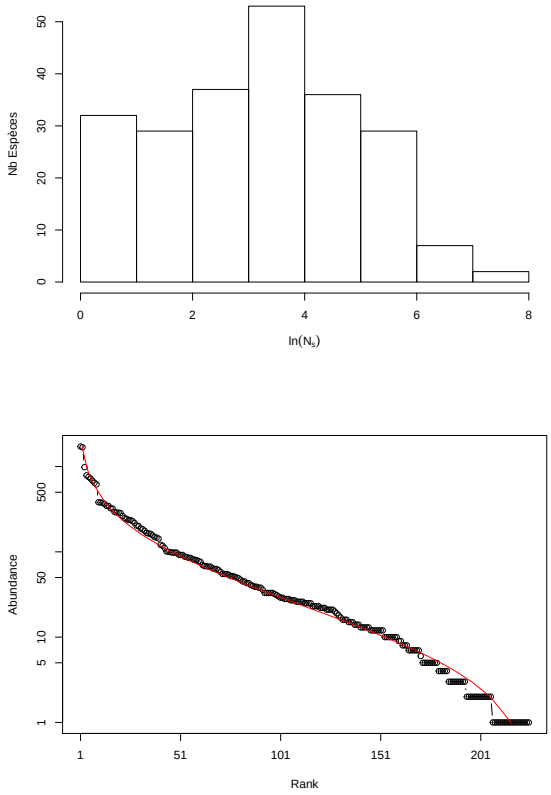


FIGURE 1.4 – Histogramme des fréquences (diagramme de Preston, en haut) des arbres du dispositif de Barro Colorado Island. En abscisse, le nombre d'arbres de chaque espèce (en logarithme), en ordonnée, le nombre d'espèces). Diagramme rang-abondance (diagramme de Whittaker, en bas) correspondant. Les points sont les données, la courbe est l'ajustement d'une distribution log-normale.

La distribution de l'abondance des espèces (SAD : *Species Abundance Distribution*) est la loi 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 1.4 [31]). Elle peut être représentée sous la

```r
library(vegan)
data(BCI)
Cumul <- apply(BCI, 2, cumsum)
Richesse <- apply(Cumul, 1, function(x) sum(x > 0))
plot(y = c(0, Richesse), x = 0:50, type = "l", xlab = "Surface (ha)")
```

[31] Le code R nécessaire pour réaliser les figures est :

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

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

forme d'un histogramme des fréquences (diagramme de Preston,[32] en haut) ou bien d'un diagramme rang-abondance (RAC : *Rank Abundance Curve* ou diagramme de Whittaker,[33] en bas). Le RAC est souvent utilisé pour reconnaître des distributions connues. Izsák et Pavoine[34] ont étudié les propriétés des RAC pour les principales SAD.

Les SAD ne sont pas traitées en détail ici : on se reportera à Magurran,[35] McGill *et al.*[36] et Izsák et Pavoine. Les SAD nécessaires à la compréhension de la suite sont :

- La distribution géométrique ;[37]
- La distribution en log-séries de Fisher ;[38]
- La distribution log-normale ;[39]
- Le modèle Broken Stick.[40]

[32] F. W. Preston (1948). « The Commonness, And Rarity, of Species ». In : *Ecology* 29.3, p. 254–283.

[33] Whittaker (1965). « Dominance and diversity in land plant communities », cf. note 16, p. 2.

[34] J. Izsák et S. Pavoine (2012). « Links between the species abundance distribution and the shape of the corresponding rank abundance curve ». In : *Ecological Indicators* 14.1, p. 1–6.

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

[36] B. J. McGill et al. (2007). « Species abundance distributions : moving beyond single prediction theories to integration within an ecological framework ». In : *Ecology Letters* 10.10, p. 995–1015.

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

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

[39] Preston (1948). Cf. note 32.

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

Chapitre 2

Mesures neutres de la diversité α ou γ

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]

Les mesures présentées ici sont les plus utilisées : dans un premier temps, celles fondées sur les relations aire-espèces, dont l'indice α de Fisher, puis les trois mesures d'entropie (richesse, Shannon, Simpson). Ces dernières sont sujettes à des biais d'estimation[4] : la correction de ces biais est traitée en détail.

L'entropie HCDT permet d'unifier ces mesures et les nombres de Hill de leur donner un sens intuitif.

2.1 Mesures fondées sur les relations aire-espèces

En forêt tempérée, toutes les espèces d'un habitat sont rencontrées sur une surface inférieure à l'hectare. L'*aire minimale* est la surface nécessaire à échantillonner pour caractériser une *association végétale*, dans l'approche de la phytosociologie fondée par Braun-Blanquet[5] : elle ne contient pas toutes les espèces mais un sous-ensemble caractéristique. Elle est de l'ordre de $200\,\mathrm{m}^2$ en Europe.[6]

En forêt tropicale, il existe de nombreuses espèces rares, au sens où la probabilité de ne pas les rencontrer à l'occasion d'un échantillonnage d'une certaine taille n'est pas négligeable. L'expérience montre que l'augmentation de l'aire d'échantillonnage

[1] R. K. Peet (1974). « The measurement of species diversity ». In : *Annual review of ecology and systematics* 5, p. 285–307.

[2] V. Devictor et al. (2010). « 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, p. 1030–40 ; J. C. Stegen et A. H. Hurlbert (2011). « Inferring ecological processes from taxonomic, phylogenetic and functional trait β-diversity. » In : *PloS one* 6.6, e20906.

[3] K. R. Clarke et R. M. Warwick (2001). « A further biodiversity index applicable to species lists : variation in taxonomic distinctness ». In : *Marine Ecology-Progress Series* 216, p. 265–278 ; C. Ricotta et G. C. Avena (2003b). « An information-theoretical measure of taxonomic diversity ». In : *Acta biotheoretica* 25.51, p. 35–41.

[4] D. Mouillot et A. Leprêtre (1999). « A comparison of species diversity estimators ». In : *Researches on Population Ecology* 41.2, p. 203–215.

[5] J. Braun-Blanquet (1928). *Pflanzensoziologie : Grundzüge der Vegetationskunde*. Berlin : Springer.

[6] M. Chytrý et Z. Otýpková (2003). « Plot sizes used for phytosociological sampling of European vegetation ». In : *Journal of Vegetation Science* 14.4, p. 563–570.

n'est pas la solution adaptée, puisqu'on rencontre toujours de nouvelles espèces,[7] même en échantillonnant des surfaces considérables (50 ha pour la Figure 2.8, page 19, au-delà du million de km^2 pour la Figure 2.11, page 25). Dans le cas général, la SAR ne tend pas vers une asymptote, et nombre d'espèces augmente tant que la surface échantillonnée croît.[8] Il ne s'agit donc pas d'estimer un nombre d'espèces asymptotique pour une communauté, mais un nombre dépendant de la surface d'échantillonnage.

[7] N. J. Gotelli et R. K. Colwell (2001). « Quantifying biodiversity : procedures and pitfalls in the measurement and comparison of species richness ». In : *Ecology Letters* 4.4, p. 379–391.

[8] M. Williamson et al. (2001). « The species–area relationship does not have an asymptote ! » In : *Journal of Biogeography* 28.7, p. 827–830.

2.1.1 La relation d'Arrhenius

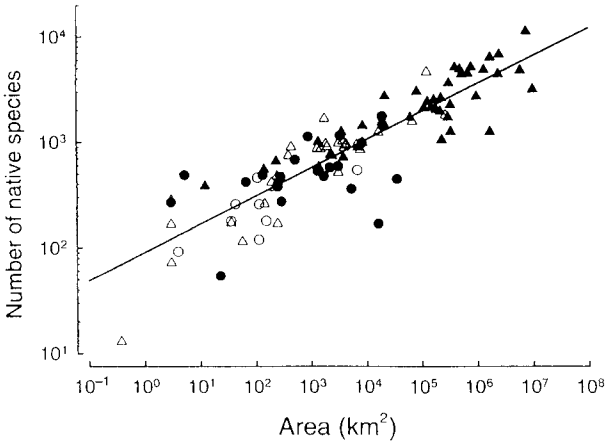


Figure 2.1 – Logarithme du nombre d'espèces en fonction du logarithme de la surface échantillonnée.[a] Les données sont des relevés d'espèces natives sur 104 sites de taille variable à travers le monde. Les symboles correspondent à divers types d'habitats (îles ou continent, réserves ou non) sans intérêt ici. La valeur de z est 0,27. Données de Lonsdale.[b]

[a] M. Williamson et al. (2001). « The species–area relationship does not have an asymptote ! » In : *Journal of Biogeography* 28.7, p. 827–830, fig. 1.

[b] W. M. Lonsdale (1999). « Global Patterns of Plant Invasions and the Concept of Invasibility ». In : *Ecology* 80.5, p. 1522–1536.

Arrhenius[9] a établi le modèle de base, dit *loi de puissance* :

$$\hat{S}^A = cA^z \tag{2.1}$$

A est la surface, S^A est le nombre d'espèces dans A, c et z des paramètres.

[9] O. Arrhenius (1921). « Species and Area ». In : *Journal of Ecology* 9.1, p. 95–99.

Preston[10] a montré que si la distribution des espèces est log-normale (sur un ensemble d'îles), la SAR (le nombre d'espèces en fonction de la surface de chaque île) suit une loi de puissance, et a établi que la puissance z valait environ 0,26 (à partir de données empiriques, pas par nécessité théorique). On notera que le nombre d'espèces est clairement fini dans ce cadre, mais que la SAR n'a pas d'asymptote : les deux notions ne sont pas équivalentes. La loi de puissance peut donc être bornée. Le glissement d'un modèle de biogéographie des îles à un cadre plus général est faisable simplement, à la manière de Hubbell[11] pour la théorie neutre : chaque île peut être considérée comme une communauté locale issue d'une méta-communauté régionale.

[10] F. W. Preston (1962). « The Canonical Distribution of Commonness and Rarity : Part I ». In : *Ecology* 43.2, p. 185–215.

[11] S. P. Hubbell (2001). *The unified neutral theory of biodiversity and biogeography*. Princeton University Press.

La distribution log-normale de Preston, dite « canonique » assumait une relation particulière entre le nombre d'espèces et le nombre d'individus, pour laquelle May[12] a montré que la valeur

[12] R. M. May (1975). « Patterns of species abundance and diversity ». In : *Ecology and Evolution of Communities*. Sous la dir. de M. L. Cody et J. M. Diamond. Harvard University Press, p. 81–120.

[13] May (1975). Cf. note 12 ; R. M. May et M. P. H. Stumpf (2000). « Species-Area Relations in Tropical Forests ». In : *Science* 290.5499, p. 2084–2086.

exacte de z est 0,25. Les résultats empiriques ultérieurs,[13] pour des lois normales quelconques, non canoniques, ont fourni des valeurs généralement comprises entre 0,2 et 0,3. Une large méta-analyse[14] à partir de 794 SAR d'origine très différentes fournit une moyenne de 0,27, comme la méta-analyse de la Figure 2.1,[15] mais avec une certaine variabilité que les auteurs relient à l'échelle spatiale, au type d'organismes et aux habitats échantillonnés, mais aussi à la technique d'échantillonnage.

[14] S. Drakare et al. (2006). « The imprint of the geographical, evolutionary and ecological context on species–area relationships ». In : *Ecology Letters* 9.2, p. 215–227.

[15] Williamson et al. (2001). « The species–area relationship does not have an asymptote ! », cf. note *a*, p. 10.

Le paramètre le plus variable de la loi de puissance reste donc c, un facteur d'échelle dépendant de la richesse de l'habitat.

Représentée en logarithmes, la relation est une droite (Figure 2.1) :

$$\ln \hat{S}^A = \ln c + z \ln A \tag{2.2}$$

2.1.2 Le modèle de Gleason

Gleason[16] réfute le travail d'Arrhenius[17] à partir de l'argument selon lequel le nombre d'espèces doit se stabiliser quand la surface augmente, au risque de prévoir une richesse beaucoup trop élevée pour de grandes surfaces. Les travaux ultérieurs lui donneront tort.[18]

[16] H. A. Gleason (1922). « On the Relation Between Species and Area ». In : *Ecology* 3.2, p. 158–162.

[17] Arrhenius (1921). Cf. note 9.

[18] Williamson et al. (2001). « The species–area relationship does not have an asymptote ! », cf. note *a*, p. 10.

Gleason propose une relation dans laquelle le nombre d'espèces (et non son logarithme) croît avec le logarithme de la surface :

$$\hat{S}^A = \ln c + z \ln A \tag{2.3}$$

Williams[19] reprend ce modèle et affirme sa validité à partir d'arguments sur la variabilité de l'habitat quand la surface d'échantillonnage augmente (par opposition à la relation d'Arrhenius applicable dans un habitat homogène, et due à la distribution log-normale des espèces selon Preston). Connor et McCoy[20] discutent de la façon de tester les deux hypothèses mais ne disposent pas des données pour le faire.

[19] C. B. Williams (1964). *Patterns in the balance of nature*. Academic Press.

[20] E. F. Connor et E. D. McCoy (1979). « The Statistics and Biology of the Species-Area Relationship ». In : *The American Naturalist* 113.6, p. 791–833.

2.1.3 Courbes d'accumulation

L'approche utilisée dans la suite est très différente de la précédente. On suppose que la communauté locale est de taille fixe et contient donc un nombre fixe d'espèces, alors que l'approche par la SAR ne fait pas d'hypothèse sur les limites de la communauté et fournit un nombre d'espèces dépendant de la surface.

Gotelli et Colwell[21] proposent une revue des problèmes les plus communément rencontrés dans l'estimation et la comparaison de la richesse spécifique. Ils suggèrent notamment de calculer la courbe de raréfaction (Figure 2.2) plutôt que d'utiliser la courbe d'accumulation obtenue en traçant simplement le nombre d'espèces découvert en fonction de l'effort d'échantillonnage. La courbe de

[21] Gotelli et Colwell (2001). « Quantifying biodiversity : procedures and pitfalls in the measurement and comparison of species richness », cf. note *a*, p. 12.

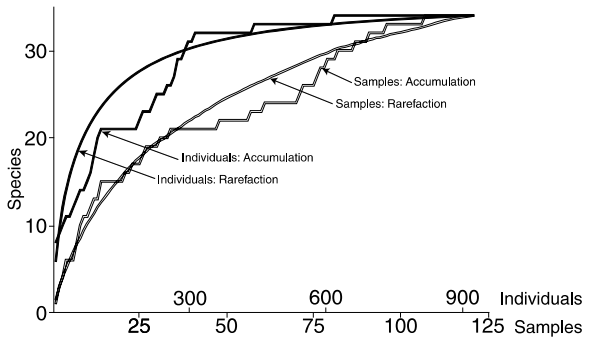
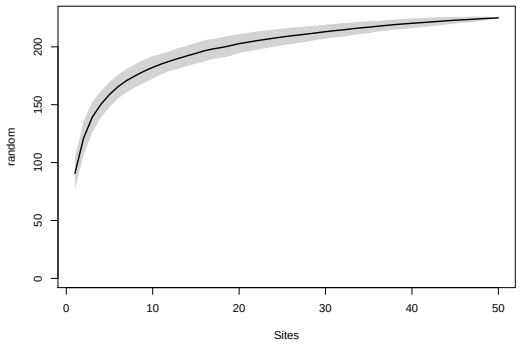
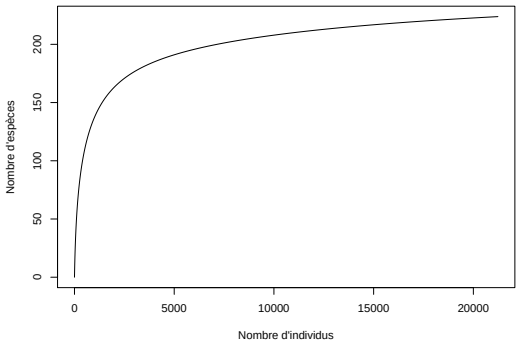


FIGURE 2.2 – Courbes d'accumulation et de raréfaction, obtenues par inventaires individuels ou groupés ("*samples*"). Les courbes de raréfaction permettent de lisser les courbes d'accumulation. Les courbes obtenues par inventaire groupé sont en général en dessous des courbes individuelles à cause de l'agrégation spatiale[a]

[a] N. J. GOTELLI et R. K. COLWELL (2001). « Quantifying biodiversity : procedures and pitfalls in the measurement and comparison of species richness ». In : *Ecology Letters* 4.4, p. 379–391, Fig. 1.



(a) Calcul par `specaccum` : accumulation des 50 carrés avec permutations.



(b) Somme des entropies de Simpson généralisées.

FIGURE 2.3 – Courbe d'accumulation des espèces d'arbres de Barro Colorado Island

raréfaction est obtenue en sous-échantillonnant dans l'inventaire complet des effectifs de toutes tailles et en calculant le nombre moyen d'espèces trouvé pour chacun, ce qui permet de lisser la courbe. La formulation exacte de la courbe de raréfaction et de sa variance est connue.[22]

La fonction `specaccum` de la librairie *vegan*[23] permet de tracer les SAC (Figure 2.3a [24]) selon différentes méthodes, avec un intervalle de confiance obtenu par permutation de l'ordre d'ajout des données ou à partir des formules de raréfaction.

Pour une surface donnée, le nombre d'individus est une variable aléatoire. Dans ce cas, la SAC (dépendant de la surface) est différente de la courbe de raréfaction (fonction du nombre d'individus) de Gotelli et Colwell, que Coleman[25] nomme « courbe du collecteur ».

[22] K. I. UGLAND et al. (2003). « The species-accumulation curve and estimation of species richness ». In : *Journal of Animal Ecology* 72, p. 888–897.

[23] J. OKSANEN et al. (2012). *vegan : Community Ecology Package*.

[25] B. D. COLEMAN (1981). « On random placement and species-area relations ». In : *Mathematical Biosciences* 54.3-4, p. 191–215.

[24] Le code R nécessaire pour réaliser la figure est :

```
SAC <- specaccum(BCI, "random")
plot(SAC, ci.type = "poly", lwd = 2, ci.lty = 0, ci.col = "lightgray")
```

La SAC théorique en fonction de la surface contient une incertitude concernant le nombre d'individus. Si la densité λ_s (nombre d'individus par unité de surface) des espèces peut être connue (estimée en comptant les individus) ou si les probabilités p_s sont issues d'un modèle (log-séries, log-normal...), le nombre d'individus observés de l'espèce s suit une loi de Poisson d'espérance $A\hat{\lambda}_s$ ou, de façon équivalente,[26] np_s (dans ce deuxième cas, n est l'estimateur de N, la variable aléatoire correspondant à la taille de l'échantillon).

La SAC théorique en fonction du nombre d'individus ne contient pas cet aléa. Le nombre d'individus de l'espèce s suit alors une loi binomiale d'espérance np_s et la distribution d'un échantillon de taille n suit la loi multinomiale correspondante. Les deux approches se réconcilient en fixant arbitrairement le nombre d'individus échantillonnés dans la courbe aire-espèce : conditionnellement au nombre total d'individus, la distribution des abondances des espèces issues de lois de Poisson de paramètres λ_s suit une loi multinomiale[27] dont les probabilités sont proportionnelles à λ_s.

Une difficulté supplémentaire consiste à prendre en compte les espèces non échantillonnées en appliquant des techniques statistiques pour réduire le biais qu'elles induisent.[28] Dans ce but, la SAC théorique peut être construite[29] comme la somme des entropies de Simpson généralisées[30] (présentées page 49).

La courbe obtenue (Figure 2.3b [31]) est la même que par raréfaction, sans les problèmes d'estimation. Son intervalle de confiance n'est pas représenté : l'estimation est sans biais et a une variance très petite ; d'autre part, la variabilité de l'échantillonnage prise en compte dans la figure 2.3a n'est pas prise en compte ici. La fonction `GenSimp.z` est fournie par le package *EntropyEstimation*.[32]

2.1.4 Indice α de Fisher

Fisher[33] a relié le nombre d'espèces $\hat{S}^n$ au nombre d'individus n échantillonnés à partir du modèle suivant :

- Étant donné la taille de l'échantillon n et la probabilité p_s d'échantillonner l'espèce s, l'espérance du nombre d'individus échantillonnés dans l'espèce s est np_s ;

[26] E. C. Pielou (1969). *An Introduction to Mathematical Ecology*. New York : Wiley, pages 204-205.

[27] G. D. Steel (1953). « Relation Between Poisson and Multinomial Distributions ». In : *Biometrics Unit Technical Reports*. Cornell University. Biometrics Unit. Chap. BU-39-M, p. 1–2.

[28] R. K. Colwell et al. (2012). « Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages ». In : *Journal of Plant Ecology* 5.1, p. 3–21.

[29] A. Chao et al. (2013). « Entropy and the species accumulation curve : a novel entropy estimator via discovery rates of new species ». In : *Methods in Ecology and Evolution* 4.11, p. 1091–1100.

[30] Z. Zhang et J. Zhou (2010). « Re-parameterization of multinomial distributions and diversity indices ». In : *Journal of Statistical Planning and Inference* 140.7, p. 1731–1738.

[31] Code R :

```
library(EntropyEstimation)
Ns <- colSums(BCI)
gsValues <- sapply(1:(sum(Ns - 1)), function(r) GenSimp.z(Ns,
    r))
SAC <- cumsum(gsValues)
plot(0:sum(Ns - 1), c(0, SAC), type = "l", xlab = "Nombre d'individus",
    ylab = "Nombre d'espèces")
```

[32] L. Cao et M. Grabchak (2014). *EntropyEstimation : Estimation of Entropy and Related Quantities*.

[33] Fisher et al. (1943). « The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population », cf. note 38, p. 8.

- Le nombre d'individus réellement observé fluctue aléatoirement. Il suit une loi de Poisson d'espérance np_s, conséquence directe des hypothèses : surface fixée et indépendance entre les échantillons ;
- Les probabilités p_s suivent toutes la même loi. La densité de probabilité de np_s est notée $f(\lambda)$, où $\lambda = np_s$ permet d'alléger l'écriture.

La probabilité π_{n_s} d'observer n_s individus d'une espèce donnée est donc la moyenne sur toutes les valeurs possibles de λ de la loi de Poisson :

$$\pi_{n_s} = \int_0^{+\infty} \frac{\lambda^{n_s} e^{-\lambda}}{n_s!} f(\lambda)\, d\lambda \tag{2.4}$$

La distribution de λ retenue par Fisher est une loi gamma, ce qui implique que π_{n_s} suit une loi binomiale négative de paramètre k.

On notera que la distribution de π_{n_s} n'est pas celle de p_s. La courbe représentant n_s en fonction de son rang n'est pas un vrai diagramme rang-abondance, mais un « pseudo-RAC ».[34]

[34] Izsák et Pavoine (2012). « Links between the species abundance distribution and the shape of the corresponding rank abundance curve », cf. note 34, p. 8.

En appliquant son modèle à des données, Fisher remarque que $k \to 0$, ce qui correspond à une distribution comprenant de nombreuses espèces et un pseudo-RAC décroissant rapidement, et retient cette propriété pour la suite de son raisonnement. Dans ce cas, et après calcul, le nombre d'espèces échantillonnées pour n individus est calculable (pour n suffisamment grand) et ne dépend que d'un paramètre α estimé à partir des données.

Le modèle est donc applicable pour un grand nombre d'observations et un grand nombre d'espèces, dont les individus sont distribués indépendamment. Alors :

$$\mathbb{E}\left(S^n\right) = \alpha \ln\left(1 + \frac{n}{\alpha}\right) \tag{2.5}$$

α est un indicateur de la biodiversité qui peut être interprété comme le nombre d'espèces nouvelles découvertes quand le nombre d'individus échantillonnés est multiplié par e : c'est en effet la pente de la courbe de $\hat{S}^n$ en fonction de $\ln n$, qui se stabilise à partir d'une valeur de n suffisante.

α est aussi égal à θ, le nombre fondamental de la biodiversité de Hubbell quand le nombre d'espèces est grand[35] : la SAD de la méta-communauté (pas celle de la communauté locale si la dispersion est limitée) du modèle neutre est en log-séries.

[35] D. Alonso et A. J. McKane (2004). « Sampling Hubbell's neutral theory of biodiversity ». In : *Ecology Letters* 7.10, p. 901–910.

Dans R, les fonctions `fisher.alpha` de la librairie *vegan* et `fishers.alpha` de la librairie *untb*[36] effectuent une résolution numérique de l'équation 2.5 pour estimer α. La fonction `optimal.theta` dans *untb* évalue la valeur de θ par la méthode du maximum de vraisemblance,[37] en utilisant toute l'information de la distribution. Le résultat est très différent parce que l'approximation de l'égalité de α et θ ne tient pas dans ce cas :

[36] R. K. S. Hankin (2007). « Introducing untb, an R Package For Simulating Ecological Drift Under the Unified Neutral Theory of Biodiversity ». In : *Journal of Statistical Software* 22.12, p. 1–15.

[37] Hubbell (2001). *The unified neutral theory of biodiversity and biogeography*, cf. note 11, p. 10, page 122.

```
# Ns est un vecteur contenant les effectifs de chaque espèce
Ns <- colSums(BCI)
fisher.alpha(Ns)

## [1] 35.05475

library(untb)
fishers.alpha(sum(Ns), length(Ns))

## [1] 35.05477

optimal.theta(Ns)

## [1] 80.95173
```

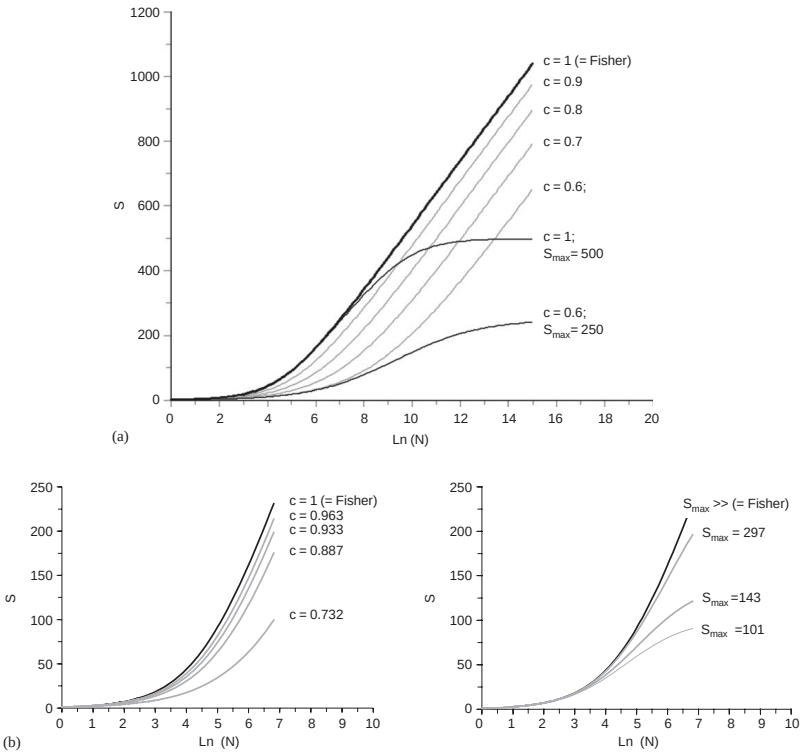


FIGURE 2.4 – Courbes[a] nombre d'individus-espèces-simulées pour illustrer l'importance des paramètres c (concentration spatiale) et $S_{\max}$ (nombre total d'espèces, S dans nos notations) dans le modèle de Fisher. α est fixé à 100 pour toutes les courbes. (a) et (b) : la concentration spatiale ($c < 1$) change drastiquement le nombre d'espèces attendu. En pratique, seuls S et N ($s^n_{\neq 0}$ et n dans nos notations) sont observés. Si c est supposé égal à 1 par erreur, α est sous-estimé. (a) et (c) : le nombre d'espèces n'est évidemment pas infini, ce qui change fortement la forme de la courbe. Si le nombre d'observations est grand, α sera aussi très sous-estimé.

[a] R. P. O. SCHULTE et al. (2005). « A new family of Fisher-curves estimates Fisher's alpha more accurately ». In : *Journal of Theoretical Biology* 232.3, p. 305–313.

Le modèle de Fisher (2.5) peut être écrit sous la forme $\hat{S}^n = \alpha \ln(\alpha + n) - \alpha \ln(\alpha)$. Pour n assez grand et sous l'hypothèse de saturation de Hubbell,[38] A est proportionnel à $\alpha + n$ et le modèle de Fisher est équivalent au modèle de Gleason.[39]

Bordenave[40] propose un indice de richesse spécifique (IRS, ou SDI : *Species Diversity Index*) égal à la pente de la relation de Gleason. Dans le cadre de l'approximation précédente, il est égal à α.

Schulte *et al.*[41] ont étendu le modèle de Fisher pour prendre en compte un nombre d'espèces fini S et une distribution non indépendante des individus, prise en compte par un paramètre d'agrégation spatiale c valable pour toutes les espèces.

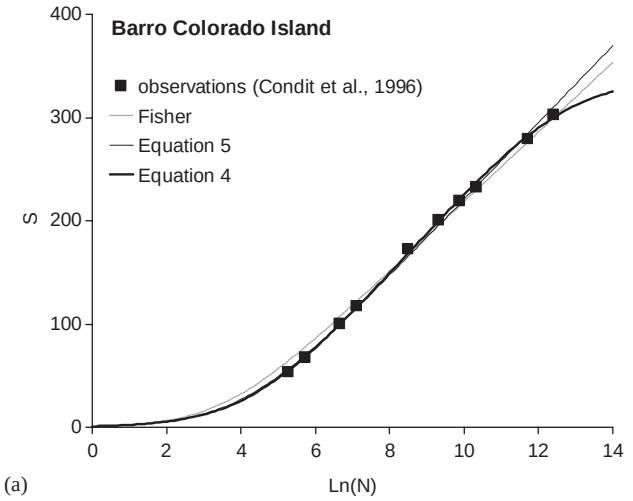
$$\hat{S}^n = \frac{\alpha}{c} \ln\left(\frac{1 + n^c\, q}{1 + n^c\, q\, e^{-cS/\alpha}}\right) \tag{2.6}$$

[38] HUBBELL (2001). *The unified neutral theory of biodiversity and biogeography*, cf. note 11, p. 10.

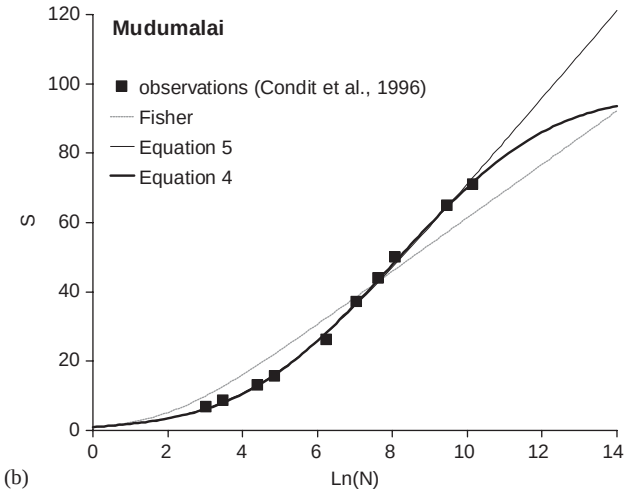
[39] S. ENGEN (1977). « Exponential and Logarithmic Species-Area Curves ». In : *The American Naturalist* 111.979, p. 591–594.

[40] B. G. BORDENAVE et J. J. DE GRANVILLE (1998). « Les mesures de la biodiversité : un outil de conservation en forêt guyanaise ». In : *JATBA, Revue d'Ethnobiologie* 40.1-2, p. 433–446 ; B. G. BORDENAVE et al. (2011). « Quantitative botanical diversity descriptors to set conservation priorities in Bakhuis Mountains rainforest, Suriname ». In : *Botanical Journal of the Linnean Society* 167.1, p. 94–130.

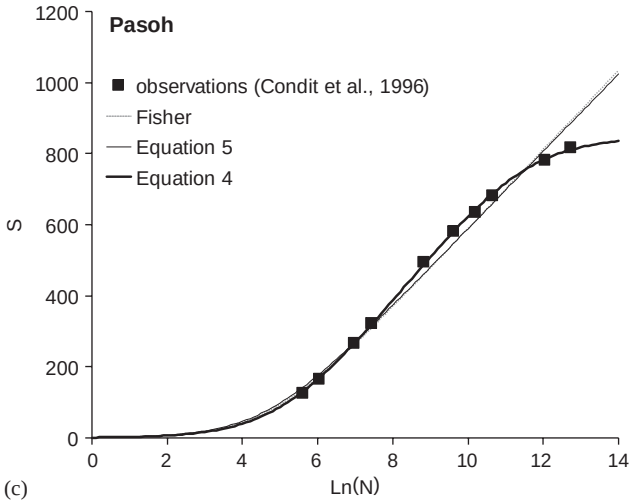
[41] R. P. O. SCHULTE et al. (2005). « A new family of Fisher-curves estimates Fisher's alpha more accurately ». In : *Journal of Theoretical Biology* 232.3, p. 305–313.



(a)



(b)



(c)

FIGURE 2.5 – Comparaison[a] entre le modèle de Fisher et le modèle étendu dans trois parcelles forestières connues. Les modèles sont ajustés aux données observées par Condit *et al.* Les courbes légendées *Equation 4* correspondent au modèle de Fisher étendu, équation (2.6), alors que l'*Equation 5* est le modèle intermédiaire, supposant le nombre total d'espèces infini.

[a] R. P. O. SCHULTE et al. (2005). « A new family of Fisher-curves estimates Fisher's alpha more accurately ». In : *Journal of Theoretical Biology* 232.3, p. 305–313.

où

$$q = \frac{e^{c/\alpha} - 1}{1 - e^{c(1-S)/\alpha}} \tag{2.7}$$

Si $S = \infty$ et $c = 1$, l'équation (2.6) se simplifie pour retrouver l'équation (2.5).

Leurs résultats (Figure 2.4) montrent que la non prise en compte de ces deux paramètres aboutit à une sous-estimation systématique de α, différente selon les sites, ce qui invalide les comparaisons inter-sites.

Appliqué à des données réelles (Figure 2.5), le modèle de Fisher étendu s'ajuste forcément mieux (il possède trois paramètres au lieu d'un). L'approximation du nombre d'espèces infini paraît être la plus pénalisante parce qu'elle exclut l'inflexion de la courbe pour les grandes valeurs de N, même si ces valeurs sont rarement atteinte dans les faits ($e^{10} > 20000$).

L'indice de Fisher repose lourdement sur l'hypothèse que la distribution réelle des espèces est conforme au modèle. Jost[42] montre par un exemple que des interprétations absurdes de l'indice peuvent être faites si l'hypothèse n'est pas respectée.

[42] L. Jost (2007). « Partitioning diversity into independent alpha and beta components ». In : *Ecology* 88.10, p. 2427–2439.

Kempton[43] a défini l'indice Q comme la pente de la courbe de $\hat{S}^n$ en fonction de $\log n$ mesurée entre le premier et le troisième quartile du nombre d'espèces observées, quelle que soit la distribution des espèces. $Q = \alpha$ si la distribution est en log-séries, et $Q = {}^{cS}/_{\sigma}$ si elle est log-normale (c est une constante dépendant de la proportion des espèces échantillonnée, S le nombre d'espèces total et σ l'écart-type de la distribution du logarithme des probabilités).

[43] R. A. Kempton et L. R. Taylor (1976). « Models and statistics for species diversity ». In : *Nature* 262.5571, p. 818–820 ; R. A. Kempton et R. W. M. Wedderburn (1978). « A comparison of three measures of species diversity ». In : *Biometrics* 34, p. 25–37.

La simulation suivante tire 10 000 arbres dans le jeu de données BCI indépendamment et selon les fréquences des espèces. L'indice Q est calculé :

```
Inventaire <- rmultinom(10000, 1, Ns/sum(Ns))
# Cumul de l'inventaire
Cumul <- apply(Inventaire, 1, cumsum)
# Nombre d'espèces cumulées
nEspeces <- apply(Cumul, 1, function(x) length(x[x > 0]))
# Nombre total d'espèces
S <- max(nEspeces)
# Premier et troisièmes quartiles
Q13 <- which(nEspeces >= max(nEspeces)/4 & nEspeces <= max(nEspeces) *
    3/4)
nq1 <- min(Q13)
nq3 <- max(Q13)
Sq1 <- nEspeces[nq1]
Sq3 <- nEspeces[nq3]
# Indice Q
(Q <- (Sq3 - Sq1)/log(nq3/nq1))

## [1] 41.47787
```

La SAC correspondante est tracée en Figure 2.6 [44].

[44] Code R pour réaliser la figure :

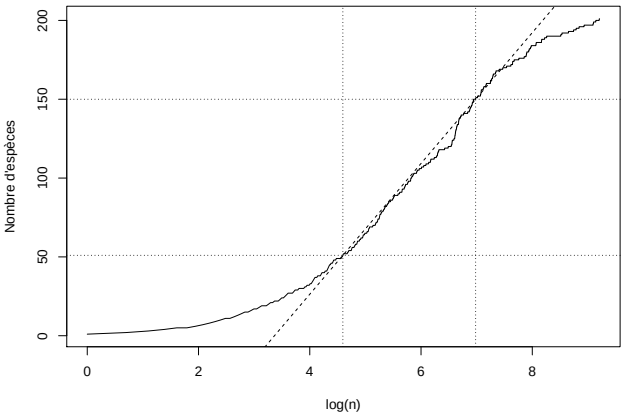


FIGURE 2.6 – SAC d'un inventaire simulé de 10 000 arbres de BCI. Le nombre d'espèces cumulé est tracé en fonction du logarithme du nombre d'individus échantillonnés. Q est la pente de la droite passant par les points de la courbe correspondant aux premier et troisième quartiles du nombre d'espèces.

2.1.5 Le modèle de Plotkin

Plotkin *et al.*[45] ont eu accès à de grands échantillons (5 placettes de 50 hectares de forêt dans l'ensemble du monde tropical, Figure 2.7, toutes déterminées correctement sur le plan botanique et dans lesquels chaque arbre est positionné). Ils montrent que le modèle d'Arrhenius comme les SAD classiques rendent mal compte de la réalité (Figure 2.8a) quand on les compare aux courbes d'accumulation à l'échelle locale.

Harte *et al.*[46] ont montré que la loi de puissance était équivalente au modèle d'autosimilarité, dérivé de la théorie des fractales. On considère un rectangle A_i dont le rapport longueur sur largeur vaut $\sqrt{2}$. On obtient le rectangle A_{i+1} en plaçant deux rectangles A_i côte à côte (et donc le rectangle A_{i-1} en coupant A_i en deux). $\hat{S}^{A_i}$ est le nombre d'espèces rencontrées dans A_i. On définit enfin $a_i = \hat{S}^{A_{i-1}}/\hat{S}^{A_i}$ le facteur de diminution du nombre d'espèces (dit *paramètre de persistance*) lié à la division par 2 de la taille du rectangle.

S'il s'agit d'une constante, a, indépendante de la taille du rectangle, l'habitat est dit auto-similaire : l'augmentation du nombre d'espèces ne dépend pas de l'échelle d'observation Si cette condition est respectée, le nombre d'espèces suit la loi de puissance, et $a = 2^{-z}$. Pour la valeur classique de $z = 0,25$, $a = 0,84$. La Figure 2.8b présente la valeur de a_i en fonction de la surface échantillonnée. Elle augmente avec la surface, ce qui infirme le modèle d'autosimilarité.

[45] J. B. PLOTKIN et al. (2000a). « Predicting species diversity in tropical forests ». In : *Proceedings of the National Academy of Sciences of the United States of America* 97.20, p. 10850–10854.

[46] J. HARTE et al. (1999). « Self-similarity in the distribution and abundance of species ». In : *Science* 284.5412, p. 334–336.

```
plot(y = nEspeces, x = log(1:ncol(Inventaire)), xlab = "log(n)",
    ylab = "Nombre d'espèces", type = "l")
abline(v = log(nq1), lty = 3)
abline(v = log(nq3), lty = 3)
abline(h = Sq1, lty = 3)
abline(h = Sq3, lty = 3)
abline(b = Q, a = Sq1 - Q * log(nq1), lty = 2)
```

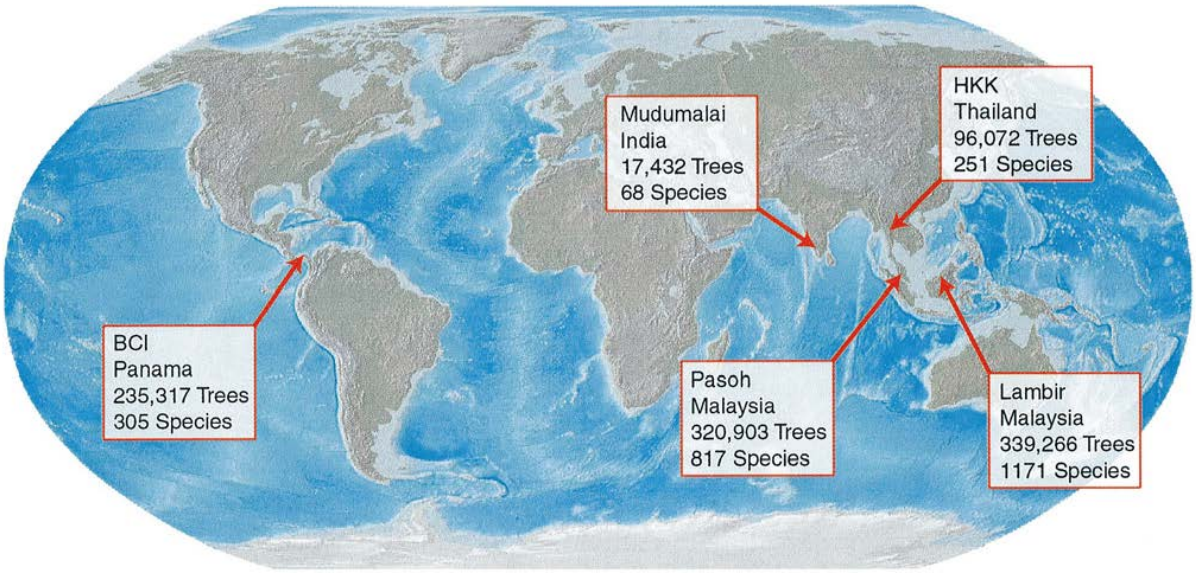


FIGURE 2.7 – Les cinq dispositifs de Plotkin *et al.* Chaque dispositif mesure 50 ha et toutes les tiges de diamètre supérieur à 1 cm sont cartographiées et déterminées.

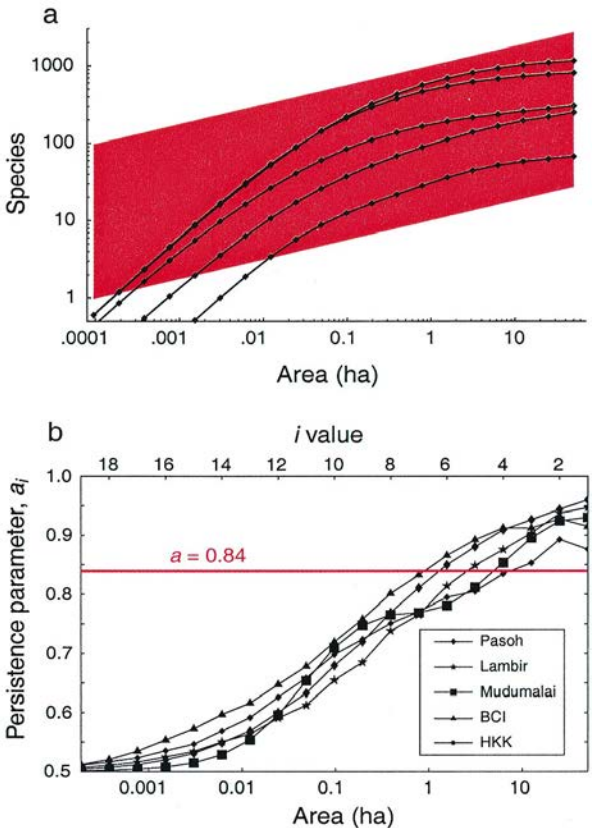


FIGURE 2.8 – Courbes aire-espèces (a) et paramètre de persistance en fonction de la surface (b) dans 5 forêts tropicales.[a] Le modèle d'Arrhenius est représentée par le trapèze sur la figure a, pour une gamme de valeurs de la constante. Elle implique que la relation aire-espèces représentée en logarithmes soit une droite, dont la position dépend de la constante, ce qui n'est visiblement pas le cas.

[a] J. B. PLOTKIN et al. (2000a). « Predicting species diversity in tropical forests ». In : *Proceedings of the National Academy of Sciences of the United States of America* 97.20, p. 10850–10854.

[47] Plotkin et al. (2000a). « Predicting species diversity in tropical forests », cf. note *a*, p. 19.

Plotkin *et al.*[47] ne supposent pas a_i constant et lui imposent simplement d'être une fonction de la surface, choisie arbitrairement pour permettre les calculs ultérieurs, et dont la forme correspond aux observations de la Figure 2.8b. Après calculs, la relation aire-espèces est $S^A \approx cA^z e^{-p(A)}$. $p(A)$ est un polynôme de degré n de A, sans constante. Les monômes sont d'autant moins importants que leur degré est élevé. L'approximation de degré 0 est la loi de puissance. Celle de degré 1 est retenue par les auteurs pour estimer le nombre d'espèces rencontrées sur la surface A :

$$\hat{S}^A = cA^z e^{-kA} \tag{2.8}$$

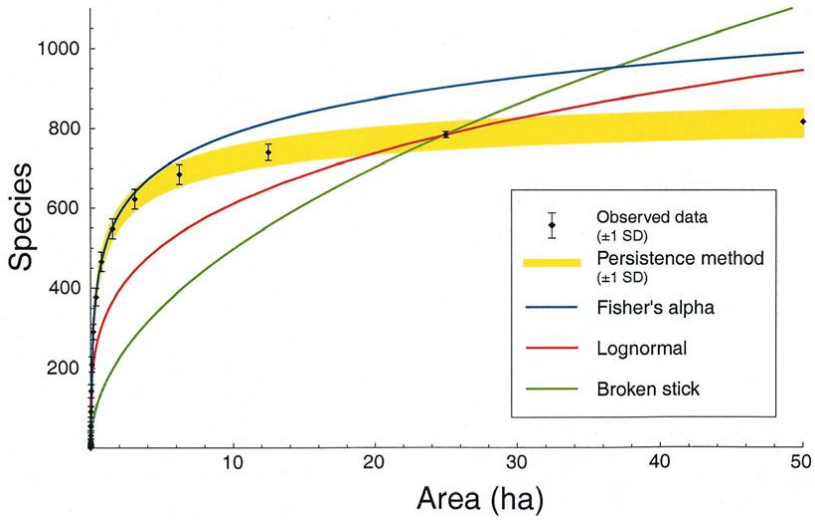


Figure 2.9 – Performance du modèle de Plotkin *et al.*[a] face à des distributions théoriques : SAD log-normales, modèles de Fisher et Broken Stick de MacArthur. La courbe représente le nombre d'espèces rencontrées à Pasoh sur des surfaces de 0 à 50 ha. Les données observées sont des tirages aléatoires d'échantillons de la surface souhaitée. L'écart-type est représenté (l'intervalle de fluctuation à 95% est de l'ordre de 2 écart-types). Les valeurs issues du modèle ont été obtenues par un paramétrage de k et z sur 25 ha, supposé valide également sur les autres forêts. Le paramétrage de c est obtenu par un échantillon d'un hectare, répété 1000 fois pour obtenir un intervalle de confiance. L'intervalle de fluctuation à 95% est ici encore de l'ordre de deux écart-types

[a] J. B. Plotkin et al. (2000a). « Predicting species diversity in tropical forests ». In : *Proceedings of the National Academy of Sciences of the United States of America* 97.20, p. 10850–10854.

Ce modèle contient un paramètre k en plus de la loi de puissance. Les auteurs ont établi que les paramètres k et z étaient très proches pour les cinq forêts ($z = 0,125$ et $k = 0,0541$ à Pasoh), et que leur estimation sur une forêt permettait de prédire avec une très bonne précision la richesse des autres, à partir d'un échantillon d'un hectare nécessaire pour obtenir le paramètre c. La Figure 2.9 compare les performances de cette méthode (*persistence method*), de la loi de puissance 0,25 (loi log-normale) et du modèle de Fisher.

2.1.6 Le modèle de placement aléatoire

En concurrence au modèle de Preston[48] prédisant une SAR en forme de loi de puissance, Coleman[49] propose le modèle de place-

[48] Preston (1962). « The Canonical Distribution of Commonness and Rarity : Part I », cf. note 10, p. 10.

[49] Coleman (1981). « On random placement and species-area relations », cf. note 25, p. 12 ; B. D. Coleman et al. (1982). « Randomness, area, and species richness ». In : *Ecology* 63.4, p. 1121–1133.

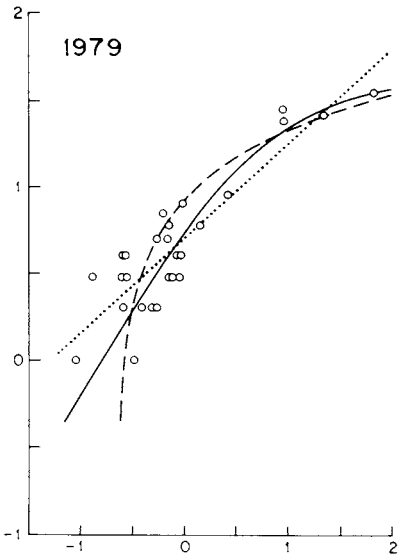


FIGURE 2.10 – Relation aire-espèce de Coleman.[a] Les axes sont en échelle logarithmique. Les points représentent des relevés de nombre d'espèces d'oiseaux (en 1979) en fonction de la taille des îles où elles se trouvaient. La ligne en pointillés courts est l'ajustement d'une loi de puissance (modèle d'Arrhenius). Les pointillés long correspondent du modèle de Gleason. La courbe continue est l'ajustement du modèle de placement aléatoire. Un test de qualité d'ajustement montre que le modèle de placement aléatoire est le plus proche des données.

[a] B. D. COLEMAN et al. (1982). « Randomness, area, and species richness ». In : *Ecology* 63.4, p. 1121–1133, fig. 3.

mant aléatoire (*random placement*).

Plusieurs îles (communautés locales) de taille relative $\{p_i\}$ appartiennent à un ensemble (méta-communautés) dans lequel la distribution des espèces $\{n_s\}$, est considéré comme connu. Chaque individu d'une espèce donnée peut se placer sur une quelconque des îles, avec une probabilité égale à p_i. La probabilité qu'un individu au moins de l'espèce s se trouve sur l'île i est $1-(1-p_i)^{n_s}$ (1 moins la probabilité que tous les individus de l'espèce se trouve sur une autre île, elle-même égale à $1-p_i$ pour tous les individus). L'espérance du nombre d'espèces représentées sur l'île i est donc :

$$\mathbb{E}\left(S^i_{\neq 0}\right) = S - \sum_s (1-p_i)^{n_s} \tag{2.9}$$

Ce modèle ne fait aucune hypothèse sur la distribution de la probabilité des espèces et n'assume que l'indépendance des individus et la proportionnalité entre la probabilité de présence et la surface. Sur des données de richesse en nombre d'espèces d'oiseaux reproducteurs dans les îles d'un lac nord-américain, le modèle s'ajuste mieux que ceux d'Arrhenius et de Gleason (figure 2.10).

La fonction `specaccum` de *vegan* permet de calculer une SAC selon ce modèle.

2.1.7 Influence de la structure spatiale

Les modèles classiques de courbes aire-espèces supposent un tirage indépendant des individus. Sur le terrain, par exemple en forêt, l'échantillonnage est continu : s'il existe une structure spatiale comme des agrégats, la probabilité que l'arbre suivant soit d'une espèce donnée dépend de ses voisins, donc de l'arbre précédent. Intuitivement, on comprend bien que le nouvel arbre de la même espèce apporte moins d'information, ce qui revient à surestimer

la taille de l'échantillon ou sous-estimer la richesse pour une taille d'échantillon fixée.

Plotkin *et al.*[50] traitent un problème un peu différent : connaissant le nombre total d'espèces dans les dispositifs de la Figure 2.7, ils tracent la courbe de raréfaction théorique qui considère simplement que la probabilité de ne pas rencontrer une espèce suit une loi binomiale, les tirages étant indépendants. Ce modèle surestime largement la diversité pour les petites surfaces : la probabilité de ne pas rencontrer une espèce agrégative est sous-estimée. Le biais diminue quand la surface d'échantillonnage augmente, parce que la taille relative des agrégats diminue.

Les effets de l'agrégation peuvent donc être opposés selon la question posée. Dans tous les cas, il suffit de considérer qu'un agrégat surestime le nombre d'arbres de son espèce qui devrait être pris en compte dans le cadre d'un modèle à tirages indépendants.

Plotkin *et al.* proposent une méthode permettant de prendre en compte la structure spatiale pour fournir des modèles de courbes aire-espèces fiables, dans un cadre particulier où le nombre total d'espèces et leur structure spatiale sont connus. La méthode nécessite une carte des individus. Le semis de point est considéré comme le résultat d'un processus de Neyman-Scott[51] : des centres d'agrégats sont tirés de façon complètement aléatoire et les individus sont répartis autour des centres selon une loi normale en deux dimensions. Les paramètres du processus sont estimés à partir du semis de points, puis la courbe aire-espèces est obtenue par simulation du processus.

Le problème est traité de façon plus approfondie par Picard *et al.*[52] qui s'appuient sur la théorie des processus ponctuels. En supposant que la distribution spatiale des espèces est homogène (c'est-à-dire stationnaire et isotrope), la forme du domaine échantillonné n'a pas d'importance : l'aire A peut être représentée par un cercle de rayon $r = \sqrt{A/\pi}$. La probabilité d'observer l'espèce s dans A est identique à la probabilité que la distance entre un individu de l'espèce s et un point arbitrairement choisi soit inférieure à r. Cette probabilité a été largement étudiée et est connue sous le nom de fonction F de Diggle,[53] ou *empty space function*. $F_s(r)$ est la fonction d'espace vide de l'espèce s, dont l'argument est la distance r. Le nombre d'espèces peut être écrit comme la somme des fonctions F_s :

$$\mathbb{E}\left(S^A\right) = \sum_s F_s\left(r\right) \tag{2.10}$$

La fonction $F_s(r)$ est connue pour les processus ponctuels courants.[54] Le processus de Poisson place les individus indépendamment les uns des autres avec la même probabilité partout. Son seul paramètre est l'intensité λ_s, qui donne l'espérance du nombre d'individus de l'espèce s par unité de surface. Si toutes les

[50] J. B. Plotkin et al. (2000b). « Species-area curves, spatial aggregation, and habitat specialization in tropical forests ». In : *Journal of Theoretical Biology* 207.1, p. 81–99.

[51] J. Neyman et E. L. Scott (1958). « Statistical Approach to Problems of Cosmology ». In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 20.1, p. 1–43.

[52] N. Picard et al. (2004). « Species-area curve and spatial pattern ». In : *Ecoscience* 11.1, p. 45–54.

[53] P. J. Diggle (1983). *Statistical analysis of spatial point patterns*. London : Academic Press, p. 1–148.

[54] N. A. Cressie (1993). *Statistics for spatial data*. New York : John Wiley & Sons, p. 1–900 ; D. Stoyan et H. Stoyan (1994). *Fractals, Random Shapes and Point Fields : Methods of Geometrical Statistics*. Chichester : John Wiley & Sons, 406 p.

espèces sont distribuées selon un processus de Poisson, la courbe aire-espèce est décrite par :

$$\mathbb{E}\left(S^A\right) = \sum_s e^{-\lambda_s A} \tag{2.11}$$

L'équation (2.11) est identique au modèle de distribution aléatoire de Coleman en remplaçant la distribution observée des effectifs par espèce par l'intensité du processus de Poisson. Les formes de F_s sont connues aussi pour des processus agrégatifs, notamment le processus de Matérn[55] ou des distributions régulières. La courbe aire-espèce peut donc être évaluée à partir d'inventaires fournissant la position des individus. Le processus ponctuel correspondant à chaque espèce peut être inféré pour fournir une forme analytique de la courbe comme dans l'équation (2.11) ou bien la fonction d'espace vide de chaque espèce peut être estimée à partir des données par la fonction `Fest` du package *spatstat* et sommée pour toutes les espèces selon l'équation (2.10).

[55] B. Matérn (1960). « Spatial variation ». In : *Meddelanden från Statens Skogsforskningsinstitut* 49.5, p. 1–144.

La relation aire-espèce contient au moins autant de paramètres que d'espèces : l'intensité λ_s si l'espèce s est distribuée spatialement selon un processus de Poisson, deux paramètres supplémentaires pour le processus de Matérn. Cette complexité peut être réduite si la distribution des fréquences des espèces (la SAD) est connue. Picard *et al.* fournissent les relations pour une SAD en log-séries et une SAD géométrique, avec une distribution spatiale suivant un processus de Poisson : elles ne contiennent que deux paramètres, un lié au nombre total d'individus, et un propre à chaque SAD. Les formules se compliquent quand la distribution spatiale n'est pas poissonnienne et ne sont pas disponibles quand la SAD (log-normale par exemple) ne prévoit pas le nombre exact d'individus par espèce.

Enfin, il est possible de prendre en compte les espèces non observées. La relation aire-espèce (2.10) s'entend pour toutes les espèces. Elle peut être réécrite sous la forme :

$$\mathbb{E}\left(S^A\right) = \sum_{s=1}^{S_{\neq 0}} F_s\left(r\right) + \sum_{s=S_{\neq 0}+1}^{S} F_s\left(r\right) \tag{2.12}$$

Le premier terme de la somme correspond aux espèces observées, le deuxième aux espèces non observées. La distribution des espèces non observées peut maintenant être estimée dans certains cas[56] et leur distribution spatiale être supposée poissonnienne puisque ce sont des espèces rares,[57] ce qui ouvre la voie à une meilleure estimation de la courbe.

[56] A. Chao et al. (2015b). « Unveiling the Species-Rank Abundance Distribution by Generalizing Good-Turing Sample Coverage Theory ». In : *Ecology* 96.5, p. 1189–1201.

[57] Picard et al. (2004). Cf. note 52.

2.1.8 Synthèse

La relation d'Arrhenius est plus robuste que celle de Gleason, selon les résultats empiriques et des considérations théoriques,[58]

[58] H. Gárcia Martín et N. Goldenfeld (2006). « On the origin and robustness of power-law species-area relationships in ecology ». In : *Proceedings of the National Academy of Sciences* 103.27, p. 10310–10315.

y compris de thermodynamique.[59]

Engen[60] généralise le modèle de Fisher en autorisant des valeurs négatives du paramètre k de la loi binomiale négative. Dans le modème de Fisher, ce paramètre est l'inverse de la variabilité de l'espérance du nombre d'individus observés par espèce, qui tend vers 0 car la variabilité est très grande. Il perd son sens quand il devient négatif mais l'espérance du nombre d'espèces en fonction de la surface a une forme plus générale :[61]

$$\mathbb{E}\left(S^n\right) = \frac{\alpha}{k}\left[1 - \left(\frac{\alpha}{\alpha + n}\right)^k\right] \tag{2.13}$$

L'équation 2.5 en est la limite pour $k \to 0$. Pour $-1 < k < 0$, avec A suffisamment grand, l'équation 2.13 revient à la relation d'Arrhenius, où $z = -k$. Les deux relations peuvent donc être obtenues à partir du même modèle mathématique dont seul le paramètre k diffère, mais aucune interprétation biologique sur le choix du paramètre n'est possible.

Empiriquement, il est difficile d'estimer lequel des modèles est le plus adapté aux données : sur 35 jeux de données, Connor et McCoy montrent que les régressions de S ou $\ln S$ en fonction de A ou $\ln A$ ont des performances similaires en termes de qualité d'ajustement (R^2), mais Triantis *et al.*[62] valident largement le modèle d'Arrhenius pour des SAR insulaires. Ce résultat est conforme à celui de Dengler[63] qui a testé 23 types de SAR et conclu que la loi de puissance est le modèle le plus universel.

Preston[64] argumente en faveur du modèle d'Arrhenius (et en défaveur de celui de Gleason) aux échelles intermédiaires et montre que la courbe aire-espèces est triphasique (Figure 2.11). Ce résultat fait maintenant consensus mais pas la délimitation des trois échelles, ni la forme de la relation aux échelles extrêmes. On notera que la partie locale de la courbe est en général une courbe d'accumulation, alors que les zones B et C sont traitées comme des SAR au sens strict.

Hubbell[65] établit que le modèle de Fisher est pertinent à l'échelle de la méta-communauté et donc de la communauté locale en absence de limitation de la dispersion (si le taux de migration m vaut 1, la communauté locale est un échantillon de la méta-communauté). Dans ce cas, la SAR a la forme de la courbe de la Figure 2.11, zone A, et l'indice α de Fisher est le bon outil de mesure. Mais la dispersion est limitée, et la distribution de la communauté locale est plus compliquée. McGill[66] montre qu'une SAD log-normale s'ajuste mieux aux données que le modèle neutre selon les résultats de Hubbell. Volkov *et al.*[67] établissent la SAD neutre de façon analytique qui s'ajuste mieux aux données que le modèle log-normal, mais la traduction en SAR n'est pas faite.

Le meilleur modèle à l'échelle locale est donc celui de Plotkin *et al.*,[68] qui réfute à la fois Gleason et Arrhenius. La SAC est

[59] P. Würtz et A. Annila (2008). « Roots of Diversity Relations ». In : *Journal of Biophysics* 2008, Article ID 654672.

[60] S. Engen (1974). « On species models frequency ». In : *Biometrika* 61.2, p. 263–270.

[61] Engen (1977). « Exponential and Logarithmic Species-Area Curves », cf. note 39, p. 15.

[62] K. A. Triantis et al. (2012). « The island species–area relationship : biology and statistics ». In : *Journal of Biogeography* 39.2, p. 215–231.

[63] Dengler (2009). « Which function describes the species-area relationship best ? A review and empirical evaluation », cf. note 28, p. 6.

[64] F. W. Preston (1960). « Time and Space and the Variation of Species ». In : *Ecology* 41.4, p. 611–627.

[65] Hubbell (2001). *The unified neutral theory of biodiversity and biogeography*, cf. note 11, p. 10, chapitre 6.

[66] B. J. McGill (2003). « A test of the unified neutral theory of biodiversity ». In : *Nature* 422.6934, p. 881–885.

[67] I. Volkov et al. (2003). « Neutral theory and relative species abundance in ecology ». In : *Nature* 424.6952, p. 1035–1037.

[68] Plotkin et al. (2000a). « Predicting species diversity in tropical forests », cf. note *a*, p. 20.

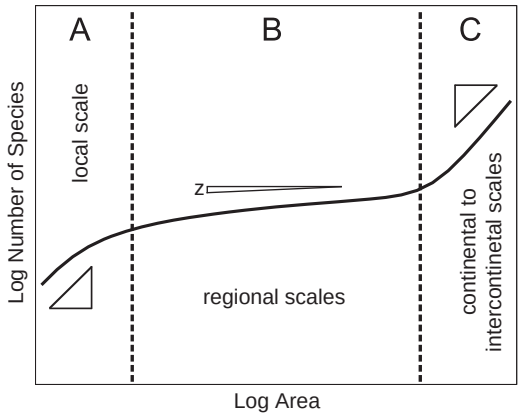


FIGURE 2.11 – Relation aire-espèce triphasique.[a] Zone B : relation d'Arrhenius.

[a] J. GRILLI et al. (2012). « Spatial aggregation and the species–area relationship across scales ». In : *Journal of Theoretical Biology* 313.0, p. 87–97.

concave sur une échelle log-log (Figure 2.8).

L'utilisation de l'indice α de Fisher est sujette à caution puisque la SAC locale ne correspond pas au modèle (figure 2.9). Sa valeur calculée plus haut pour BCI est 35, très différente de la valeur de θ du modèle neutre pour les mêmes données,[69] entre 47 et 50.

[69] VOLKOV et al. (2003). Cf. note 67, figure 1.

L'échelle plus grande (zone B de la Figure 2.11) est pour Hubbell un assemblage d'un certain nombre de communautés composant la méta-communauté (page 186). La SAR obtenue par simulation suit la loi de puissance d'Arrhenius. Grilli *et al.*[70] développent un modèle de SAR valable à toutes les échelles. Ses hypothèses sont peu réalistes dans la zone A, mais il valide la loi de puissance analytiquement à l'échelle intermédiaire.

[70] J. GRILLI et al. (2012). « Spatial aggregation and the species–area relationship across scales ». In : *Journal of Theoretical Biology* 313.0, p. 87–97.

La relation d'Arrhenius est donc systématiquement validée à l'échelle régionale (Figure 2.1). On remarquera que la distribution de Fisher, validée par Hubbell pour la méta-communauté, n'est jamais observée dans la SAR à cause de l'agrégation spatiale : en augmentant la surface d'échantillonnage, de nouvelles communautés sont ajoutées progressivement, mais l'échantillonnage de toute la méta-communauté n'est jamais réalisé (sauf au point marquant le passage de la zone B à la zone C). Les seuls exemples empiriques sont des inventaires d'arbres à l'échelle de l'Amazonie entière,[71] ou de la totalité des forêts tropicales[72] qui correspondent bien à la distribution de Fisher. Les valeurs de α à ces échelles sont bien supérieures aux valeurs locales (35 pour BCI, voir page 14) : 754 pour l'Amazonie, de l'ordre de 1000 pour les néotropiques, mais seulement de 250 pour l'Afrique, et de l'ordre de 2000 pour l'ensemble des tropiques.

[71] H. ter STEEGE et al. (2013). « Hyperdominance in the Amazonian Tree Flora ». In : *Science* 342.6156, p. 1243092.

[72] J. W. F. SLIK et al. (2015). « An estimate of the number of tropical tree species ». In : *Proceedings of the National Academy of Sciences* in press.

Enfin, quand l'échelle est encore plus grande, le nombre d'espèces augmente à nouveau plus vite parce que d'autres méta-communautés sont intégrées.

2.2 Taux de couverture

2.2.1 Formule des fréquences de Good-Turing

La relation fondamentale entre les fréquences des espèces est due à Turing et a été publiée par Good.[73] 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 moyenne α_ν d'une espèce représentée ν fois (c'est-à-dire par ν individus) au rapport entre les nombres d'espèces représentées $\nu+1$ fois et ν fois.

[73] I. J. Good (1953). « The Population Frequency of Species and the Estimation of Population Parameters ». In : *Biometrika* 40.3/4, p. 237–264.

$$\alpha_\nu \approx \frac{(\nu+1)}{n}\frac{s^n_{\nu+1}}{s^n_\nu} \tag{2.14}$$

Les singletons (s^n_1 : le nombre d'espèces observées une seule fois) et les doubletons (s^n_2 : le nombre d'espèces observées deux fois) ont une importance centrale. Pour $\nu = 1$, on a : $\alpha_1 = {}^{2s^n_2}/_{s^n_1}$: la fréquence d'une espèce typiquement représentée par un singleton est environ égale à deux fois le rapport entre le nombre des singletons et des doubletons. Pour $\nu = 0$, l'ignorance du nombre d'espèces non échantillonnées s^n_0 pose problème mais le produit $\alpha_0 \times s^n_0 = \pi_0$, la probabilité totale des espèces non représentées, peut être estimée par ${}^{s^n_1}/_n$. Ces relations sont le fondement des estimateurs de richesse de Chao présentées ci-dessous.

La relation a été précisée[74] en limitant les approximations dans les calculs. La seule nécessaire est que les probabilités des espèces représentées le même nombre de fois ν varient peu et puissent être considérées toutes égales à α_ν. Alors, α_ν est estimé par :

[74] C.-H. Chiu et al. (2014a). « An Improved Nonparametric Lower Bound of Species Richness via a Modified Good-Turing Frequency Formula ». In : *Biometrics* 70.3, p. 671–682, eq. 6 et 7a.

$$\hat{\alpha}_\nu = \frac{(\nu+1)s^n_{\nu+1}}{(n-\nu)s^n_\nu + (\nu+1)s^n_{\nu+1}} \tag{2.15}$$

Ce nouvel estimateur est à la base de l'estimateur de Chao amélioré et des estimateurs d'entropie de Chao et Jost (pages 48 et 67).

2.2.2 Taux de couverture et déficit de couverture

Good[75] définit le taux de couverture (*sample coverage*) de l'échantillonnage comme la proportion des espèces découvertes $C = \sum_{s=1}^{S} \mathbf{1}(n_s > 0)\, p_s$ où $\mathbf{1}(\cdot)$ est la fonction indicatrice. Son complément à 1 est appelé déficit de couverture (*coverage deficit*).

[75] Good (1953). Cf. note 73.

Le taux de couverture augmente avec l'effort d'échantillonnage. Plus il est élevé, meilleures seront les estimations de la diversité. Pour comparer deux communautés par des courbes de raréfaction, Chao et Jost[76] montrent que le taux de couverture plutôt que la taille de l'échantillon doit être identique. Les estimateurs de la diversité développés plus loin reposent largement sur cette notion

[76] A. Chao et L. Jost (2012). « Coverage-based rarefaction and extrapolation : standardizing samples by completeness rather than size ». In : *Ecology* 93.12, p. 2533–2547.

pour la correction du biais d'échantillonnage[77] (la sous-estimation systématique de la diversité due aux espèces non observées, un des éléments du biais d'estimation).

L'estimateur du taux de couverture, que Good attribue à Turing, est selon la relation des fréquences vues plus haut :

$$\hat{C} = 1 - \frac{s_1^n}{n} \tag{2.16}$$

Cet estimateur est biasé.[78] En réalité :

$$C = 1 - \frac{\mathbb{E}(S_1^n) - \pi_1}{n} \tag{2.17}$$

L'estimateur de Good néglige le terme π_1, la somme des probabilités des espèces observées une fois. Ce terme peut être estimé avec un biais plus petit. Chao *et al.*[79] puis Zhang et Huang proposent l'estimateur suivant, qui utilise toute l'information disponible et a le plus petit biais possible :

$$\hat{C} = 1 - \sum_{\nu=1}^{n} (-1)^{\nu+1} \binom{n}{\nu}^{-1} s_\nu^n \tag{2.18}$$

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,[80] complété par Zhang et Zhang,[81] a montré que l'estimateur était asymptotiquement normal et a calculé l'intervalle de confiance de $\hat{C}$:

$$C = \hat{C} \pm t_{1-\alpha/2}^n \frac{\sqrt{s_1^n \left(1 - \frac{s_1^n}{n}\right) + 2s_2^n}}{n} \tag{2.19}$$

Où $t_{1-\alpha/2}^n$ 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[82] disponible sur internet[83]. Il est la base des estimateurs d'entropie de Chao et Jost (page 67). L'estimation de l'équation (2.17) donne la relation :

$$\hat{C} = 1 - \frac{s_1^n - \hat{\pi}_1}{n} \tag{2.20}$$

Or, $\hat{\pi}_1 = s_1^n \hat{\alpha}_1$. α_1 peut être estimé par la relation de Good-Turing (2.15), en remplaçant s_0^n par l'estimateur Chao1 (équation 2.23). Alors :

$$\hat{C} = 1 - \frac{s_1^n}{n}(1 - \hat{\alpha}_1) = 1 - \frac{s_1^n}{n}\left[\frac{(n-1)s_1^n}{(n-1)s_1^n + 2s_2^n}\right] \tag{2.21}$$

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

[77] G. Dauby et O. J. Hardy (2012). « Sampled-based estimation of diversity sensu stricto by transforming Hurlbert diversities into effective number of species ». In : *Ecography* 35.7, p. 661–672.

[78] Z. Zhang et H. Huang (2007). « Turing's formula revisited ». In : *Journal of Quantitative Linguistics* 14.2-3, p. 222–241.

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

[80] W. W. Esty (1983). « A Normal Limit Law for a Nonparametric Estimator of the Coverage of a Random Sample ». In : *The Annals of Statistics* 11.3, p. 905–912.

[81] C.-H. Zhang et Z. Zhang (2009). « Asymptotic Normality of a Nonparametric Estimator of Sample Coverage ». English. In : *Annals of Statistics* 37.5A, p. 2582–2595.

[82] A. Chao et al. (2010). « Phylogenetic diversity measures based on Hill numbers ». In : *Philosophical Transactions of the Royal Society B* 365.1558, p. 3599–3609.

[83] `http://chao.stat.nthu.edu.tw/softwareCE.html`

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

## ZhangHuang
##  0.9220438
```

Chao et Jost[84] montrent que 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 2.2) est égale au déficit de couverture :

$$1 - \mathbb{E}\left(C^n\right) = \mathbb{E}\left(S^{n+1}\right) - \mathbb{E}\left(S^n\right) \tag{2.22}$$

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.

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 et Lin[85] et Hwang *et al.*[86]

2.3 Richesse spécifique

On ne fait pas de supposition sur la forme de la SAD quand on utilise des méthodes d'estimation non paramétriques. Les estimateurs les plus connus sont ceux de Chao[87] et le *jackknife*.[88]

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[89] et il peut être efficace d'utiliser plusieurs méthodes d'estimation de façon concurrente sur les mêmes données.[90]

2.3.1 Techniques d'estimation non paramétrique

Dans le cadre d'un échantillonnage de n individus, on observe $S^n_{\neq 0}$ 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 $1 - (1 - p_s)^n$.

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

[84] Chao et Jost (2012). « Coverage-based rarefaction and extrapolation : standardizing samples by completeness rather than size », cf. note 76, p. 26.

[85] A. Chao et C.-W. Lin (2012). « Nonparametric Lower Bounds for Species Richness and Shared Species Richness under Sampling without Replacement ». English. In : *Biometrics* 68.3, p. 912–921.

[86] W.-H. Hwang et al. (2014). « Good-Turing frequency estimation in a finite population ». In : *Biometrical journal* 57.2, p. 321–339.

[87] A. Chao (1984). « Nonparametric estimation of the number of classes in a population ». In : *Scandinavian Journal of Statistics* 11, p. 265–270.

[88] K. P. Burnham et W. S. Overton (1979). « Robust Estimation of Population Size When Capture Probabilities Vary Among Animals ». In : *Ecology* 60.5, p. 927–936.

[89] R. B. O'Hara (2005). « Species richness estimators : How many species can dance on the head of a pin ? » In : *Journal of Animal Ecology* 74, p. 375–386.

[90] Y. Basset et al. (2012). « Arthropod Diversity in a Tropical Forest ». In : *Science* 338.6113, p. 1481–1484.

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 et Colwell[91] qui concluent notamment que les estimateurs ne peuvent fournir qu'une borne inférieure de l'intervalle des possibles valeurs du nombre réel d'espèces.

[91] C. X. Mao et R. K. Colwell (2005). « Estimation of Species Richness : Mixture Models, the Role of Rare Species, and Inferential Challenges ». In : *Ecology* 86.5, p. 1143–1153.

Chao1 et Chao2

Chao[92] estime le nombre d'espèces non observées à partir de celles observées 1 ou 2 fois. Il s'agit d'un estimateur minimum, valide à condition que les singletons et doubletons représentent une part importante de l'information. L'estimateur est :

[92] Chao (1984). Cf. note 87.

$$\hat{S}_{Chao1} = s^n_{\neq 0} + \frac{(n-1)\,(s^n_1)^2}{2n s^n_2} \tag{2.23}$$

$s^n_{\neq 0}$ est le nombre d'espèces différentes observé, s_ν le nombre d'espèces observées ν fois. L'échantillon est de n individus.

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

$$\hat{S}_{Chao1} = s^n_{\neq 0} + \frac{(n-1)\,s^n_1\,(s^n_1-1)}{2n} \tag{2.24}$$

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

$$\hat{S}_{Chao1} = s^n_{\neq 0} + \frac{(s^n_1)^2}{2 s^n_2} \tag{2.25}$$

Si aucune espèce n'est observée deux fois, l'estimateur est remplacé[93] par $\hat{S}_{Chao1} = s^n_{\neq 0} + s^n_1\,(s^n_1-1)/2$.

[93] A. Chao (2004). « Species richness estimation. » In : *Encyclopedia of Statistical Sciences*. Sous la dir. de N Balakrishnan et al. 2nd ed. New York : Wiley.

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

$$\mathrm{Var}\left(\hat{S}_{Chao1}\right) = s^n_2 \left[\frac{1}{2}\left(\frac{s^n_1}{s^n_2}\right)^2 + \left(\frac{s^n_1}{s^n_2}\right)^3 + \frac{1}{4}\left(\frac{s^n_1}{s^n_2}\right)^4\right] \tag{2.26}$$

Si aucune espèce n'est observée deux fois :

$$\mathrm{Var}\left(\hat{S}_{Chao1}\right) = \frac{s^n_1\,(s^n_1-1)}{2} + \frac{s^n_1(2s^n_1-1)^2}{4} + \frac{(s^n_1)^4}{4 s^n_{\neq 0}} \tag{2.27}$$

Chao[94] calcule une approximation de l'intervalle de confiance à 95% :

[94] A. Chao (1987). « Estimating the population size for capture-recapture data with unequal catchability ». In : *Biometrics* 43.4, p. 783–791.

$$s^n_{\neq 0} + \frac{\hat{S}_{Chao1} - s^n_{\neq 0}}{c} \leq S \leq s^n_{\neq 0} + \left(\hat{S}_{Chao1} - s^n_{\neq 0}\right) c \qquad (2.28)$$

Où :

$$c = e^{t^n_{1-\alpha/2}\sqrt{\ln\left(1+\frac{\mathrm{Var}(\hat{S}_{Chao1})}{\left(\hat{S}_{Chao1}-s^n_{\neq 0}\right)^2}\right)}} \qquad (2.29)$$

Eren *et al.*[95] (eq. 8) 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[96] 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.*[97] améliorent l'estimateur en reprenant la démarche originale de Chao mais en utilisant un estimateur plus précis du taux de couverture, (2.21) au lieu de (2.16) :

$$\hat{S}_{iChao1} = \hat{S}_{Chao1} + \frac{s^n_3}{4s^n_4} \max\left(s^n_1 - \frac{s^n_2 s^n_3}{2s^n_4}; 0\right) \qquad (2.30)$$

L'estimateur ACE

Chao et Lee[98] 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 ${}^{\nu}s$ correspondant aux espèces rares : concrètement, la valeur 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 :

$$\hat{S}_{ACE} = s^n_{>\kappa} + \frac{s^n_{\leq\kappa}}{\hat{C}_{rares}} + \frac{s^n_1}{\hat{C}_{rares}}\hat{\gamma}_{rares} \qquad (2.31)$$

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

L'estimateur du coefficient de variation est :

$$\hat{\gamma}^2_{rares} = \max\left(\frac{s^n_{\leq\kappa}\sum_{\nu=1}^{\kappa}\nu\left(\nu-1\right)s^n_\nu}{\hat{C}_{rares}\left(\sum_{\nu=1}^{\kappa}\nu s^n_\nu\right)\left(\sum_{\nu=1}^{\kappa}\nu s^n_\nu - 1\right)} - 1; 0\right) \qquad (2.32)$$

[95] M. I. Eren et al. (2012). « Estimating the Richness of a Population When the Maximum Number of Classes Is Fixed : A Nonparametric Solution to an Archaeological Problem ». English. In : *Plos One* 7.5.

[96] Chao (1987). « Estimating the population size for capture-recapture data with unequal catchability », cf. note 94, p. 29.

[97] Chiu et al. (2014a). « An Improved Nonparametric Lower Bound of Species Richness via a Modified Good-Turing Frequency Formula », cf. note 74, p. 26.

[98] A. Chao et S.-M. Lee (1992). « Estimating the Number of Classes Via Sample Coverage ». In : *Journal of the American Statistical Association* 87.417, p. 210–217.

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

$$\tilde{\gamma}^2_{rares} = \max\left(\widehat{\gamma}^2_{rares}\left(1 + \frac{\left(1-\hat{C}_{rares}\right)\sum_{\nu=1}^{\kappa}\nu\left(\nu-1\right)s^n_\nu}{\hat{C}_{rares}\left(\sum_{\nu=1}^{\kappa}\nu s^n_\nu - 1\right)}\right);0\right) \tag{2.33}$$

Chao et Shen[99] conseillent d'utiliser le deuxième estimateur dès que $\widehat{\gamma}^2_{rares}$ 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.

[99] A. Chao et T. J. Shen (2010). *Program SPADE : Species Prediction And Diversity Estimation. Program and user's guide.* Hsin-Chu, Taiwan.

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[100] ont utilisé cette technique pour obtenir des estimateurs du nombre d'espèces, appelés jackknife à l'ordre j, prenant en compte les valeurs de s^n_1 à s^n_j. Les estimateurs du premier et du deuxième ordre sont les plus utilisés en pratique :

[100] A. K. P. Burnham et W. S. Overton (1978). « Estimation of the Size of a Closed Population when Capture Probabilities vary Among Animals ». In : *Biometrika* 65.3, p. 625–633 ; Burnham et Overton (1979). « Robust Estimation of Population Size When Capture Probabilities Vary Among Animals », cf. note 88, p. 28.

$$\hat{S}_{J1} = s^n_{\neq 0} + \frac{(n-1)s^n_1}{n} \tag{2.34}$$

$$\hat{S}_{J2} = s^n_{\neq 0} + \frac{(2n-3)s^n_1}{n} - \frac{(n-2)^2 s^n_2}{n\,(n-1)} \tag{2.35}$$

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

Chao[101] a montré que les estimateurs jackknife pouvaient être retrouvés par approximation de l'indice Chao1.

[101] Chao (1984). « Nonparametric estimation of the number of classes in a population », cf. note 87, p. 28.

La variance du jackknife d'ordre 1 est[102] :

[102] J. F. Heltshe et N. E. Forrester (1983). « Estimating species richness using the jackknife procedure ». In : *Biometrics* 39.1, p. 1–11.

$$\mathrm{Var}\left(\hat{S}_{J1}\right) = \frac{n-1}{n}\left(\sum_{j=1}^{n} j^2 s^n_j - \frac{\left(s^n_{\neq 0}\right)^2}{n}\right) \tag{2.36}$$

L'estimateur jackknife n'a pas de support théorique solide[103] mais est efficace en pratique, notamment parce que son ordre peut être adapté aux données.

[103] R. M. Cormack (1989). « Log-Linear Models for Capture-Recapture ». In : *Biometrics* 45.2, p. 395–413.

L'estimateur du bootstrap

L'estimateur du bootstrap[104] est :

[104] E. P. Smith et G. V. Belle (1984). « Nonparametric Estimation of Species Richness ». In : *Biometrics* 40.1, p. 119–129.

$$\hat{S}_b = s^n_{\neq 0} + \sum_s (1-P_s)^n \tag{2.37}$$

Il est peu utilisé parce que le jackknife est plus performant.[105]

[105] R. K. Colwell et J. A. Coddington (1994). « Estimating Terrestrial Biodiversity through Extrapolation ». In : *Philosophical Transactions : Biological Sciences* 345.1311, p. 101–118.

Calcul

Ces estimateurs peuvent être calculés de façon relativement simple à l'aide du logiciel SPADE. 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 sous R. Le package *vegan* 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 (non détaillée ici). 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*[106] 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.

[106] J.-P. Wang (2011). « SPECIES : An R Package for Species Richness Estimation ». In : *Journal of Statistical Software* 40.9, p. 1–15.

Comparaison des fonctions sur l'ensemble du dispositif BCI ($s^n_{\neq 0} = 225, s_1 = 19$) :

```
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)
Ns <- 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 AbdFreqCount dans entropart
jackknife(AbdFreqCount(Ns))

##
## 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 (2.34) :

```
# Nombre d'espèces par efffectif observé
DistNs <- tapply(Ns, Ns, length)
# Calcul direct de Jack1
sum(DistNs) + DistNs[1] * (sum(BCI) - 1)/sum(BCI)

##        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 :

```
Ps <- Ns/sum(Ns)
length(Ps) + sum((1 - Ps)^N)

## [1] 234.3517
```

Estimation à partir de placettes distantes

Krishnamani *et al.*[107] développent une méthode permettant l'extrapolation de la richesse mesurée sur de petites surfaces (48 placettes de 0,25 ha) à de très grandes zones (60 000 km^2, la taille des Western Ghats en Inde).

Partant de la relation d'Arrhenius vue précédemment, le nombre d'espèces dans une surface A_1 est :

[107] R. Krishnamani et al. (2004). « Estimating species richness at large spatial scales using data from small discrete plots ». In : *Ecography* 27.5, p. 637–642.

$$S^{A_1} = (A_1/A_0)^z S^{A_0} \tag{2.38}$$

c'est-à-dire le nombre d'espèces dans la surface A_0 plus petite multipliée par le rapport des surfaces à la puissance z. La valeur de z est constante pour des variations de surface limitées. L'idée est donc de procéder par étapes, en partant du nombre d'espèces mesuré sur les petites placettes, et extrapolant vers une surface plus grande. Cette surface est ensuite prise pour référence pour une nouvelle extrapolation avec une nouvelle valeur de z. En indiçant les surfaces de 0 à n, on obtient :

$$S^{A_n} = (A_n/A_{n-1})^{z_{n-1}}(A_{n-1}/A_{n-2})^{z_{n-2}} \dots (A_1/A_0)^{z_0} S^{A_0} \tag{2.39}$$

Il reste à évaluer les différentes valeurs de z. À l'intérieur de chaque placette, l'équation (2.38) est suffisante. À plus grande distance, les auteurs utilisent l'indice de Sørensen, c'est-à-dire deux fois la fraction d'espèces communes entre deux placettes : $\chi = 2(S^1 \cap S^2)/(S^1 + S^2)$. L'indice dépend principalement de la surface des placettes A et de leur éloignement d. Pour certaines plages de distances $[d_{\min}; d_{\max}]$ vérifiant $A/d_{\min}^2 \ll 1$ (placettes éloignées), χ est proportionnel à $(A/d^2)^z$. La valeur de z peut donc être calculée en comptant le nombre d'espèces communes entre chaque paire de placettes, et ajustant $\ln \chi$ à $\ln (A/d^2)$ sur des plages de distances convenables pour que z reste constant, et donc que la relation soit linéaire.

Cette méthode permet l'extrapolation à de très grandes surfaces de mesures de richesse faites sur de petites placettes à condition qu'elles soient assez nombreuses et éloignées les unes des autres.

Choix de l'estimateur

Des tests poussés ont été menés par Brose *et al.*[108] pour permettre le choix du meilleur estimateur de la richesse en fonction de la complétude de l'échantillonnage $s^n_{\neq 0}/S$. Les auteurs appellent cette proportion couverture (*coverage*). Le terme *completeness* a été proposé par Beck et Schwanghart[109] pour éviter la confusion avec le taux de couverture défini par Good (vu page 26). 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 2.12. Le choix 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

[108] U. Brose et al. (2003). « Estimating species richness : Sensitivity to sample coverage and insensitivity to spatial patterns ». In : *Ecology* 84.9, p. 2364–2377.

[109] J. Beck et W. Schwanghart (2010). « Comparing measures of species diversity from incomplete inventories : an update ». In : *Methods in Ecology and Evolution* 1.1, p. 38–44.

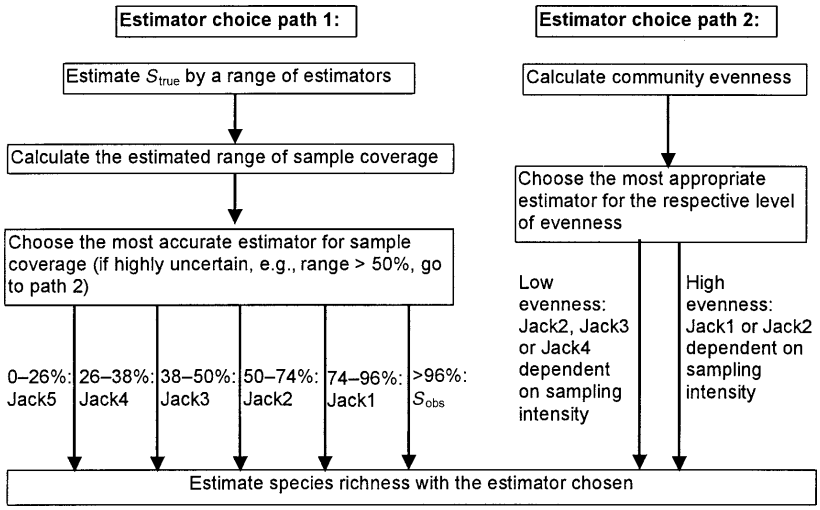


FIGURE 2.12 – Arbre de décision du meilleur estimateur du nombre d'espèces, *in* Brose *et al.*.[a]

[a] U. BROSE et al. (2003). « Estimating species richness : Sensitivity to sample coverage and insensitivity to spatial patterns ». In : *Ecology* 84.9, p. 2364–2377.

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 page 76). Si elle est faible (de l'ordre de 0,5-0,6), 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 (0,8-0,9), on préférera jackknife 1 ou 2.

Exemple de 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.

Chiu *et al.*,[110] 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 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.[111]

[110] CHIU et al. (2014a). « An Improved Nonparametric Lower Bound of Species Richness via a Modified Good-Turing Frequency Formula », cf. note 74, p. 26.

[111] É. MARCON (2015). « Practical Estimation of Diversity from Abundance Data ». In : *HAL* 01212435.version 2.

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.*[112] proposent un estimateur et le confrontent avec succès à des estimateurs antérieurs. On note $\hat{S}_0^n$ 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 :

[112] T. J. SHEN et al. (2003). « Predicting the number of new species in a further taxonomic sampling ». In : *Ecology* 84, p. 798–804.

$$\hat{S}' = \hat{S}_0^n \left[1 - \left(1 - \frac{1 - \hat{C}}{\hat{S}_0^n} \right)^{n'} \right] \tag{2.40}$$

$\hat{S}_0^n$ est obtenu par la différence entre les nombres d'espèces estimé et observé : $\hat{S}_0^n = \hat{S} - s_{\neq 0}^n$.

Exemple de BCI, suite du code de la page 33 : combien de nouvelles espèces seront découvertes en échantillonnant plus ?

```
# Espèces non observées
(NonObs <- jackknife(AbdFreqCount(Ns))$Nhat - length(Ns))

##
## Your specified order is larger than that determined by the test,
## Therefore the order from the test is used.
## [1] 19

# Taux de couverture
(C <- Coverage(Ns))

## ZhangHuang
##  0.9991146
```

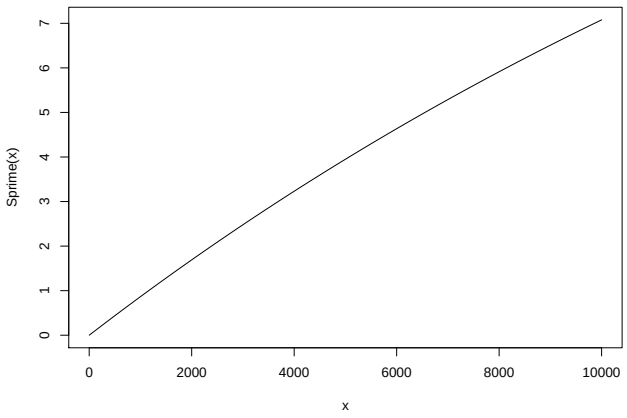


FIGURE 2.13 – Nombres de nouvelles espèces découvertes en fonction de l'effort d'échantillonnage supplémentaire (données de BCI). Seulement 7 nouvelles espèces seront observées en échantillonnant 10000 arbres supplémentaires (environ 25 ha en plus des 50 ha de la parcelle qui contiennent 225 espèces).

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 2.13 [113].

[113] Le code R nécessaire pour réaliser la figure est :

```
# Nouvelles espèces en fonction du nombre de nouveaux
# individus
Sprime <- function(Nprime) NonObs * (1 - (1 - (1 - C)/NonObs)^Nprime)
curve(Sprime, 1, 10000)
```

2.3.2 Inférence du nombre d'espèces à partir de la SAD

Distribution de Preston

Preston[114] 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* :

[114] Preston (1948). « The Commonness, And Rarity, of Species », cf. note 32, p. 8.

```
library(vegan)
data(BCI)
veiledspec(colSums(BCI))

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

L'ajustement direct du modèle aux données, sans regroupement par octaves,[115] est également possible (Figure 2.14 [116]) :

[115] M. Williamson et K. J. Gaston (2005). « The lognormal distribution is not an appropriate null hypothesis for the species-abundance distribution ». In : *Journal of Animal Ecology* 74.2001, p. 409–422.

```
mod.ll <- prestondistr(colSums(BCI))
veiledspec(mod.ll)

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

Maximum de vraisemblance d'une distribution de Fisher

Norris et Pollock[117] supposent que la distribution des espèces suit le modèle de Fisher (voir page 14) 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).

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

```
library(SPECIES)
# Distribution du nombre d'espèces (vecteur: noms = nombre
# d'individus valeurs = nombres d'espèces ayant ce nombre
# d'individus)
Ns <- colSums(BCI)
# Mise au format requis (matrice: colonne 1 = nombre
# d'individus colonne 2 = nombres d'espèces ayant ce nombre
# d'individus)
DistNs.SPECIES <- AbdFreqCount(Ns)
# Regroupement de la queue de distribution: la longueur du
# vecteur est limitée à 25 pour alléger les calculs.
DistNs.SPECIES[25, 2] <- sum(DistNs.SPECIES[25:nrow(DistNs.SPECIES),
    2])
DistNs.SPECIES <- DistNs.SPECIES[1:25, ]
unpmle(DistNs.SPECIES)
```

```
## Method: Unconditional NPMLE method
##  by Norris and Pollock 1996, 1998,
##  using algorithm by Wang and Lindsay 2005:
##
##  MLE=                      239
```

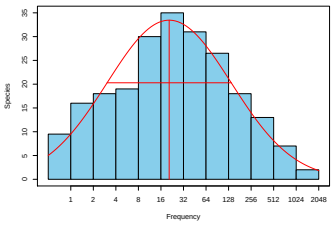


Figure 2.14 – Ajustement du modèle de Preston aux données de BCI.

[117] J. L. Norris et K. H. Pollock (1998). « Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species ». In : *Environmental and Ecological Statistics* 5.4, p. 391–402.

[116] Le code R nécessaire pour réaliser la figure est :

```
plot(mod.ll)
```

```
##  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.*[118] ont amélioré sa stabilité en pénalisant le calcul de la vraisemblance :

[118] J.-P. Wang et al. (2005). « Gene capture prediction and overlap estimation in EST sequencing from one or multiple libraries ». In : *BMC Bioinformatics* 6.1, p. 300.

```
pnpmle(DistNs.SPECIES)
```

```
## Method: Penalized NPMLE method by  Wang and Lindsay 2005.
##
##  MLE=                    243
##  Estimated Poisson mixture components:
##  p=                      0.9033625 3.397048 10.59029
##  pi=                     0.2771211 0.3208847 0.4019942
##
## $Nhat
## [1] 243
```

Enfin, Wang[119] 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.

[119] J.-P. Wang (2010). « Estimating species richness by a Poisson-compound gamma model ». In : *Biometrika* 97.3, p. 727–740.

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(DistNs.SPECIES)
```

```
## Method: Poisson-Compound Gamma method by Wang 2010.
## Alpha grid used: 1 2 3 4 5 6 7 8 9 10 .
##
##  MLE=                    236
##  Selected alpha model:   Inf
##  Estimated Gamma components:
##  p=                 1.351635 4.049317 10.85329 25.92447
##  pi=                0.1464461 0.09874953 0.1547005 0.6001039
##
## $Nhat
## [1] 236
```

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

Cette approche consiste à extrapoler la courbe aire-espèces, en contradiction avec les modèles présentés plus haut (Fisher, Arrhenius) qui prévoient que le nombre d'espèces augmente indéfiniment.

Le modèle le plus connu est celui de Michaelis-Menten[120] proposé par Clench.[121] 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 :

$$S^n = S\frac{n}{K+n} \tag{2.41}$$

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[122]. Les 50 carrés de BCI sont utilisés pour fabriquer une courbe d'accumulation :

```
data(BCI)
# Cumul du nombre d'arbres
nArbres <- cumsum(rowSums(BCI))
# Cumul de l'inventaire
Cumul <- apply(BCI, 2, cumsum)
# Nombre d'espèces cumulées
nEspeces <- apply(Cumul, 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 2.16.

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

## Nonlinear regression model
##   model: nEspeces ~ S * nArbres/(K + nArbres)
##    data: list(nArbres, nEspeces)
##      K      S
## 1251.0  232.8
##  residual sum-of-squares: 3066
##
## Number of iterations to convergence: 6
## Achieved convergence tolerance: 7.881e-06
```

L'estimation précédente utilise la méthode des moindres carrés, qui suppose l'indépendance des résidus, hypothèses évidemment violée par une courbe d'accumulation.[123] L'estimation par le maximum de vraisemblance est plus convenable.[124] Elle utilise la totalité des points de la courbe d'accumulation. La courbe d'accumulation de BCI a été calculée page 12. Ses données sont utilisées ici :

```
N <- length(SAC)
# Calculs intermédiaires
Yi <- SAC
Xi <- Yi/(1:N)
Xbar <- mean(Xi)
Ybar <- mean(Yi)
Syy <- sum((Yi - Ybar)^2)
Sxx <- sum((Xi - Xbar)^2)
Sxy <- sum((Xi - Xbar) * (Yi - Ybar))
# Estimations
(Khat <- (Xbar * Syy - Ybar * Sxy)/(Ybar * Sxx - Xbar * Sxy))
```

[120] L. Michaelis et M. L. Menten (1913). « Die Kinetik der Invertinwirkung ». In : *Biochemische Zeitschrift* 49, p. 333–369.

[121] H. K. Clench (1979). « How to make regional lists of butterflies : some thoughts ». In : *Journal of the Lepidopterists' Society* 33.4, p. 216–231.

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

[123] Colwell et Coddington (1994). « Estimating Terrestrial Biodiversity through Extrapolation », cf. note 105, p. 32.

[124] J. G. W. Raaijmakers (1987). « Statistical Analysis of the Michaelis-Menten Equation ». In : *Biometrics* 43.4, p. 793–803.

```
## [1] 407.5794

(Shat <- Ybar + Khat * Xbar)

## [1] 215.5661
```

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 Khat doit annuler
f <- function(Khat) Sxy + Khat * Sxx - (Syy + 2 * Khat * Sxy +
    Khat * Khat * Sxx) * sum(Xi/(Yi + Khat * Xi)/N)
# Résolution numérique, l'intervalle de recherche doit être
# fourni
Solution <- uniroot(f, c(0, 1e+07))
(Khat <- Solution$root)

## [1] 407.6541

(Shat <- Ybar + Khat * Xbar)

## [1] 215.5691
```

Le nombre d'espèces estimé est 216, inférieur au nombre d'espèces observé.

Pour calculer l'intervalle de confiance, il est plus simple de passer par une transformation linéaire du modèle[125] :

$$\left[\frac{1}{\hat{S}^n}\right] = \frac{K}{\hat{S}}\left[\frac{1}{n}\right] + \frac{1}{\hat{S}} \tag{2.42}$$

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

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

## (Intercept)
##     217.002
```

[125] H. Lineweaver et D. Burk (1934). « The Determination of Enzyme Dissociation Constants ». In : *Journal of the American Chemical Society* 56.3, p. 658–666.

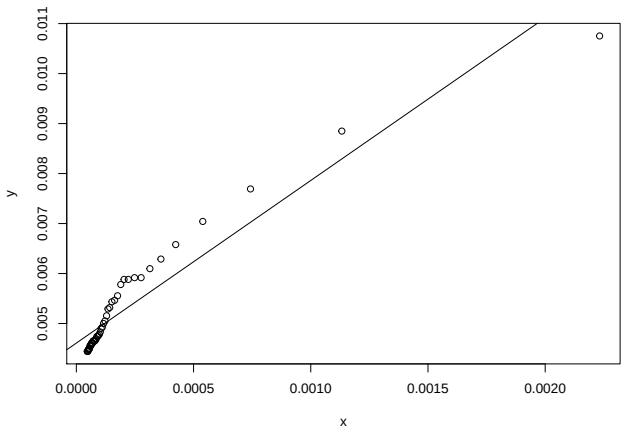


Figure 2.15 – Ajustement du même modèle de Michaelis-Menten transformé selon Lineweaver et Burk.

On voit assez clairement que le modèle (Figure 2.15 [126]) s'ajuste mal quand il est représenté sous cette forme.[127]

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 et Llorente[128] 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 :

$$S^n = S\left(1 - e^{Kn}\right) \tag{2.43}$$

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

```
(nlsexp <- nls(nEspeces ~ S * (1 - exp(K * nArbres)), data = list(nArbres,
    nEspeces), start = list(S = max(nEspeces), K = -1/1000)))

## Nonlinear regression model
##   model: nEspeces ~ S * (1 - exp(K * nArbres))
##    data: list(nArbres, nEspeces)
##          S          K
## 212.198082  -0.000494
##  residual sum-of-squares: 10664
##
## Number of iterations to convergence: 13
## Achieved convergence tolerance: 8.042e-06
```

Ce modèle, proposé par Holdridge *et al.*,[129] 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 :

$$S^n = \frac{1}{z} \ln \left[\frac{a}{c} - \frac{a-c}{c} e^{-czn} \right] \tag{2.44}$$

Les paramètres à estimer sont z, a et c.

```
(nlslog <- nls(nEspeces ~ 1/z * log(a/c - (a - c)/c * exp(-c *
    z * nArbres)), data = list(nArbres, nEspeces), start = list(z = 0.05,
    a = 1, c = 0.001)))
```

[126] Le code R nécessaire pour réaliser la figure est :

```
plot(y ~ x)
abline(lm1$coef)
```

[127] Raaijmakers (1987). « Statistical Analysis of the Michaelis-Menten Equation », cf. note 124, p. 39.

[128] J. Soberón M. et J. Llorente B. (1993). « The Use of Species Accumulation Functions for the Prediction of Species Richness ». In : *Conservation Biology* 7.3, p. 480–488.

[129] L. Holdridge et al. (1971). *Forest environments in tropical life zones*. Oxford : Pergamon Press.

```
## Nonlinear regression model
##   model: nEspeces ~ 1/z * log(a/c - (a - c)/c * exp(-c * z * nArbres))
##    data: list(nArbres, nEspeces)
##         z         a         c
## 0.025139 0.755365 0.001114
##  residual sum-of-squares: 446.1
##
## Number of iterations to convergence: 4
## Achieved convergence tolerance: 4.813e-06

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

##        a
## 259.3128
```

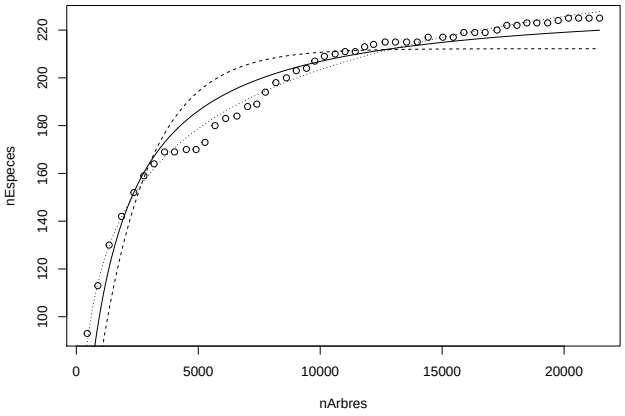


FIGURE 2.16 – Ajustement des modèle de Michaelis-Menten (trait plein) et de de Soberón et Llorente (pointillés longs : modèle exponentiel négatif ; pointillés courts : modèle à trois paramètres) aux données de BCI. Les cercles représentent le nombre d'espèces cumulées en fonction du nombre d'arbres. 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 2.16 [130] 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.

2.3.4 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.[131] Le

[130] Le code R nécessaire pour réaliser la figure est :

```
plot(nArbres, nEspeces)
# Courbe ajustée
x <- seq(from = 0, to = max(nArbres), length = 255)
lines(x = x, y = predict(nlsfit, newdata = list(nArbres = x)))
lines(x = x, y = predict(nlsexp, newdata = list(nArbres = x)),
    lty = 2)
lines(x = x, y = predict(nlslog, newdata = list(nArbres = x)),
    lty = 3)
```

[131] A. BALMFORD et al. (1996b). « Using higher-taxon richness as a surrogate for species richness : II. Local applications ». In : *Proceedings of the Royal Society B* 263, p. 1571–1575.

choix d'estimer la diversité de taxons de rang supérieur (genres ou même familles au lieu des espèces) est envisageable.[132]

Empiriquement, la corrélation entre la richesse générique et la richesse spécifique (des angiospermes, des oiseaux et des mammifères) est bonne en forêt tropicale,[133] 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.*[134] 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,[135] 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.

[132] P. Williams et K. J. Gaston (1994). « Measuring more of biodiversity : can higher-taxon richness predict wholesale species richness ? » In : *Biological Conservation* 67, p. 211–217.

[133] A. Balmford et al. (1996a). « Using higher-taxon richness as a surrogate for species richness : I. Regional tests ». In : *Proceedings of the Royal Society B* 263, p. 1267–1274.

[134] C. C. Cartozo et al. (2008). « Quantifying the taxonomic diversity in real species communities ». In : *Journal of Physics A : Mathematical and Theoretical* 41.22.

[135] G. Caldarelli et al. (2002). « Scale-free networks from varying vertex intrinsic fitness. » In : *Physical Review Letters* 89.25, p. 258702.

2.3.5 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.*[136] 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 2.17, A à E). Le nombre de taxon de chaque niveau est lié à celui du niveau précédent, ce qui est représenté par la Figure 2.17, G 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.[137]

En se limitant aux arbres, l'estimation se monte à 16 000 espèces pour l'Amazonie,[138] de l'ordre de 5000 pour l'Afrique et entre 40000 et 53000 pour l'ensemble des tropiques[139] (donc pour l'ensemble de la planète, le nombre d'espèces non-tropicales étant

[136] C. Mora et al. (2011). « How Many Species Are There on Earth and in the Ocean ? » In : *PLoS Biol* 9.8, e1001127.

[137] R. M. May (2011). « Why Worry about How Many Species and Their Loss ? » In : *PLoS Biol* 9.8, e1001130.

[138] Steege et al. (2013). « Hyperdominance in the Amazonian Tree Flora », cf. note 71, p. 25.

[139] Slik et al. (2015). « An estimate of the number of tropical tree species », cf. note 72, p. 25.

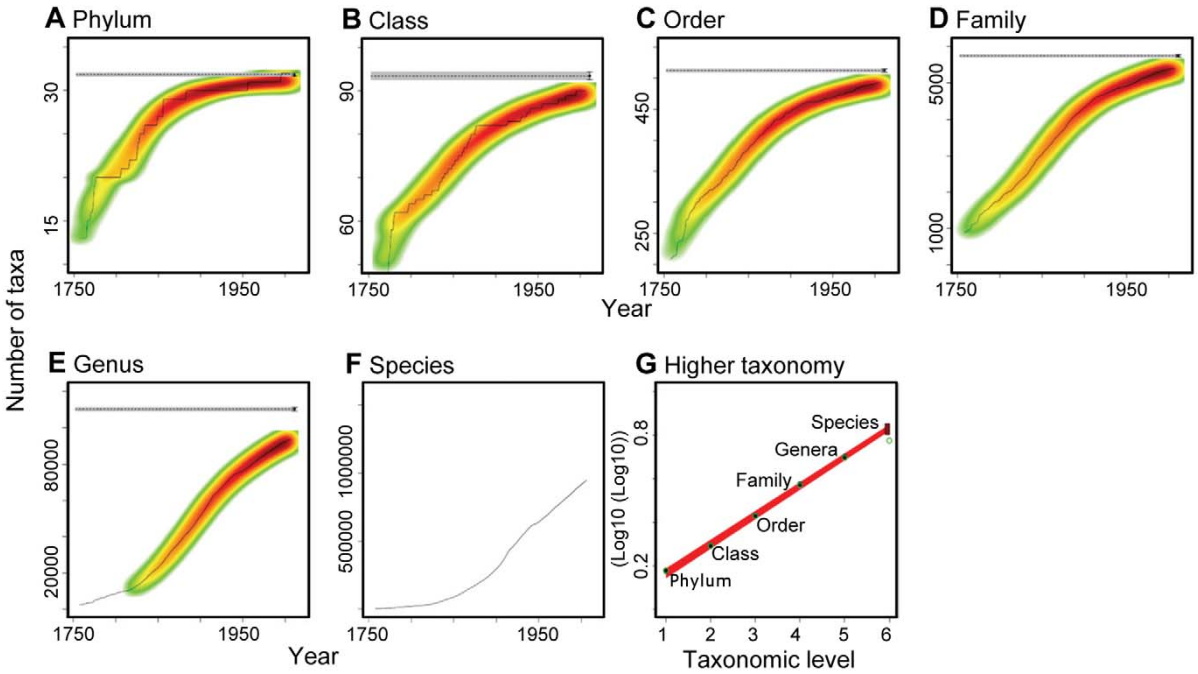


FIGURE 2.17 – Évolution du nombre de taxons connus en fonction du temps et valeur estimée du nombre total (ligne horizontale grise, A à E). Le nombre d'espèces connues est trop faible pour utiliser cette méthode (F). Il est obtenu par extrapolation de la relation du nombre de taxons d'un niveau à celui du niveau supérieur (G). *In* C. MORA et al. (2011). « How Many Species Are There on Earth and in the Ocean ? » In : *PLoS Biol* 9.8, e1001127, fig. 1 , Animalia seulement.

négligeable).

Wilson *et al.*[140] 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[141]). 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.*

2.3.6 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.[142] 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.[143] L'autre aspect, avec des conséquences pratiques plus immédiates, concerne leur délimitation. Mayden[144] recense 22 définitions différentes du concept d'espèce.

Le concept le plus répandu est celui d'espèce *biologique*[145] :

[140] J. B. WILSON et al. (2012). « Plant species richness : The world records ». In : *Journal of Vegetation Science* 23.4, p. 796–802.

[141] J. H. CONNELL (1978). « Diversity in Tropical Rain Forests and Coral Reefs ». In : *Science* 199.4335, p. 1302–1310.

[142] E. CASETTA (2014). « Évaluer et conserver la biodiversité face au problème des espèces ». In : *La biodiversité en question. Enjeux philosophiques, éthiques et scientifiques*. Sous la dir. d'E. CASETTA et J. DELORD. Paris : Editions Matériologiques. Chap. 5, p. 139–154.

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

[144] R. L. MAYDEN (1997). « A hierarchy of species concepts : the denouement in the saga of the species problem ». In : *Species. The units of biodiversity*. Sous la dir. de M. F. CLARIDGE et al. London : Chapman et Hall. Chap. 19, p. 381–424.

[145] E. MAYR (1942). *Systematics and the origin of species from the viewpoint of a zoologist*. New York : Columbia University Press ; T. DOBZHANSKY (1937). *Genetics and the Origin of Species*. New York : Columbia University Press.

un « groupe de populations naturelles isolées reproductivement les unes des autres ».[146] Lorsque les populations d'une espèce sont isolées géographiquement, leur capacité à se reproduire ensemble est toute théorique (et d'ailleurs 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 le concept d'espèce *phylogénétique* : « le plus petit groupe identifiable d'individus avec un pattern commun d'ancêtres et de descendants ».[147] Le nombre d'espèces phylogénétiques est bien supérieur au nombre d'espèces biologiques. Enfin, Van Valen[148] 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) : le concept *écologique* d'espèce est proche de celui de complexe d'espèces (ensemble d'espèces voisines échangeant des gènes[149]).

Le choix de la définition modifie considérablement sur la quantification de la richesse.[150] Des problèmes méthodologiques s'ajoutent aux problèmes conceptuels[151] : 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.[152]

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 (page 102) règle 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.

[146] E. Mayr (1970). *Populations, species, and evolution.* Cambridge, Massachussets : Belknap Press of Harvard University Press.

[147] J. Cracraft (1983). « Species Concepts and Speciation Analysis ». English. In : *Current Ornithology Volume 1*. Sous la dir. de R. F. Johnston. T. 1. Current Ornithology. Springer US, p. 159–187.

[148] L. Van Valen (1976). « Ecological Species, Multispecies, and Oaks ». In : *Taxon* 25.2/3, p. 233–239.

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

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

[151] Hey (2001). « The mind of the species problem », cf. note 8, p. 2.

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

2.4 Indice de Simpson

2.4.1 Définition

On note p_s la probabilité qu'un individu tiré au hasard appartienne à l'espèce s. L'indice de Simpson,[153] ou Gini-Simpson, est :

$$E = 1 - \sum_{s=1}^{S} p_s^2 \tag{2.45}$$

[153] E. H. Simpson (1949). « Measurement of diversity ». In : *Nature* 163.4148, p. 688.

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 :

$$D = \sum_{s=1}^{S} p_s^2 \tag{2.46}$$

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

$$\hat{E} = 1 - \sum_{s=1}^{s_{\neq 0}^n} \hat{p}_s^2 \tag{2.47}$$

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 `Simpson` du package *entropart* :

```
data(BCI)
# as.ProbaVector() transforme les abondances en probabilités
Ps <- as.ProbaVector(colSums(BCI))
Simpson(Ps)

##      None
## 0.9736755
```

Un historique de la définition de l'indice, de Gini[154] à Simpson, inspiré par Turing, est fourni par Ellerman.[155]

[154] C. Gini (1912). *Variabilità e mutabilità*. Bologna : C. Cuppini.

[155] D. Ellerman (2013). « An Introduction To Logical Entropy and Its Relation To Shannon Entropy ». In : *International Journal of Semantic Computing* 7.02, p. 121–145.

2.4.2 Biais d'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[156] :

$$\tilde{E} = \left(\frac{n}{n-1}\right)\left(1 - \sum_{s=1}^{s_{\neq 0}^n} \hat{p}_s^2\right) \tag{2.48}$$

[156] Good (1953). « The Population Frequency of Species and the Estimation of Population Parameters », cf. note 73, p. 26 ; R. Lande (1996). « Statistics and partitioning of species diversity, and similarity among multiple communities ». In : *Oikos* 76, p. 5–13.

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 :

$$\tilde{D} = \frac{\sum_{s=1}^{S} n_s\,(n_s - 1)}{n\,(n-1)} \tag{2.49}$$

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

La fonction `Simpson` de *entropart* accepte comme argument un vecteur d'abondances et propose par défaut la correction de Lande :

```
Simpson(colSums(BCI))

##     Lande
## 0.9737209
```

2.5 Indice de Shannon

2.5.1 Définition

L'indice de Shannon,[157] aussi appelé indice de Shannon-Weaver ou Shannon-Wiener, ou simplement *entropie* est dérivé de la théorie de l'information.

[157] C. E. Shannon (1948). « A Mathematical Theory of Communication ». In : *The Bell System Technical Journal* 27, p. 379–423, 623–656 ; C. E. Shannon et W. Weaver (1963). *The Mathematical Theory of Communication*. University of Illinois Press.

$$H = -\sum_{s=1}^{S} p_s \ln p_s \tag{2.50}$$

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 :

$$L = \frac{n!}{\prod_{i=1}^{S} (np_s)!} \tag{2.51}$$

Démonstration :

Le nombre de positions possibles dans la liste pour les individus de la première espèce est $\binom{n}{np_1}$. 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 (2.51).

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 :

$$\ln L = -n \sum_{s=1}^{S} p_s \ln p_s \tag{2.52}$$

On ne s'intéresse pas aux listes de plantes ne respectant pas les probabilités individuelles.

$H = (\ln L)/n$ est l'indice de Shannon. Ce résultat est connu sous le nom de formule de Brillouin.[158] À l'origine, Shannon a

[158] L. Brillouin (1962). *Science and information theory*. 2nd ed. Oxford : Academic Press. Chap. 351.

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.[159]

La formule (2.52) est celle de l'indice de Theil,[160] présenté en détail par Conceição et Ferreira,[161] 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.

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

$$\hat{H} = -\sum_{s=1}^{s^n_{\neq 0}} \hat{p}_s \ln \hat{p}_s \tag{2.53}$$

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 `Shannon` de *entropart* :

```
Ps <- as.ProbaVector(colSums(BCI))
Shannon(Ps)

##     None
## 4.270409
```

La distribution de l'estimateur est connue[162] mais elle est inutile en pratique à cause du biais d'estimation.

Bulmer[163] 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 :

$$\hat{H} = \Psi(\hat{\alpha} + 1) - \Psi(1) \tag{2.54}$$

$\Psi(\cdot)$ est la fonction digamma, et $\hat{\alpha}$ est l'estimateur de l'indice de Fisher (voir page 14) :

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

## [1] 4.148322
```

La sous-estimation est assez sévère sur cet exemple.

2.5.2 Biais d'estimation

Basharin[164] a montré que l'estimateur de l'indice de Shannon était biaisé parce que des espèces ne sont pas échantillonnées. Si S est

[159] E. C. Pielou (1966b). « The measurement of diversity in different types of biological collections ». In : *Journal of Theoretical Biology* 13.C, p. 131–144.

[160] H. Theil (1967). *Economics and Information Theory*. Chicago : Rand McNally et Company.

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

[162] K. Hutcheson (1970). « A Test for Comparing Diversities Based on the Shannon Formula ». In : *Journal of Theoretical Biology* 29, p. 151–154.

[163] M. G. Bulmer (1974). « On Fitting the Poisson Lognormal Distribution to Species-Abundance Data ». In : *Biometrics* 30.1, p. 101–110.

[164] G. P. Basharin (1959). « On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables ». In : *Theory of Probability and its Applications* 4.3, p. 333–336.

le nombre d'espèces réel et n le nombre d'individus échantillonnés, le biais est :

$$\mathbb{E}\left(\hat{H}\right) - H = -\frac{S-1}{2n} + O\left(n^{-2}\right) \tag{2.55}$$

$O\left(n^{-2}\right)$ 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[165] utilise l'information disponible, en sous-estimant le nombre d'espèces et donc l'entropie :

$$\tilde{H} = -\sum_{s=1}^{s^n_{\neq 0}} \hat{p}_s \ln \hat{p}_s + \frac{s^n_{\neq 0} - 1}{2n} \tag{2.56}$$

[165] G. A. Miller (1955). « Note on the bias of information estimates ». In : *Information Theory in Psychology : Problems and Methods*. Sous la dir. de H. Quastler. Free Press, p. 95–100.

Chao et Shen[166] établissent un estimateur moins biaisé à partir du taux de couverture de l'échantillonnage $\hat{C}$:

$$\tilde{H} = -\sum_{s=1}^{s^n_{\neq 0}} \frac{\hat{C}\hat{p}_s \ln\left(\hat{C}\hat{p}_s\right)}{1 - \left(1 - \hat{C}\hat{p}_s\right)^n} \tag{2.57}$$

[166] A. Chao et T.-J. Shen (2003). « Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample ». In : *Environmental and Ecological Statistics* 10.4, p. 429–443.

Multiplier les fréquences observées par le taux de couverture permet d'obtenir un estimateur non biaisé des probabilités conditionnellement aux espèces non observées.[167]

[167] J. Ashbridge et I. B. J. Goudie (2000). « Coverage-adjusted estimators for mark-recapture in heterogeneous populations ». In : *Communications in Statistics - Simulation and Computation* 29.4, p. 1215–1237.

Le terme au dénominateur est la correction de Horvitz et Thompson[168] : chaque terme de la somme est divisé par la probabilité d'observer au moins une fois l'espèce correspondante. Il tend vers 1 quand la taille de l'échantillon augmente.

[168] D. G. Horvitz et D. J. Thompson (1952). « A generalization of sampling without replacement from a finite universe ». In : *Journal of the American Statistical Association* 47.260, p. 663–685.

Beck et Schwanghart[169] montrent que la correction du biais est efficace, même à des niveaux de complétude de l'échantillonnage (voir page 35) très faibles. Vu *et al.*[170] étudient la vitesse de convergence de l'estimateur.

[169] Beck et Schwanghart (2010). « Comparing measures of species diversity from incomplete inventories : an update », cf. note 109, p. 34.

[170] V. Q. Vu et al. (2007). « Coverage-adjusted entropy estimation ». In : *Statistics in Medicine* 26.21, p. 4039–4060.

Zhang[171] définit l'indice de Simpson généralisé :

$$\zeta_{u,v} = \sum_{s=1}^{S} p_s^u (1 - p_s)^v \tag{2.58}$$

$\zeta_{u,v}$ 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 $\zeta_{1,v}$:

[171] Z. Zhang (2012). « Entropy Estimation in Turing's Perspective ». In : *Neural Computation* 24.5, p. 1368–1389.

$$H = \sum_{v=1}^{\infty} \frac{1}{v} \zeta_{1,v} \tag{2.59}$$

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 caculé en pratique par :

$$H_z = \sum_{v=1}^{n-1} \frac{1}{v} \left\{ \frac{n^{v+1}\,[n-(v+1)]!}{n!} \sum_{s=1}^{s_{\neq 0}^n} p_s \prod_{j=0}^{v-1} \left(1 - \hat{p}_s - \frac{j}{n}\right) \right\} \tag{2.60}$$

Zhang[172] montre que le biais de l'estimateur H_z est asymptotiquement normal et calcule sa variance. Zhang et Grabchak[173] améliorent l'estimateur en le complétant par un estimateur de son biais, mais les calculs deviennent excessivement complexes. Vinck *et al.*[174] appliquent la même démarche avec un estimateur bayesien 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[175] 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 moyenne calculée en ajoutant tous les relevés disponibles, est $\tilde{H}$.

Chao *et al.*[176] utilisent l'estimateur de la pente de la courbe de raréfaction, calculé précédemment[177] pour estimer la richesse spécifique, pour fournir un estimateur extrêmement performant :

$$\begin{aligned} \tilde{H} = &- \sum_{s=1}^{s_{\neq 0}^n} \frac{n_s}{n} \left(\Psi(n) - \Psi(n_s)\right) \\ &- \frac{s_1}{n}(1-A)^{1-n} \left(-\ln(A) - \sum_{r=1}^{n-1} \frac{1}{r}(1-A)^r\right) \end{aligned} \tag{2.61}$$

Où $\Psi(\cdot)$ 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.*[178] en font une revue). Le problème traité est la non-linéarité de l'indice de Shannon par rapport aux probabilités qui entraine un biais d'estimation. La fonction logarithme fournit un exemple simple : l'espérance de

[172] Z. Zhang (2013). « Asymptotic normality of an entropy estimator with exponentially decaying bias ». In : *IEEE Transactions on Information Theory* 59.1, p. 504–508.

[173] Z. Zhang et M. Grabchak (2013). « Bias adjustment for a nonparametric entropy estimator ». In : *Entropy* 15.6, p. 1999–2011.

[174] M. Vinck et al. (2012). « Estimation of the entropy based on its polynomial representation ». In : *Physical Review E* 85.5.

[175] E. C. Pielou (1966a). « Species-diversity and pattern-diversity in the study of ecological succession ». In : *Journal of Theoretical Biology* 10.2, p. 370–383.

[176] Chao et al. (2013). « Entropy and the species accumulation curve : a novel entropy estimator via discovery rates of new species », cf. note 29, p. 13.

[177] Chao et Jost (2012). « Coverage-based rarefaction and extrapolation : standardizing samples by completeness rather than size », cf. note 76, p. 26.

[178] J. A. Bonachela et al. (2008). « Entropy estimates of small data sets ». In : *Journal of Physics A : Mathematical and Theoretical* 41.202001, p. 1–9.

$\ln(p_s)$ 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.[179] L'indice de Shannon est concave (Figure 2.18 [180]) donc son estimateur (2.50) est biaisé négativement, même sans prendre en considération les espèces non observées.

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 2.19 [181]. 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.

Grassberger[182] a fourni la correction de référence :

$$\tilde{H} = -\sum_{s=1}^{s_{\neq 0}^n} \frac{n_s}{n} \left(\ln(n) - \Psi(n_s) - \frac{(-1)^{n_s}}{n_s + 1} \right) \tag{2.62}$$

$\Psi(n_s)$ est la fonction digamma.

Grassberger[183] l'a perfectionnée :

$$\tilde{H} = -\sum_{s=1}^{s_{\neq 0}^n} \frac{n_s}{n} \left(\Psi(n) - \Psi(n_s) - (-1)^{n_s} \int_0^1 \frac{t^{n_s - 1}}{1+t}\, dt \right) \tag{2.63}$$

Enfin, Schürmann[184] l'a généralisée pour définir une famille de corrections dépendant d'un paramètre ξ :

$$\tilde{H} = -\sum_{s=1}^{s_{\neq 0}^n} \frac{n_s}{n} \left(\Psi(n) - \Psi(n_s) - (-1)^{n_s} \int_0^{\frac{1}{\xi}-1} \frac{t^{n_s - 1}}{1+t}\, dt \right) \tag{2.64}$$

[179] J. L. W. V. Jensen (1906). « Sur les fonctions convexes et les inégalités entre les valeurs moyennes ». French. In : *Acta Mathematica* 30.1, p. 175–193.

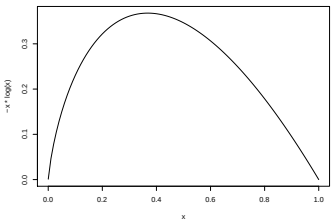


Figure 2.18 – courbe de $x \ln x$ entre 0 et 1.).

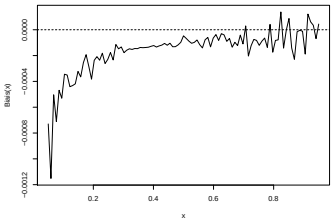


Figure 2.19 – Biais de $-\hat{p}_s \ln \hat{p}_s$ entre 0 et 1.

[180] Le code R nécessaire pour réaliser la figure est :

```
curve(-x * log(x), 1e-04, 1)
```

[181] Code R :

```
Biais.p <- function(p) {
    Ps <- rnorm(10000, p, 0.01)
    p * log(p) - mean(Ps * log(Ps))
}
Biais <- function(Ps) {
    # Applique Biais.p à chaque valeur de Ps
    sapply(Ps, Biais.p)
}
curve(Biais, 0.05, 0.95)
abline(h = 0, lty = 2)
```

[182] P. Grassberger (1988). « Finite sample corrections to entropy and dimension estimates ». In : *Physics Letters A* 128.6-7, p. 369–373.

[183] P. Grassberger (2003). « Entropy Estimates from Insufficient Samplings ». In : *ArXiv Physics e-prints* 0307138.version 2.

[184] T. Schürmann (2004). « Bias analysis in entropy estimation ». In : *Journal of Physics A : Mathematical and General* 37.27, p. L295–L301.

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 `Shannon` permet toutes ces corrections. Elle accepte comme argument un vecteur d'abondances (et non de probabilités) :

```
Shannon(colSums(BCI), Correction = "ChaoWangJost")

## ChaoWangJost
##     4.276193

Shannon(colSums(BCI), Correction = "Grassberger")

## Grassberger
##    4.276599

Shannon(colSums(BCI), Correction = "Grassberger2003")

## Grassberger2003
##        4.276491

Shannon(colSums(BCI), Correction = "Schurmann")

## Schurmann
##  4.276094

Shannon(colSums(BCI), Correction = "ZhangHz")

##  ZhangHz
## 4.272529
```

D'autres estimateurs peu utilisés en écologie sont disponibles dans le package *entropy*.[185] La contraction de Stein[186] consiste à estimer la distribution des probabilités d'occurence 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 :

$$\tilde{p}_s = \hat{\lambda}\frac{1}{\hat{S}} + \left(1 - \hat{\lambda}\right)\hat{p}_s \tag{2.65}$$

$$\hat{\lambda} = \frac{1 - \sum_{s=1}^{s^n_{\neq 0}} \left(\hat{p}_s\right)^2}{(n-1)\sum_{s=1}^{s^n_{\neq 0}} \left(\frac{1}{\hat{S}} - \hat{p}_s\right)^2} \tag{2.66}$$

L'entropie est ensuite simplement estimée par l'estimateur plug-in : $\tilde{H} = -\sum_{s=1}^{s^n_{\neq 0}} \tilde{p}_s \ln \tilde{p}_s$. 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

[185] J. Hausser et K. Strimmer (2009). « Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks ». In : *Journal of Machine Learning Research* 10, p. 1469–1484.

[186] W. James et C. Stein (1961). « Estimation with Quadratic Loss ». In : *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Sous la dir. de J. Neyman. T. Volume 1 : Berkeley, California : University of California Press, p. 361–379.

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 :

```
(S0 <- jackknife(AbdFreqCount(Ns))$Nhat - ncol(BCI))

##
## Your specified order is larger than that determined by the test,
## Therefore the order from the test is used.
## [1] 19

entropy.shrink(c(colSums(BCI), rep(0, S0)))

## 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 :

```
PsJS <- freqs.shrink(c(colSums(BCI), rep(0, S0)))

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

Shannon(as.ProbaVector(PsJS))

##     None
## 4.276543
```

Appliqué à des données de biodiversité aquatique, l'estimateur de James-Stein obtient de meilleurs résultats que celui de Chao et Shen[187] quand l'échantillonnage est réduit.

[187] D. Liu et al. (2015). « Entropy of hydrological systems under small samples : Uncertainty and variability ». In : *Journal of Hydrology* in press.

2.6 Indice de Hurlbert

2.6.1 Définition

L'indice de Hurlbert[188] est l'espérance du nombre d'espèces observées dans un échantillon de taille k choisie :

[188] S. H. Hurlbert (1971). « The Nonconcept of Species Diversity : A Critique and Alternative Parameters ». In : *Ecology* 52.4, p. 577–586.

$$ {}_kS = \sum_{s=1}^{S} \left[1 - (1 - p_s)^k \right] \tag{2.67} $$

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[189] (c'est-à-dire le nombre d'espèces équiprobables nécessaire pour obtenir la même diversité) notés ${}_kD$ à partir de la relation :

[189] Dauby et Hardy (2012). « Sampled-based estimation of diversity sensu stricto by transforming Hurlbert diversities into effective number of species », cf. note 77, p. 27.

$$ {}_kS = {}_kD \left[1 - \left(1 - \frac{1}{{}_kD} \right)^k \right] \tag{2.68} $$

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 de diversité, il existe une correspondance parfaite entre les deux mesures[190] : la connaissance de l'une suffit permet d'obtenir l'autre. Pour $k > 1$, on a[191] :

$$ {}_kS = k + \sum_{q=2}^{k} \binom{k}{q} (-1)^{q+1} ({}^qD)^{1-q} \tag{2.69} $$

On a vu que ${}^0D$ égale ${}_\infty D$, donc ${}_\infty S = {}^0D$ (équation 2.68).

Inversement :

$$ ({}^qD)^{1-q} = q + \sum_{k=2}^{q} \binom{q}{k} (-1)^{k+1} {}_kS \tag{2.70} $$

Pour $q = 1$, la relation est[192] :

$$ {}^1H = -1 + \sum_{k=2}^{\infty} \frac{{}_kS}{k(k-1)} \tag{2.71} $$

2.6.2 Biais d'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) :

$$ {}_k\tilde{S} = \sum_{s=1}^{s^n_{\neq 0}} \left[1 - \binom{n - n_s}{k} \Big/ \binom{n}{k} \right] \tag{2.72} $$

Dauby et Hardy[193] 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. Smith et Grassle[194] ont calculé sa variance.

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

```
# Indice de Hurlbert (probabilités)
Hurlbert(as.ProbaVector(colSums(BCI)), k = 2)

##   Biased
## 1.973676

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

## Unbiased
## 1.973721

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

[190] A. Chao et al. (2014a). « Rarefaction and extrapolation with Hill numbers : A framework for sampling and estimation in species diversity studies ». In : *Ecological Monographs* 84.1, p. 45–67, Annexe S2.

[191] T. Leinster et C. Cobbold (2012). « Measuring diversity : the importance of species similarity ». In : *Ecology* 93.3, p. 477–489.

[192] C. X. Mao (2007). « Estimating Species Accumulation Curves and Diversity Indices ». In : *Statistica Sinica* 17, p. 761–774.

[193] Dauby et Hardy (2012). « Sampled-based estimation of diversity sensu stricto by transforming Hurlbert diversities into effective number of species », cf. note 77, p. 27.

[194] W. Smith et J. F. Grassle (1977). « Sampling properties of a family of diversity measures ». In : *Biometrics* 33.2, p. 283–292.

```
## Unbiased
## 38.05308
```

Chapitre 3

Entropie

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.

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

3.1 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]

[2] H. T. Davis (1941). *The theory of econometrics*. Bloomington, Indiana : The Principia Press.

[3] Theil (1967). *Economics and Information Theory*, cf. note 160, p. 48.

[4] Shannon (1948). « A Mathematical Theory of Communication », cf. note 157, p. 47 ; Shannon et Weaver (1963). *The Mathematical Theory of Communication*, cf. note 157, p. 47.

[5] E. Maasoumi (1993). « A compendium to information theory in economics and econometrics ». In : *Econometric Reviews* 12.2, p. 137–181.

Considérons une expérience dont les résultats possibles sont $\{r_1, r_2, \ldots, r_S\}$. La probabilité d'obtenir r_s est p_s, et $\mathcal{P} = \{p_1, p_2, \ldots, p_S\}$. 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 $I(0) = +\infty$ (ou éventuellement une valeur strictement positive finie) à $I(1) = 0$. Chaque valeur observée dans l'échantillon apporte une certaine quantité d'information, dont la somme est l'information de l'échantillon. Patil et Taillie[6] appellent l'information « rareté ».

[6] G. P. Patil et C. Taillie (1982). « Diversity as a concept and its measurement ». In : *Journal of the American Statistical Association* 77.379, p. 548–561.

La quantité d'information attendue de l'expérience est $\sum_{s=1}^{S} p_s I(p_s) = H(\mathcal{P})$. Si on choisit $I(p_s) = -\ln(p_s)$, $H(\mathcal{P})$ est l'indice de Shannon, mais bien d'autres formes de $I(p_s)$ sont possibles. $H(\mathcal{P})$ 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 $\mathcal{P}$ est la distribution des probabilité des espèces dans une communauté, Patil et Taillie[7] montrent que :

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

[7] Patil et Taillie (1982). « Diversity as a concept and its measurement », cf. note 6, p. 57.

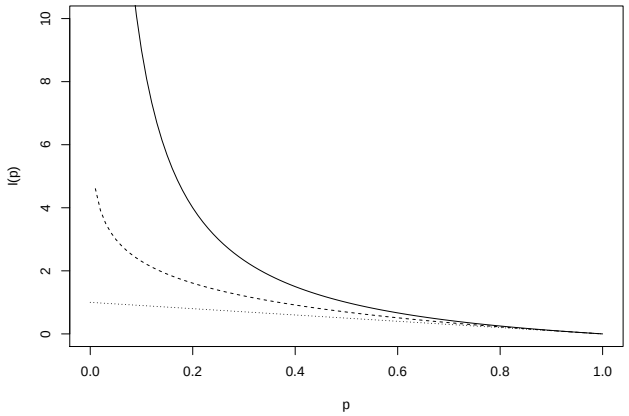


Figure 3.1 – Fonctions d'information utilisées dans le nombre d'espèces (trait plein), l'indice de Shannon (pointillés longs) et l'indice de Simpson (pointillés). L'information apportée par l'observation d'espèces rares décroît du nombre d'espèces à l'indice de Simpson.

Ces trois fonctions d'information sont représentées en Figure 3.1 [8].

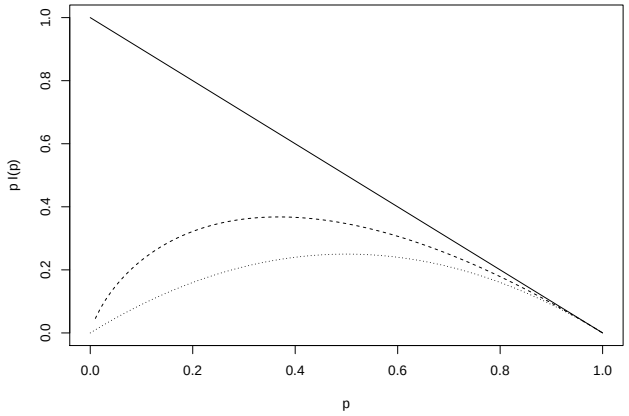


Figure 3.2 – Valeur de $p_s I(p_s)$ dans le nombre d'espèces (trait plein), l'indice de Shannon (pointillés longs) et l'indice de Simpson (pointillés). Les espèces rares contribuent peu, sauf pour le nombre d'espèces

La contribution de chaque espèce à la valeur totale de l'entropie est représentée Figure 3.2 [9].

[8] Le code R nécessaire pour réaliser la figure est :

```
curve((1 - x)/x, 0, 1, lty = 1, xlab = "p", ylab = "I(p)", ylim = c(0,
    10))
curve(-log(x), 0, 1, lty = 2, add = TRUE)
curve(1 - x, 0, 1, lty = 3, add = TRUE)
```

[9] Code R :

3.2 Entropie relative

Considérons maintenant les probabilités q_s formant l'ensemble $\mathcal{Q}$ 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, $\sum_{s=1}^{S} q_s I(q_s, p_s) = H(\mathcal{Q}, \mathcal{P})$, 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 $\mathcal{P}$ et $\mathcal{Q}$ 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 $\hat{H}(\mathcal{Q}, \mathcal{P})$ seront nécessaires.

Quelques formes possibles de $H(\mathcal{Q}, \mathcal{P})$ sont :

- La divergence de Kullback-Leibler[10] connue par les économistes comme l'indice de dissimilarité de Theil[11] :

$$T = \sum_{s=1}^{S} q_s \ln \frac{q_s}{p_s} \tag{3.1}$$

- Sa proche parente, appelée parfois deuxième mesure de Theil,[12] qui inverse simplement les rôles de p et q :

$$L = \sum_{s=1}^{S} p_s \ln \frac{p_s}{q_s} \tag{3.2}$$

[10] S. Kullback et R. A. Leibler (1951). « On Information and Sufficiency ». In : *The Annals of Mathematical Statistics* 22.1, p. 79–86.

[11] Theil (1967). *Economics and Information Theory*, cf. note 160, p. 48.

[12] Conceição et Ferreira (2000). *The Young Person's Guide to the Theil Index : Suggesting Intuitive Interpretations and Exploring Analytical Applications*, cf. note 161, p. 48.

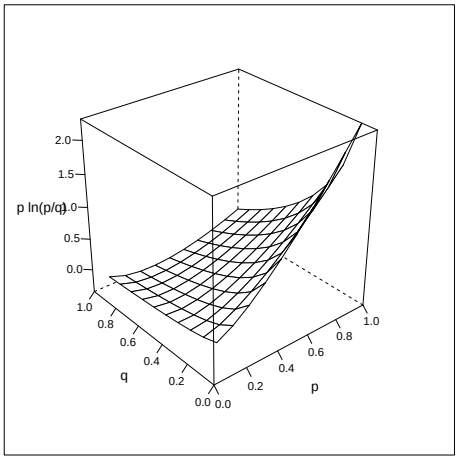


Figure 3.3 – Valeur de $q_s \ln(q_s/p_s)$ en fonction de p_s et q_s. L'entropie de Kullback-Leibler est la somme de cette valeur pour toutes les espèces

```
curve(1 - x, 0, 1, lty = 1, xlab = "p", ylab = "p I(p)")
curve(-x * log(x), 0, 1, lty = 2, add = TRUE)
curve(x * (1 - x), 0, 1, lty = 3, add = TRUE)
```

L'entropie relative est essentielle pour la définition de la diversité β présentée dans le chapitre 5.3. 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.[13] Les valeurs de chaque terme de la divergence sont représentées Figure 3.3 [14].

[13] É. Marcon et al. (2012). « The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity ». In : *Oikos* 121.4, p. 516–522.

3.3 Entropie et biodiversité : historique

MacArthur[15] est le premier à avoir introduit la théorie de l'information en écologie.[16] 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.

[15] R. H. MacArthur (1955). « Fluctuations of Animal Populations and a Measure of Community Stability ». In : *Ecology* 36.3, p. 533–536.

[16] R. E. Ulanowicz (2001). « Information theory in ecology ». In : *Computers & Chemistry* 25.4, p. 393–399.

Les premiers travaux consistant à généraliser l'indice de Shannon sont dus à Rényi.[17] L'entropie d'ordre q de Rényi est :

$$ {}^qR = \frac{1}{1 - q \ln \sum_{q=1}^{S} p_s^q} \tag{3.3} $$

[17] A. Rényi (1961). « On Measures of Entropy and Information ». In : *4th Berkeley Symposium on Mathematical Statistics and Probability*. Sous la dir. de J. Neyman. T. 1. Berkeley, USA : University of California Press, p. 547–561.

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

- 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 ${}^qR$ respecte les 3 axiomes.

Patil et Taillie[18] 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 $R(\mathcal{P})$ soit concave. Dans la littérature économique sur les inégalités, cette propriété est connue sous le nom de Pigou-Dalton.[19]

[18] Patil et Taillie (1982). « Diversity as a concept and its measurement », cf. note 6, p. 57.

[19] H. Dalton (1920). « The measurement of the inequality of incomes ». In : *The Economic Journal* 30.119, p. 348–361.

Hill[20] transforme l'entropie de Rényi en *nombres de Hill*, qui en sont simplement l'exponentielle :

[20] M. O. Hill (1973). « Diversity and Evenness : A Unifying Notation and Its Consequences ». In : *Ecology* 54.2, p. 427–432.

[14] Le code R nécessaire pour réaliser la figure est :

```
grid <- expand.grid(x=seq(0, 1, .1) , y=seq(0, 1, .1))
grid$z <- grid$x*log(grid$x/grid$y)
print(wireframe(z ~ x * y, grid, scales=list(arrows = FALSE),
                xlab="p", ylab="q", zlab="p ln(p/q)"))
```

$$ {}^qD = \left(\sum_{s=1}^{S} p_s^q \right)^{\frac{1}{1-q}} \tag{3.4} $$

Le souci de Hill était de rendre les indices de diversité intelligibles après l'article remarqué de Hurlbert[21] intitulé « le nonconcept 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.[22]

Ces résultats avaient déjà été obtenus avec une autre approche par MacArthur[23] et repris par Adelman[24] dans la littérature économique.

Les nombres de Hill sont des « nombres effectifs » ou « nombres équivalents ». Le concept a été défini rigoureusement par Gregorius,[25] d'après Wright[26] (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[27] 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 au chapitre 5. 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.

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 :

[21] Hurlbert (1971). « The Nonconcept of Species Diversity : A Critique and Alternative Parameters », cf. note 188, p. 53.

[22] J. A. Stoddart (1983). « A genotypic diversity measure ». In : *Journal of Heredity* 74, p. 489–490.

[23] R. H. MacArthur (1965). « Patterns of species diversity ». In : *Biological Reviews* 40.4, p. 510–533.

[24] M. A. Adelman (1969). « Comment on the "H" Concentration Measure as a Numbers-Equivalent ». In : *The Review of Economics and Statistics* 51.1, p. 99–101.

[25] H.-R. Gregorius (1991). « On the concept of effective number. » In : *Theoretical population biology* 40.2, p. 269–83.

[26] S. Wright (1931). « Evolution in Mendelian Populations ». In : *Genetics* 16.2, p. 97–159.

[27] H.-R. Gregorius (2014). « Partitioning of diversity : the "within communities" component ». In : *Web Ecology* 14, p. 51–60.

$$D = \frac{1}{I^{-1}\left(\sum_{s=1}^{S} p_s I(p_s)\right)} \tag{3.5}$$

3.4 Entropie HCDT

Tsallis[28] propose une classe de mesures appelée entropie généralisée, définie par Havrda et Charvát[29] pour la première fois et redécouverte plusieurs fois, notamment par Daróczy,[30] d'où son nom *entropie HCDT* (voir Mendes *et al.*,[31] page 451, pour un historique complet) :

$$^{q}H = \frac{1}{q-1}\left(1 - \sum_{s=1}^{S} p_s^q\right) \tag{3.6}$$

Tsallis a montré que les indices de Simpson et de Shannon étaient des cas particuliers d'entropie généralisée, retrouvant, sans faire le rapprochement,[32] la définition d'un indice de diversité de Patil et Taillie.[33]

Ces résultats ont été complétés par d'autres et repris en écologie par Keylock[34] et Jost.[35] 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 seulement pour les cas particuliers (nombre d'espèces, Shannon, Simpson). Enfin, sa décomposition sera présentée en détail dans le chapitre 5.3.

3.4.1 Logarithmes déformés

L'écriture de l'entropie HCDT est largement simplifiée en introduisant le formalisme des logarithmes déformés.[36] Le logarithme d'ordre q est défini par :

$$\ln_q x = \frac{x^{1-q} - 1}{1-q} \tag{3.7}$$

Le logarithme déformé converge vers le logarithme naturel quand $q \to 1$ (Figure 3.4 [37]).

Sa fonction inverse est l'exponentielle d'ordre q :

$$e_q^x = [1 + (1-q)\,x]^{\frac{1}{1-q}} \tag{3.8}$$

[28] C. Tsallis (1988). « Possible generalization of Boltzmann-Gibbs statistics ». In : *Journal of Statistical Physics* 52.1, p. 479–487.

[29] J. Havrda et F. Charvát (1967). « Quantification method of classification processes. Concept of structural a-entropy ». In : *Kybernetika* 3.1, p. 30–35.

[30] Z. Daróczy (1970). « Generalized information functions ». In : *Information and Control* 16.1, p. 36–51.

[31] R. S. Mendes et al. (2008). « A unified index to measure ecological diversity and species rarity ». In : *Ecography* 31.4, p. 450–456.

[32] C. Ricotta (2005d). « On parametric diversity indices in ecology : A historical note ». In : *Community Ecology* 6.2, p. 241–244.

[33] Patil et Taillie (1982). « Diversity as a concept and its measurement », cf. note 6, p. 57.

[34] C. J. Keylock (2005). « Simpson diversity and the Shannon-Wiener index as special cases of a generalized entropy ». In : *Oikos* 109.1, p. 203–207.

[35] L. Jost (2006). « Entropy and diversity ». In : *Oikos* 113.2, p. 363–375 ; Jost (2007). « Partitioning diversity into independent alpha and beta components », cf. note 42, p. 17.

[36] C. Tsallis (1994). « What are the numbers that experiments provide ? » In : *Química Nova* 17.6, p. 468–471.

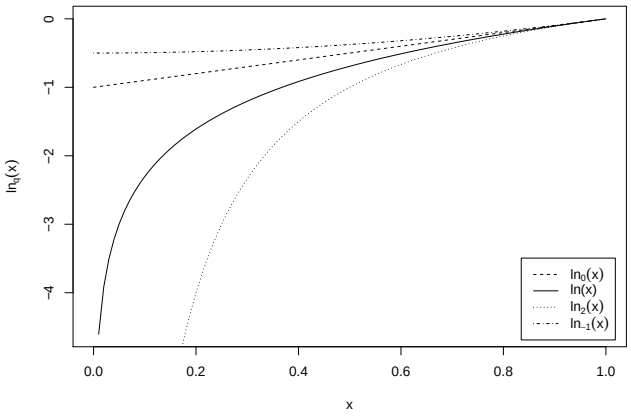


FIGURE 3.4 – Valeur du logarithme d'ordre q de probabilités entre 0 et 1 pour différentes valeurs de q : $q = 0$ (pointillés longs), la courbe est une droite ; $q = 1$ (trait plein) : logarithme naturel ; $q = 2$ (pointillés courts) : la courbe a la même forme que le logarithme naturel pour les valeurs positives de q ; $q = -1$ (pointillés alternés) : la courbe est convexe pour les valeurs négatives de q.

Enfin, le logarithme déformé est subadditif :

$$\ln_q(xy) = \ln_q x + \ln_q y - (q-1)(\ln_q x)(\ln_q y) \tag{3.9}$$

Ses propriétés sont les suivantes :

$$\ln_q \frac{1}{x} = -x^{q-1} \ln_q x \tag{3.10}$$

$$\ln_q(xy) = \ln_q x + x^{1-q} \ln_q y \tag{3.11}$$

$$\ln_q\left(\frac{x}{y}\right) = \ln_q x - \left(\frac{x}{y}\right)^{1-q} \ln_q y \tag{3.12}$$

Et :

$$e_q^{x+y} = e_q^x e_q^{\frac{y}{1+(1-q)x}} \tag{3.13}$$

Si $q > 1$, $\lim_{x\to+\infty}(\ln_q x) = 1/(q-1)$, donc e_q^x n'est pas défini pour $x > 1/(q-1)$.

Ces fonctions sont implémentées dans le package *entropart* : `lnq(x, q)` et `expq(x, q)`.

L'entropie d'ordre q s'écrit :

[37] Le code R nécessaire pour réaliser la figure est :

```
curve(log(x), 0, 1, lty = 1, ylab = expression(ln[q](x)))
curve(lnq(x, 0), 0, 1, lty = 2, add = TRUE)
curve(lnq(x, 2), 0, 1, lty = 3, add = TRUE)
curve(lnq(x, -1), 0, 1, lty = 4, add = TRUE)
legend("bottomright", legend = c(expression(ln[0](x)), "ln(x)",
    expression(ln[2](x)), expression(ln[-1](x))), lty = c(2,
    1, 3, 4), inset = 0.02)
```

$$ {}^{q}H = \frac{1}{q-1}\left(1 - \sum_{s=1}^{S} p_s^q\right) = -\sum_s p_s^q \ln_q p_s = \sum_s p_s \ln_q \frac{1}{p_s} \quad (3.14) $$

Ces trois formes sont équivalentes mais les deux dernières s'interprètent comme une généralisation de l'entropie de Shannon.[38]

[38] É. Marcon et al. (2014a). « Generalization of the Partitioning of Shannon Diversity ». In : *Plos One* 9.3, e90289.

Le calcul de ${}^{q}H$ peut se faire avec la fonction `Tsallis` de la librairie *entropart* :

```
Ps <- as.ProbaVector(colSums(BCI))
q <- 1.5
Tsallis(Ps, q)

##     None
## 1.715954
```

3.4.2 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.

$$ {}^{q}H = \ln_q {}^{q}D \quad (3.15) $$

$$ {}^{q}D = e_q^{{}^{q}H} \quad (3.16) $$

L'entropie est utile pour les calculs : la correction des biais d'estimation notamment. Les nombres de Hill, ou *nombres équivalent d'espèces* ou *nombres d'espèces effectives* permettent une appréhension plus intuitive de la notion de biodiversité.[39] En raison de leurs propriétés, notamment de décomposition (voir le chapitre 5.3), Jost[40] les appelle « vraie diversité ». Hoffmann et Hoffmann[41] critiquent cette définition totalitaire[42] et fournissent une revue historique plus lointaine sur les origines de ces mesures.

[39] Jost (2006). « Entropy and diversity », cf. note 35, p. 62.

[40] Jost (2007). « Partitioning diversity into independent alpha and beta components », cf. note 42, p. 17.

[41] S. Hoffmann et A. Hoffmann (2008). « Is there a "true" diversity ? » In : *Ecological Economics* 65.2, p. 213–215.

[42] Jost (L. Jost [2009b]. « Mismeasuring biological diversity : Response to Hoffmann and Hoffmann (2008) ». In : *Ecological Economics* 68, p. 925–928) reconnaît qu'un autre terme aurait pu être choisi (« diversité neutre » ou « diversité mathématique » par exemple).

Dauby et Hardy[43] écrivent « diversité au sens strict » ; Gregorius[44] « diversité explicite ». Quoi qu'il en soit, les nombres de Hill respectent le principe de réplication (voir Chao *et al.*,[45] 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 $I \times D$.

[43] Dauby et Hardy (2012). « Sampled-based estimation of diversity sensu stricto by transforming Hurlbert diversities into effective number of species », cf. note 77, p. 27.

[44] H.-R. Gregorius (2010). « Linking Diversity and Differentiation ». In : *Diversity* 2.3, p. 370–394.

[45] Chao et al. (2010). « Phylogenetic diversity measures based on Hill numbers », cf. note 82, p. 27.

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 ;

- ${}^0D$ est le nombre d'espèces (alors que ${}^0H$ 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 important, et sa résolution a généré une littérature en soi ;
- ${}^1D$ 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[46] 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[47] qui est la proportion de la communauté représentée par l'espèce la plus abondante : $d = \max(p_S)$. Toutes les autres espèces sont ignorées.

[46] Hill (1973). « Diversity and Evenness : A Unifying Notation and Its Consequences », cf. note 20, p. 60.

[47] W. H. Berger et F. L. Parker (1970). « Diversity of planktonic foraminifera in deep-sea sediments ». In : *Science* 168.3937, p. 1345–1347.

Le calcul de ${}^qD$ peut se faire avec la fonction `Diversity` de la librairie *entropart* :

```
Ps <- as.ProbaVector(colSums(BCI))
q <- 1.5
Diversity(Ps, q)

##     None
## 49.57724
```

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,[48] 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 (chapitre 5) comme la moyenne des entropies de plusieurs communautés, ou l'entropie phylogénétique (chapitre 4) 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.

[48] R. Gadagkar (1989). « An undesirable property of Hill ' s diversity index N2 ». In : *Oecologia* 80, p. 140–141.

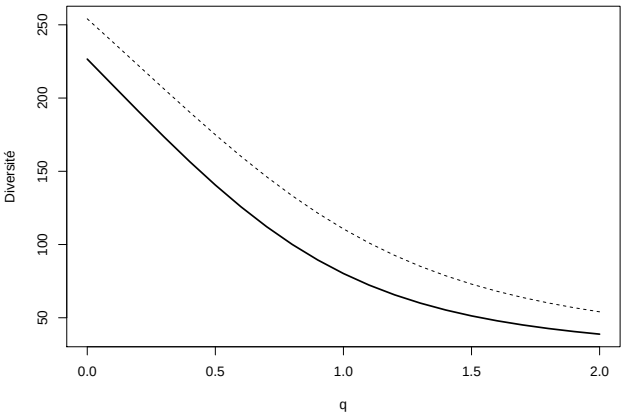


FIGURE 3.5 – Profil de diversité calculé pour deux parcelles de Paracou (Parcelle 6 : trait plein et Parcelle 18 : trait pointillé). La correction du biais d'estimation est celle de Chao et Jost.

3.4.3 Profils de diversité

Leinster et Cobbold,[49] après Hill,[50] Patil et Taillie,[51] Tothmeresz[52] et Kindt *et al.*,[53] recommandent de tracer des profils de diversité, c'est-à-dire la valeur de la diversité ${}^qD$ en fonction de l'ordre q (Figure 3.5 [54]) 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.[55]

Lande *et al.*[56] montrent que si la diversité de Simpson et la richesse de deux communautés n'ont pas le même ordre, alors les courbes d'accumulation du nombre d'espèces en fonction du nombre d'individus échantillonnés se croisent aussi.

Pallmann *et al.*[57] ont développé un test statistique pour comparer la diversité de deux communautés pour plusieurs valeurs de q simultanément.

Liu *et al.*[58] 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 (page 2.6), et inversement. Ils montrent que les communautés séparables selon un troisième type de profil, celui de la queue de distribution,[59] 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

[49] LEINSTER et COBBOLD (2012). « Measuring diversity : the importance of species similarity », cf. note 191, p. 54.

[50] HILL (1973). « Diversity and Evenness : A Unifying Notation and Its Consequences », cf. note 20, p. 60.

[51] PATIL et TAILLIE (1982). « Diversity as a concept and its measurement », cf. note 6, p. 57.

[52] B. TOTHMERESZ (1995). « Comparison of different methods for diversity ordering ». In : *Journal of Vegetation Science* 6.2, p. 283–290.

[53] R. KINDT et al. (2006). « Tree diversity in western Kenya : Using profiles to characterise richness and evenness ». In : *Biodiversity and Conservation* 15.4, p. 1253–1270.

[55] TOTHMERESZ (1995). Cf. note 52.

[56] R. LANDE et al. (2000). « When species accumulation curves intersect : implications for ranking diversity using small samples ». In : *Oikos* 89.3, p. 601–605.

[57] P. PALLMANN et al. (2012). « Assessing group differences in biodiversity by simultaneously testing a user-defined selection of diversity indices ». In : *Molecular Ecology Resources* 12.6, p. 1068–1078.

[58] C. LIU et al. (2006). « Unifying and distinguishing diversity ordering methods for comparing communities ». In : *Population Ecology* 49.2, p. 89–100.

[59] PATIL et TAILLIE (1982). « Diversity as a concept and its measurement », cf. note 6, p. 57.

[54] Code R :

```
q.seq <- seq(0, 2, 0.1)
P6D <- CommunityProfile(Diversity, Paracou618.MC$Nsi[, 1], q.seq)
P18D <- CommunityProfile(Diversity, Paracou618.MC$Nsi[, 2], q.seq)
plot(P6D, xlab = "q", ylab = "Diversité", main = "", ylim = c(min(P6D$y),
    max(P18D$y)))
lines(P18D, lty = 2)
```

une espèce plus rare que l'espèce en abscisse (Figure 3.6 [60]). Les coordonnées des points du profil sont définies par :

$$y(x) = \sum_{s=x+1}^{S} p_{[s]},\ x \in \{0, 1, \ldots, S\} \tag{3.17}$$

$p_{[s]}$ est la probabilité de l'espèce s ; les espèces sont classées par probabilité décroissante.

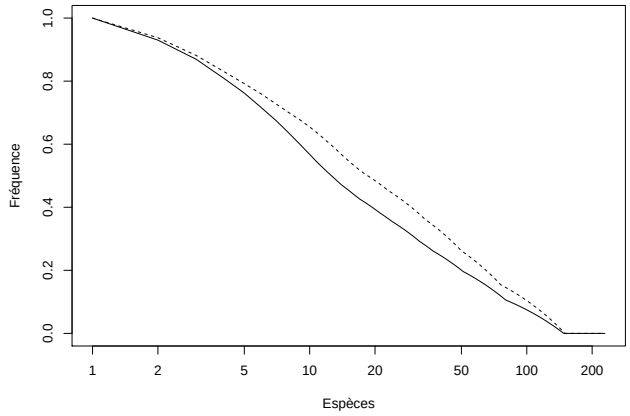


FIGURE 3.6 – Profil de queue de distribution calculé pour les deux parcelles de Paracou (Parcelle 6 : trait plein et Parcelle 18 : trait pointillé). En abscisse : rang de l'espèce dans le classement de la plus fréquente à la plus rare ; en ordonnée : probabilité qu'un individu de la communauté appartienne à une espèce plus rare.

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.

3.4.4 Biais d'estimation

Zhang et Grabchak

Zhang et Grabchak[61] montrent que toutes les mesures d'entropie usuelles sont des combinaisons linéaires de $\zeta_{u,v}$ (voir page 49). À partir des estimateurs de $\zeta_{u,v}$,[62] l'estimateur suivant est proposé pour $h_q = \sum_{s=1}^{S} p_s^q$:

$$\tilde{h}_q = 1 + \sum_{s=1}^{s_{\neq 0}^n} \frac{n_s}{n} \sum_{v=1}^{n-n_s} \left[\prod_{i=1}^{v} \frac{i-q}{i} \right] \left[\prod_{j=1}^{v} \left(1 - \frac{n_s - 1}{n-j} \right) \right] \tag{3.18}$$

[60] Code R :

```
# Elimination des espèces absentes des deux parcelles
Psi <- Paracou618.MC$Psi[Paracou618.MC$Ps > 0, ]
P6Ps <- sort(Psi[, 1])
P18Ps <- sort(Psi[, 2])
plot(rev(cumsum(P6Ps)), type = "l", log = "x", xlab = "Espèces",
    ylab = "Fréquence")
lines(rev(cumsum(P18Ps)), lty = 2)
```

[61] Z. ZHANG et M. GRABCHAK (2014a). « Entropic Representation and Estimation of Diversity Indices ». In : *arXiv* 1403.3031.version 2.

[62] ZHANG et ZHOU (2010). « Re-parameterization of multinomial distributions and diversity indices », cf. note 30, p. 13.

$\tilde{h}_q$ peut ensuite être utilisé pour l'estimation de l'entropie HCDT :

$$ {}^q\tilde{H} = \frac{1 - \tilde{h}_q}{q - 1} \tag{3.19} $$

L'estimateur est asymptotiquement normal et son biais décroit exponentiellement vite.

$\tilde{h}_q$ n'est défini que pour $q \neq 1$. L'estimateur corrigé de ${}^1H$ est :

$$ {}^1\tilde{H} = \sum_{s=1}^{s_{\neq 0}^n} \frac{n_s}{n} \sum_{v=1}^{n-n_s} \frac{1}{v} \left[\prod_{j=1}^{v} 1 - \frac{n_s - 1}{n - j} \right] \tag{3.20} $$

Cet estimateur est différent de H_z.[63] Il a été également obtenu par d'autres auteurs[64,65] avec d'autres approches.

Dans *entropart*, utiliser `Tsallis` et `Diversity` :

```
Tsallis(Paracou618.MC$Ns, q = 1, Correction = "ZhangGrabchak")

## ZhangGrabchak
##      4.846377

Diversity(Paracou618.MC$Ns, q = 1, Correction = "ZhangGrabchak")

## ZhangGrabchak
##      127.2784
```

[63] Zhang (2012). « Entropy Estimation in Turing's Perspective », cf. note 171, p. 49.

[64] Mao (2007). « Estimating Species Accumulation Curves and Diversity Indices », cf. note 192, p. 54 ; Vinck et al. (2012). « Estimation of the entropy based on its polynomial representation », cf. note 174, p. 50.

[65] Chao et al. (2013). « Entropy and the species accumulation curve : a novel entropy estimator via discovery rates of new species », cf. note 29, p. 13, Annexe S1.

Chao et Jost

Chao et Jost[66] complètent cet estimateur par une estimation du biais résiduel, selon la même démarche que pour l'entropie de Shannon[67] (voir page 50) :

$$ {}^q\tilde{H} = \frac{1}{q-1} \left[1 - \tilde{h}_q - \frac{s_1}{n}(1-A)^{1-n} \left(A^{q-1} - \sum_{r=0}^{n-1} \binom{q-1}{r} (A-1)^r \right) \right] \tag{3.21} $$

La formule du nombre de combinaisons est généralisée aux valeurs de $q-1$ non entières, et vaut par convention 1 pour $r = 0$. 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.

Cet estimateur est le plus performant. Quand $q \to 1$, il tend vers l'estimateur de l'entropie de Shannon de Chao, Wang et Jost,[68] équation 2.61 page 50. Quand $q = 0$, l'estimateur se réduit à Chao1. Pour toutes les valeurs entières de q supérieures ou égales à 2, il se réduit à l'estimateur (3.21), identique à l'estimateur de Lande pour $q = 2$.

[66] A. Chao et L. Jost (2015). « Estimating diversity and entropy profiles via discovery rates of new species ». In : *Methods in Ecology and Evolution* 6.8, p. 873–882.

[67] Chao et al. (2013). « Entropy and the species accumulation curve : a novel entropy estimator via discovery rates of new species », cf. note 29, p. 13.

[68] Ibid.

L'estimateur est obtenu à partir de celui du taux de couverture extrapolé (page 26) fourni par Chao et Jost[69] sans le détail de la démonstration présenté ici. L'espérance conditionnelle du taux de couverture dans un échantillon de taille $n+m$ compte-tenu des effectifs observés, notés $\{n_s\}$, est :

[69] Chao et Jost (2012). « Coverage-based rarefaction and extrapolation : standardizing samples by completeness rather than size », cf. note 76, p. 26, annexe E.

$$\mathbb{E}\left(C^{n+m}|\{n_s\}\right) = C^n + \sum_s p_s[1-(1-p_s)^m]\mathbf{1}(n_s=0) \quad (3.22)$$

Le taux de couverture C^n est augmenté par l'échantillonnage de m individus supplémentaires si les espèces non échantillonnées (vérifiant $\mathbf{1}(n_s=0)$) sont découvertes (la probabilité de le faire est $1-(1-p_s)^m$). $\hat{C}^n + \sum_s p_s \mathbf{1}(n_s=0)$ vaut par définition 1. Le deuxième terme de la somme est estimé en posant $\hat{p}_s = \hat{\alpha}_0$ pour toutes les espèces non observées, ce qui constitue la même approximation que pour l'estimation de la relation de Good-Turing, équation (2.15) page 26. Alors, $\sum_s p_s$ est estimé par $\hat{\alpha}_0 s_0^n = \frac{s_1^n}{n}(1-\hat{\alpha}_1)$ et $(1-p_s)^m$ par $(1-\hat{\alpha}_0)^m$. Les deux estimateurs $\hat{\alpha}_1$ et $\hat{\alpha}_0$ sont égaux à A si le nombre d'espèces non échantillonnées est estimé par l'estimateur Chao1. On obtient :

$$\hat{C}^{n+m} = 1 - \frac{s_1^n}{n}\left[\frac{(n-1)s_1^n}{(n-1)s_1^n + 2s_2^n}\right]^{n+m} \quad (3.23)$$

Autres méthodes

Deux autres approches sont possibles[70] :

- Celle de Chao et Shen,[71] étendue au-delà de l'entropie de Shannon qui prend en compte les espèces non observées ;
- Celle de Grassberger[72] qui prend en compte la non-linéarité de l'estimateur.

[70] Marcon et al. (2014a). « Generalization of the Partitioning of Shannon Diversity », cf. note 38, p. 64.

[71] Chao et Shen (2003). « Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample », cf. note 166, p. 49.

[72] Grassberger (1988). « Finite sample corrections to entropy and dimension estimates », cf. note 182, p. 51.

Le domaine de validité des deux corrections est différent. Quand q est petit, les espèces rares non observées ont une grande importance alors que ${}^qH$ est presque linéaire : la correction de Grassberger disparaît quand $q=0$. Quand q est grand (au-delà de 2), l'influence des espèces rares est faible, la correction de Chao et Shen devient négligeable alors que la non-linéarité augmente. Un compromis raisonnable consiste à utiliser le plus grand des deux estimateurs.

L'estimateur de Chao et Shen est :

$${}^q\tilde{H} = -\sum_{s=1}^{s_{\neq 0}^n} \frac{\hat{C}\hat{p}_s \ln_q\left(\hat{C}\hat{p}_s\right)}{1-\left(1-\hat{C}\hat{p}_s\right)^n} \quad (3.24)$$

Celui de Grassberger est :

$${}^q\tilde{H} = \frac{1-\sum_{s=1}^{s_{\neq 0}^n}\widetilde{p_s^q}}{q-1} \quad (3.25)$$

Où :

$$\widetilde{p_s^q} = {n_s}^{-q} \left(\frac{\Gamma(n_s + 1)}{\Gamma(n_s - q + 1)} + \frac{(-1)^n \Gamma(1 + q) \sin \pi q}{\pi (n + 1)} \right) \tag{3.26}$$

$\Gamma(\cdot)$ est la fonction gamma. L'estimateur est linéaire par rapport à $\widetilde{p_s^q}$: le problème original est donc résolu par la construction de l'estimateur de p_s^q, mais celui-ci est biaisé également.

L'estimateur de Chao et Shen peut être amélioré[73] en estimant mieux les probabilités. Chao *et al.*[74] ont développé un estimateur de la distribution des probabilités à partir de l'estimation du taux de couverture généralisé. $\hat{C}\hat{p}_s$ peut être remplacé par cet estimateur pour réduire le biais de l'estimateur de Chao et Shen.

Enfin, la distribution des espèces non échantillonnées peut être ajoutée en estimant leur nombre et en choisissant une loi. Chao *et al.*[75] ont utilisé l'estimateur Chao1 et une distribution géométrique. L'estimateur jackknife[76] est plus versatile parce que son ordre peut être adapté aux données quand l'effort d'échantillonnage est trop faible pour que l'estimateur Chao1 soit performant.[77] Un simple estimateur plug-in, appelé « estimateur révélé », peut ensuite être appliqué à cette distribution révélée.[78]

L'estimateur du taux de couverture généralisé ne fait aucune hypothèse sur les espèces non observées, et n'utilise que le taux de couverture et la technique de Horvitz-Thompson pour estimer leur contribution. L'estimateur révélé ne tente aucune correction mais utilise l'estimation la meilleure possible pour la distribution des probabilités.

Biais des estimateurs

Étant donné la distribution des probabilités des espèces dans la communauté, l'échantillon inventorié peut être considéré comme un tirage dans une loi multinomiale, qui contient pas toutes les espèces. Par simulation à partir d'une distribution connue, il est possible de tirer un grand nombre d'échantillons et de répéter l'estimation de la diversité sur chacun d'eux. La distribution de la valeur de l'échantillon permet de juger ses performances : son biais moyen (mesuré par rapport à la valeur de référence calculée à partir des probabilités choisies), et sa variabilité (évaluée par sa variance empirique sur les simulations). Un estimateur performant doit être peu biaisé mais aussi peu variable. L'erreur quadratique moyenne, somme du carré du biais et de la variance, ou sa racine carrée dont la dimension est celle des données (*root mean square error* : RMSE), est un bon critère de test.[79]

Haegeman *et al.*[80] proposent deux estimateurs de la diversité ${}^qD$ à partir des courbes de raréfaction et d'extrapolation du nombre d'espèces. Les deux estimateurs bornent la vraie valeur de ${}^qD$ (leur variabilité n'est pas prise en compte, mais elle est

[73] Marcon (2015). « Practical Estimation of Diversity from Abundance Data », cf. note 111, p. 35.

[74] Chao et al. (2015b). « Unveiling the Species-Rank Abundance Distribution by Generalizing Good-Turing Sample Coverage Theory », cf. note 56, p. 23.

[75] Ibid.

[76] Burnham et Overton (1979). « Robust Estimation of Population Size When Capture Probabilities Vary Among Animals », cf. note 88, p. 28.

[77] Brose et al. (2003). « Estimating species richness : Sensitivity to sample coverage and insensitivity to spatial patterns », cf. note *a*, p. 35.

[78] Marcon (2015). « Practical Estimation of Diversity from Abundance Data », cf. note 111, p. 35.

[79] Voir par exemple Chao et al. (2013). « Entropy and the species accumulation curve : a novel entropy estimator via discovery rates of new species », cf. note 29, p. 13, figure 1.

[80] B. Haegeman et al. (2013). « Robust estimation of microbial diversity in theory and in practice ». In : *The ISME journal* 7.6, p. 1092–101.

négligeable dans le cadre utilisé). L'estimateur minimal est très proche de celui de Chao et Jost, à la différence que les termes estimant le biais résiduel sont évalués par une approximation numérique. L'estimateur maximal suppose que tous les individus non échantillonnés appartiennent à une espèce nouvelle, ce qui est effectivement le pire cas possible. Les deux estimateurs sont appliqués à la diversité de communautés microbiennes (jusqu'à un million d'espèces distribuées selon une loi géométrique) et très sous-échantillonnées (un individu sur 10^{10} dans le pire des cas).

Les résultats présentent de très grands intervalles d'estimation (de plusieurs ordres de grandeur) entre la borne inférieure et la borne supérieure, ce qui montre que la méthode est inutilisable pour $q < 1$ approximativement. L'estimateur minimal, qui correspond à l'état de l'art, sous-estime sévèrement la diversité quand l'échantillonnage est faible. L'estimateur maximal est excessivement conservateur (il est conçu pour borner l'estimation dans la pire hypothèse).

Aux ordres de diversité supérieurs, les estimateurs convergent : même avec de petits échantillons, même en supposant que la richesse de la communauté est la plus grande possible, l'estimation de la diversité est précise dès $q = 1$.

Dans des conditions moins difficiles, correspondant à des communautés moins riches (typiquement des forêts tropicales), l'estimation est précise dès $q = 0,5$.[81] Ces résultats sont développés plus loin.

[81] Marcon (2015). « Practical Estimation of Diversity from Abundance Data », cf. note 111, p. 35.

Intervalle de confiance des estimateurs

À partir de données d'inventaire, l'estimation du biais est impossible (sinon, le biais serait simplement ajouté à l'estimateur pour le rendre sans biais). La variabilité de l'estimateur peut être évaluée en générant des communautés par bootstrap et en estimant leur diversité comme dans la méthode précédente. Les communautés peuvent être simulées par tirage dans une loi multinomiale respectant les probabilités observées.[82] L'inconvénient est que les espèces dont les probabilités sont petites ne sont pas tirées, ce qui entraîne un nouveau biais d'estimation. L'estimateur le corrige en partie, le biais d'estimation dû au non échantillonnage des espèces rares de la communauté (corrigé par l'estimateur appliqué aux données réelles) n'est pas accessible. Il est donc nécessaire de recentrer les valeurs obtenues par simulation autour de la valeur obtenue sur les données.

[82] Marcon et al. (2012). « The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity », cf. note 13, p. 60 ; Marcon et al. (2014a). « Generalization of the Partitioning of Shannon Diversity », cf. note 38, p. 64.

Chao et Jost[83] proposent une méthode plus élaborée qui consiste à reconstituer la distribution des probabilités de la communauté réelle le mieux possible avant d'en tirer les échantillons simulés. L'estimateur naïf de la probabilité d'une espèe observée s est $\hat{p}_s = {}^{n_s}/_{n}$. $\hat{p}_s$ est sans biais mais comme on ne dispose que d'une réalisation de la loi, il est préférable de chercher un estima-

[83] Chao et Jost (2015). « Estimating diversity and entropy profiles via discovery rates of new species », cf. note 66, p. 68.

teur sans biais conditionnellement aux espèces observées. Il est obtenu par Chao *et al.*[84] à partir de l'espérance conditionnelle de N_s/n :

$$\mathbb{E}\left(\frac{N_s}{n}|n_s > 0\right) = \frac{p_s}{1-(1-p_s)^n} \tag{3.27}$$

L'estimateur de p_s obtenu est :

$$\tilde{p}_s = \frac{n_s}{n}\left[1 - \hat{\lambda}\left(1 - \frac{n_s}{n}\right)^n\right] \tag{3.28}$$

Où :

$$\hat{\lambda} = \frac{1-\hat{C}}{\sum_{s=1}^{s^n \neq 0} \frac{n_s}{n}\left(1-\frac{n_s}{n}\right)^n} \tag{3.29}$$

λ est un paramètre d'un modèle permettant d'estimer les valeurs de p_i à partir de leur espérance conditionnelle et des données. Le modèle complet est présenté par Chao *et al.*[85]

Les probabilités des espèces non observées sont considérées comme égales par souci de simplification et parce qu'elles influencent peu la variance des estimateurs. Leur somme est égale au déficit de couverture, et leur nombre estimé par Chao1.

L'intervalle de confiance de l'estimateur obtenu par cette technique à partir des données permet de connaître sa variabilité, mais pas son biais. Il n'est absolument pas garanti (et peu probable si l'échantillonnage est faible et q petit) que la vraie valeur de la diversité se trouve dans l'intervalle de confiance.

Pratique de l'estimation

La conversion de l'estimateur de l'entropie en estimateur de la diversité pose un nouveau problème puisque ${}^qD$ n'est pas une transformation linéaire de ${}^qH$. Ce dernier biais peut être négligé[86] parce que ${}^qH$, l'entropie, est la valeur moyenne de l'information de nombreuses espèces indépendantes et fluctue donc peu à cause de l'échantillonnage, contrairement aux estimateurs des probabilités.

Des simulations menées sur des distributions log-normale ou géométrique de richesse diverses avec des efforts d'échantillonnage variés permettent de tester les estimateurs dans des cas théoriques réalistes.[87] Un résultat représentatif est donné ici.

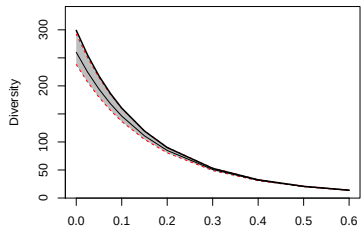
La figure 3.7 compare les résultats des deux meilleurs estimateurs à la vraie valeur de la diversité d'une communauté de 300 espèces, dont la distribution log-normale (avec un écart-type dont le logarithme vaut 2) mime une forêt tropicale similaire à Barro-Colorado. L'effort d'inventaire simulé est de 500 ou 5000 individus (de l'ordre d'un ou dix hectares pour les arbres de diamètre supérieur ou égal à 10 cm). L'estimation ne pose pas de problème

[84] Chao et al. (2013). « Entropy and the species accumulation curve : a novel entropy estimator via discovery rates of new species », cf. note 29, p. 13.

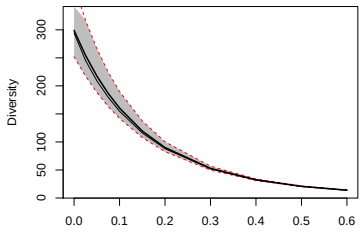
[85] Chao et al. (2015b). « Unveiling the Species-Rank Abundance Distribution by Generalizing Good-Turing Sample Coverage Theory », cf. note 56, p. 23.

[86] Grassberger (1988). « Finite sample corrections to entropy and dimension estimates », cf. note 182, p. 51.

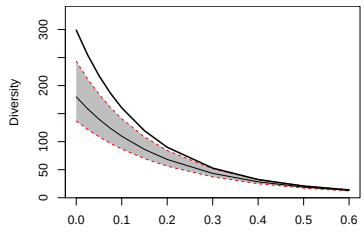
[87] Marcon (2015). « Practical Estimation of Diversity from Abundance Data », cf. note 111, p. 35.



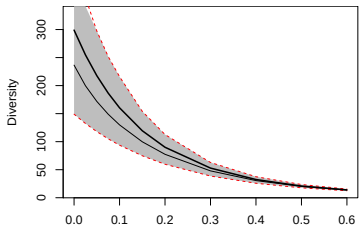
(a) 5000 individus, estimateur de Chao-Wang-Jost



(b) 5000 individus, estimateur révélé



(c) 500 individus, estimateur de Chao-Wang-Jost



(d) 500 individus, estimateur révélé

FIGURE 3.7 – Estimation de la diversité d'une communauté log-normale de 300 espèces. Un inventaire de 500 (en bas) ou 5000 (en haut) individus est répété 1000 fois et la diversité estimée à chaque fois par l'estimateur de Chao-Wang-Jost (à gauche) ou l'estimateur révélé (à droite). La courbe noire en gras représente la vraie diversité de la communauté. La courbe maigre est la moyenne de l'estimation, l'enveloppe grise limitée par les pointillés rouges contient 95% des valeurs simulées.

avec 5000 individus, bien que Chao-Wang-Jost sous-estime un peu le nombre d'espèces. L'estimateur révélé est centré sur la vraie valeur, mais avec une plus grande variabilité.

En limitant l'échantillonnage à 500 individus, l'estimateur de Chao-Wang-Jost sous-estime le nombre d'espèces de moitié. L'estimateur révélé sous-estime beaucoup moins la diversité d'ordre faible, mais son intervalle de confiance est très large.

Dès $q = 0,5$, l'estimation est très précise dans tous les cas, même en réduisant l'échantillonnage à 200 individus ou pour des communautés dont la queue de distribution est beaucoup plus longue (résultats non présentés ici).

Les conclusions de ce travail d'évaluation tiennent en deux points. Dans des conditions réalistes d'inventaire d'arbres en forêt tropicale, les estimateurs de diversité les plus performants sont celui de Chao, Wang et Jost et l'estimateur révélé. Le premier a une variance plus faible pour les valeurs de q proches de 0 mais est limité par son estimation du nombre d'espèces par l'estimateur Chao1. Le second est préférable quand le nombre d'espèces estimé par l'estimateur jackknife d'ordre optimal est clairement supérieur (c'est-à-dire, en pratique, quand l'ordre du jacknife optimal est supérieur à 1). Son biais est alors bien inférieur, au prix d'une variance plus grande.

Dans tous les cas, l'estimation de la diversité aux ordres inférieurs à 0,5 est imprécise, sauf à disposer d'inventaires de taille considérable (de l'ordre de la dizaine d'hectares au moins).

Les fonctions `Tsallis` et `Diversity` de *entropart* acceptent comme argument un vecteur d'abondances (et non de probabilités) et proposent par défaut la correction de Chao et Jost :

```
Tsallis(colSums(BCI), q = 1)

## ChaoWangJost
##     4.276193

Diversity(colSums(BCI), q = 1)

## ChaoWangJost
##     71.96597
```

MSOM et MSAM

L'utilisation de modèles hiérarchiques bayesiens fondés sur des modèles d'occupation multispécifiques (MSOM) ou des modèles d'abondance multispécifiques (MSAM) progresse dans la littérature récente. Ils nécessitent de choisir une loi de probabilité d'observation des espèces compatible avec un échantillonnage non exhaustif, par exemple une loi ZIP (*zero-inflated Poisson*),[88] dans laquelle l'espérance de la présence de chaque espèce peut être modélisée à son tour comme la conséquence de variables environnementales. Une approche alternative consiste à modéliser la probabilité d'observation comme la combinaison d'une probabilité de présence et d'une probabilité de détection.[89] Ces méthodes ont l'avantage de fournir une distribution de probabilité des paramètres des modèles, et donc de la diversité qui est calculée comme une quantité dérivée. Leur faiblesse est qu'ils dépendent de la loi choisie. Une revue complète est fournie par Iknayan *et al.*.[90]

[88] J. Zhang et al. (2014). « Partitioning of α and β diversity using hierarchical Bayesian modeling of species distribution and abundance ». In : *Environmental and Ecological Statistics* 21.4, p. 611–625.

[89] K. M. Broms et al. (2014). « Accounting for Imperfect Detection in Hill Numbers for Biodiversity Studies ». In : *Methods in Ecology and Evolution* 6.1, p. 99–108.

[90] K. J. Iknayan et al. (2014). « Detecting diversity : emerging methods to estimate species diversity ». In : *Trends in ecology & evolution* 29.2, p. 97–106.

3.4.5 Raréfaction et extrapolation

Chao *et al.*[91] étendent l'approche de Gotelli et Colwell[92] aux nombres de Hill et fournissent les méthodes nécessaires au calcul de courbes de raréfaction et d'extrapolation de l'estimateur de ${}^qD$ en fonction de la taille de l'échantillon ou du taux de couverture.

Les courbes de raréfaction sont obtenues à partir de l'échantillon disponible en simulant la diminution de sa taille. Les courbes d'extrapolation simulent l'augmentation de la taille de l'échantillon. Elles sont en réalité calculées par raréfaction d'un échantillon hypothétique de taille infinie : un estimateur de ${}^qD$ corrigé du biais d'estimation est nécessaire. L'article se limite aux valeurs de q égales à 0, 1 ou 2 ou non entières mais supérieures à 2.

[91] Chao et al. (2014a). « Rarefaction and extrapolation with Hill numbers : A framework for sampling and estimation in species diversity studies », cf. note 190, p. 54.

[92] Gotelli et Colwell (2001). « Quantifying biodiversity : procedures and pitfalls in the measurement and comparison of species richness », cf. note *a*, p. 12.

3.5 Autres approches

3.5.1 L'entropie de Rényi

L'entropie d'ordre q de Rényi[93] est :

$$ {}^qR = \frac{1}{1 - q \ln \sum_{q=1}^{S} p_s^q} \tag{3.30} $$

C'est le logarithme naturel des nombres de Hill[94] qui ont été

[93] Rényi (1961). « On Measures of Entropy and Information », cf. note 17, p. 60.

[94] Hill (1973). « Diversity and Evenness : A Unifying Notation and Its Consequences », cf. note 20, p. 60.

définis à partir d'elle. L'entropie de Rényi a connu un succès important en tant que seule mesure d'entropie généralisée jusqu'à la diffusion de l'entropie HCDT. Son défaut est qu'elle n'est pas forcément concave,[95] ce qui empêche sa décomposition (voir chapitre 5, page 123).

[95] C. Beck (2009). « Generalised Information and Entropy Measures in Physics ». In : *Contemporary Physics* 50.4, p. 495–510.

3.5.2 L'entropie généralisée des mesures d'inégalité

L'entropie généralisée des économistes, d'ordre q, est définie comme :

$$H_q = \frac{1}{q\,(q-1)} \left[\frac{1}{S} \sum_{s=1}^{S} \left(\frac{n_s}{\bar{n_s}} \right)^q - 1 \right] \tag{3.31}$$

C'est en réalité une entropie relative, qui mesure l'écart entre la distribution observée $\{n_s\}$ et sa valeur moyenne. Elle tend vers l'entropie de Theil quand $q \to 1$ et la moyenne des $\ln(n_s/\bar{n_s})$ quand $q \to 0$. Elle a été conçue pour mesurer les inégalités,[96] puis utilisée par exemple pour mesurer la concentration spatiale.[97]

Maasoumi[98] explicite cet usage et pour comparer la distribution de probabilité observée $\{q_s\}$ à une distribution attendue $\{p_s\}$, plutôt qu'à la distribution uniforme dont toutes les valeurs seraient $1/S$. En posant $r = q - 1$, en transformant les abondances $\{n_s\}$ en probabilités $\{q_s\}$ et en choisissant $\{p_s\}$ arbitrairement, on obtient :

$$H_r\left(\mathcal{Q}, \mathcal{P}\right) = \frac{1}{r\,(r+1)} \left[\sum_{s=1}^{S} q_s \left(\frac{q_s}{p_s} \right)^r - 1 \right] \tag{3.32}$$

[96] F. Cowell (2000). « Measurement of Inequality ». In : *Handbook of Income Distribution*. Sous la dir. d'A. B. Atkison et F Bourguignon. T. 1. 6. North Holland. Chap. 2, p. 87–166.

[97] M. Brülhart et R. Traeger (2005). « An Account of Geographic Concentration Patterns in Europe ». In : *Regional Science and Urban Economics* 35.6, p. 597–624.

[98] Maasoumi (1993). « A compendium to information theory in economics and econometrics », cf. note 5, p. 57.

Cressie et Read[99] ont défini une mesure généralisée de qualité d'ajustement (*Goodness of Fit*) de la distribution $\{q_s\}$ à celle de $\{p_s\}$. Elle est identique à l'entropie généralisée (à un coefficient 2 près) bien que développée indépendamment pour d'autres motivations. La statistique de Cressie et Read d'ordre 1 est la divergence de Kullback-Leibler,[100] celle d'ordre 2 le χ^2 de Pearson.[101]

[99] N. Cressie et T. R. C. Read (1984). « Multinomial Goodness-of-Fit Tests ». In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 46.3, p. 440–464.

[100] Kullback et Leibler (1951). « On Information and Sufficiency », cf. note 10, p. 59.

[101] K. Pearson (1900). « On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling ». In : *Philosophical Magazine Series 5* 50.302, p. 157–175.

Studeny *et al.*[102] définissent une mesure d'équitabilité comme la divergence entre la distribution observée $\{p_s\}$ et la distribution uniforme, ce qui les ramène à l'entropie de Cowell :

$$I_r = \frac{1}{r\,(r+1)} \left[\sum_{s=1}^{S} p_s \left(\frac{p_s}{1/S} \right)^r - 1 \right] \tag{3.33}$$

[102] A. C. Studeny et al. (2011). « Goodness-of-fit measures of evenness : a new tool for exploring changes in community structure ». In : *Ecosphere* 2.2, art.15.

Cette mesure inclut à la fois l'équitabilité et la richesse de la distribution. Il s'agit donc d'une mesure de diversité, qui est, à une normalisation près, l'écart entre la valeur de l'entropie HCDT d'ordre $q = r + 1$ et sa valeur maximale. Elle généralise donc l'indice de Theil :[103]

[103] Theil (1967). *Economics and Information Theory*, cf. note 160, p. 48.

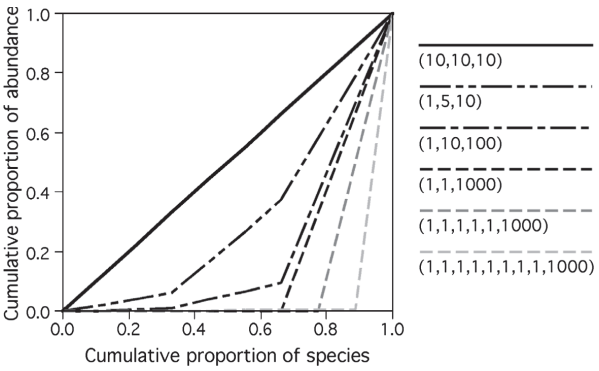


FIGURE 3.8 – Courbes de Lorenz de plusieurs jeux de données. Les espèces sont classées par fréquence croissante, et la courbe représente la fréquence cumulée des espèces. Plus la courbe est loin de la bissectrice, plus le jeu de données est inéquitable. Figure extraite de Tuomisto.[a]

[a] H. TUOMISTO (2012). « An updated consumer's guide to evenness and related indices ». In : *Oikos* 121.8, p. 1203–1218.

$$I_r = \frac{1}{qS^{1-q}} \left(\ln_q S - {}^qH \right) \tag{3.34}$$

3.6 Équitabilité

3.6.1 Approche axiomatique

La régularité d'une distribution est une notion intuitivement assez simple : la faiblesse de l'écart entre la distribution réelle et une distribution parfaitement régulière,[104] vérifiant $p_s = 1/S$.

Une approche axiomatique a été développée par la littérature, revue par Jost[105] ou Tuomisto[106] Les propriétés fondamentales sont les critères de Dalton[107] :

- L'équitabilité augmente quand un individu est transféré d'une espèce fréquente à une espèce rare ;
- L'équitabilité diminue quand une espèce rare est ajoutée ;
- Invariance d'échelle : l'équitabilité ne dépend que des fréquences, pas des abondances absolues.

Un critère supplémentaire établi par Smith et Wilson,[108] complété par Gosselin[109] est la compatibilité avec la courbe de Lorenz,[110] construite en classant les espèces de la moins fréquente à la plus fréquente, et en traçant le cumul des fréquences en fonction du rang (Figure 3.8) : une communauté parfaitement équitable est représentée par la bissectrice, la surface entre la courbe réelle et la bissectrice est très utilisée en économie pour mesurer l'inégalité sous le nom d'« indice de Gini » :[111]

- Compatibilité avec la courbe de Lorenz : L'équitabilité augmente pour une communauté si sa courbe de Lorenz est plus proche de la bissectrice. Si les courbes de deux communautés se croisent, l'ordre de leurs équitabilités est imprévisible.

Ce critère est remis en cause par Ricotta :[112] il est incompatible[113] avec le critère de continuité retenu pour les mesures de diversité : l'introduction d'une espèce supplémentaire avec une probabilité infinitésimale peut modifier considérablement l'équitabilité.

[104] M. LLOYD et R. J. GHELARDI (1964). « A Table for Calculating the 'Equitability' Component of Species Diversity ». In : *Journal of Animal Ecology* 33.2, p. 217–225.

[105] JOST (2010b). « The Relation between Evenness and Diversity », cf. note 22, p. 4.

[106] H. TUOMISTO (2012). « An updated consumer's guide to evenness and related indices ». In : *Oikos* 121.8, p. 1203–1218.

[107] DALTON (1920). « The measurement of the inequality of incomes », cf. note 19, p. 60.

[108] B. SMITH et J. B. WILSON (1996). « A consumer's guide to evenness indices ». In : *Oikos* 76.1, p. 70–82.

[109] F. GOSSELIN (2001). « Lorenz partial order : the best known logical framework to define evenness indices ». In : *Community Ecology* 2.2, p. 197–207.

[110] M. O. LORENZ (1905). « Methods of Measuring the concentration of Wealth ». In : *Quarterly Publications of the American Statistical Association* 9, p. 209–219.

[111] GINI (1912). *Variabilità e mutabilità*, cf. note 154, p. 46 ; L. CERIANI et P. VERME (2012). « The origins of the Gini index : extracts from Variabilità e Mutabilità (1912) by Corrado Gini ». In : *Journal of Economic Inequality* 10.3, p. 421–443.

[112] C. RICOTTA (2004b). « A recipe for unconventional evenness measures ». In : *Acta Biotheoretica* 52.2, p. 95–104.

[113] R. D. ROUTLEDGE (1983). « Evenness indices : Are any admissible ? » In : *Oikos* 40.1, p. 149–151.

L'invariance par réplication est une propriété nécessaire (mais pas suffisante) à la compatibilité avec la courbe de Lorenz :[114]

- Invariance par réplication : l'équitabilité ne change pas si chaque espèce est remplacée par plusieurs nouvelles espèces ayant la même abondance.

[114] Ricotta (2004b). Cf. note 112.

Cette propriété est retenue par Tuomisto mais rejetée par Jost comme par Ricotta et d'autres. L'argumentation de Jost est la plus convaincante : une communauté d'une richesse donnée dans laquelle tous les individus (à l'exception d'un nombre négligeable) sont concentrés dans une seule espèce atteint le niveau minimum d'équitabilité. Si elle est dupliquée, les individus seront répartis dans deux espèces et non une : la nouvelle communauté n'atteindra pas le niveau minimum d'équitabilité, il est donc souhaitable que la mesure d'équitabilité augmente.

3.6.2 Indice classiques

Indice de Bulla

L'indice d'équitabilité de Bulla[115] respecte les critères de Smith et Wilson :

$$O = \sum_s \min(p_s, \frac{1}{S}) \tag{3.35}$$

[115] L. Bulla (1994). « An index of evenness and its associated diversity measure ». In : *Oikos* 70.1, p. 167–172.

Il mesure l'écart entre la distribution réelle et une distribution parfaitement équitable. Bulla le normalise pour qu'il soit compris entre 0 et 1 :

$$E_O = \frac{O - \min O}{1 - \min O} \tag{3.36}$$

La valeur minimale de O est $1/S$. Bulla considère que ce n'est pas le cas si les effectifs sont entiers et fournit une valeur exacte dépendant du nombre d'espèces.

Indice de Gini

L'indice de Gini est le double de la surface comprise entre la courbe de Lorenz et la bissectrice (Figure 3.8).

Rapports de moyennes

Taillie[116] propose le rapport de la moyenne géométrique ou de la moyenne harmonique à la moyenne arithmétique des probabilités.

[116] C. Taillie (1979). « Species equitability : a comparative approach ». In : *Ecological Diversity in Theory and Practice*. Sous la dir. de J. F. Grassle et al. Fairland, MD : International Cooperative Publishing House, p. 51–62.

Ratios de Hill

Hill[117] utilise les rapports entre nombres effectifs d'espèces d'ordres différents comme mesure d'équitabilité. $E_{1,0} = {}^1D/{}^0D$ avait déjà été proposé par Sheldon.[118] Peet[119] le normalise en retirant au numérateur et au dénominateur leur valeur minimale, c'est-à-dire 1.

La version normalisée de $E_{2,1}$ est recommandée par Alatalo.[120]

[117] Hill (1973). « Diversity and Evenness : A Unifying Notation and Its Consequences », cf. note 20, p. 60.

[118] A. L. Sheldon (1969). « Equitability indices : dependence on the species count ». In : *Ecology* 50.3, p. 466–467.

[119] Peet (1974). « The measurement of species diversity », cf. note 1, p. 9.

[120] R. V. Alatalo (1981). « Problems in the Measurement of Evenness in Ecology ». In : *Oikos* 37.2, p. 199–204.

Synthèse

Riccota *et al.*[121] testent ces indices sur 65 jeux de données différents pour observer leurs corrélations. Tous sont compatibles avec la courbe de Lorenz. Ils concluent que les indices se répartissent en deux groupes : la plupart sont sensibles aux abondances relatives des espèces rares. $E_{\infty,0}$ et, dans une moindre mesure, $E_{2,0}$ sont plus sensibles aux espèces fréquentes.

[121] C. Ricotta et al. (2001). « On the mutual relatedness of evenness measures ». In : *Community Ecology* 2.1, p. 51–56.

3.6.3 Indice alternatifs

Variance de la rareté

Engen[122] définit comme mesure d'équitabilité la variance de la rareté (la fonction d'information) de l'entropie de Shannon ($-\ln p$) :

$$E_{VS} = \sum_s p_s (\ln p_s)^2 - (\sum_s p_s \ln p_s)^2 \tag{3.37}$$

Sa valeur est nulle si les espèces sont équiprobables.

[122] Engen (1977). « Exponential and Logarithmic Species-Area Curves », cf. note 39, p. 15.

Riccota[123] généralise cette mesure à d'autres fonctions de rareté. Toutes peuvent être utilisées.

$$E_{VR} = \sum_s p_s I(p_s)^2 - [\sum_s p_s I(p_s)]^2 \tag{3.38}$$

Ces mesures ne sont pas invariantes par réplication mais sont continues par rapport aux probabilités.

[123] C. Ricotta (2003). « On parametric evenness measures ». In : *Journal of Theoretical Biology* 222.2, p. 189–197.

Non-spécificité

Riccota[124] propose un indice fondé sur les ensembles flous, utilisant les mesures de spécificité.[125] Les espèces sont classées par probabilité décroissante. La mesure de non-spécificité utilisable comme mesure d'équitabilité est :

$$E_{NSp} = \sum_{s=1}^{S-1} \frac{p_s - p_{s+1}}{s} \tag{3.39}$$

Cet indice n'est pas invariant par réplication mais est continu par rapport aux probabilités.

[124] Ricotta (2004b). « A recipe for unconventional evenness measures », cf. note 112, p. 76.

[125] R. R. Yager (1992). « On the specificity of a possibility distribution ». In : *Fuzzy Sets and Systems* 50.3, p. 279–292.

3.6.4 Indice de Pielou

Une expression de l'équitabilité est souvent donnée à partir de l'indice de Shannon.[126] La valeur maximale de l'indice de Shannon est obtenue quand la distribution est parfaitement régulière. Alors : $H_{\max} = \ln S$. On a donc défini l'indice, souvent appelé « indice de Pielou » :

$$J = \frac{H}{H_{\max}} = \frac{H}{\ln S} \tag{3.40}$$

[126] Lloyd et Ghelardi (1964). « A Table for Calculating the 'Equitability' Component of Species Diversity », cf. note 104, p. 76 ; Sheldon (1969). « Equitability indices : dependence on the species count », cf. note 118, p. 77 ; Pielou (1966b). « The measurement of diversity in different types of biological collections », cf. note 159, p. 48 ; Pielou (1975). *Ecological Diversity*, cf. note 1, p. 57.

J est compris entre 0 (une seule espèce a une probabilité de 1) et 1 (toutes les espèces ont la même probabilité).

3.6.5 Généralisation de l'indice de Pielou

Mendes *et al.*

Mendes *et al.*[127] généralisent l'indice de Pielou en remplaçant l'indice de Shannon par l'entropie de Tsallis. Ils définissent :

$$ {}^{q}J^{*} = \frac{{}^{q}H}{{}^{q}H_{\max}} = \frac{{}^{q}H}{\ln_q S} \tag{3.41} $$

et tracent des profils d'équitabilité, c'est-à-dire la valeur de ${}^{q}J^{*}$ en fonction de q. Ils observent que l'équitabilité est minimale pour une valeur particulière de q qu'ils notent q^{*} et montrent sur des exemples simulés que q^{*} diminue avec la richesse de la communauté, mais sans en explorer complètement les propriétés.

[127] Mendes et al. (2008). « A unified index to measure ecological diversity and species rarity », cf. note 31, p. 62.

Tuomisto

Tuomisto argumente en faveur d'une définition stricte de l'équitabilité (*evenness* en anglais, *equitability* pouvant être utilisé dans un sens plus large) en tant que ratio de Hill :

$$ {}^{q}E = E_{q,0} = \frac{{}^{q}D}{{}^{0}D} = \frac{{}^{q}D}{S} \tag{3.42} $$

L'équitabilité d'ordre q est le rapport de la diversité d'ordre q sur la richesse.

L'idée fondamentale de la notion d'équitabilité est que la mesure de la diversité doit pouvoir être partitionnée en deux composantes indépendantes : la richesse et l'équitabilité.[128]

[128] Smith et Wilson (1996). « A consumer's guide to evenness indices », cf. note 108, p. 76.

Jost

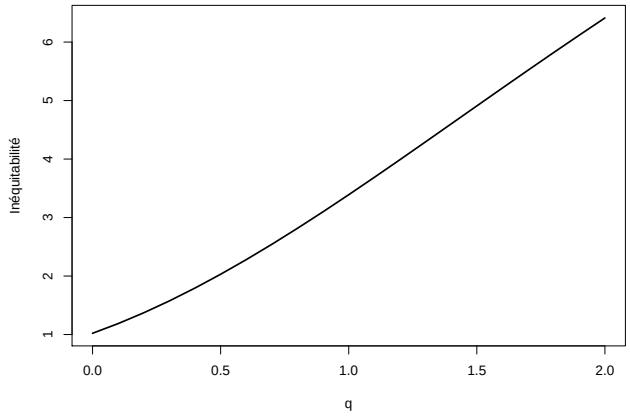


Figure 3.9 – Profil d'inéquitabilité de BCI : L'inéquitabilité augmente quand on néglige de plus en plus les espèces rares. Pour q=2, la forêt de BCI est aussi inéquitable qu'une communauté de 6 espèces dans laquelle une seule serait représentée au-delà d'un nombre négligeable d'individus.

Jost montre que l'indépendance est impossible dans le cadre de la décomposition multiplicative : si ${}^qE = {}^qD/S$, sa valeur minimale est contrainte par le nombre d'espèces, une communauté pauvre ne peut pas être aussi inéquitable qu'une communauté riche (propriété déjà montrée par Sheldon[129] ou Alatalo[130]). En revanche, la richesse peut être partitionnée de façon indépendante en diversité et inéquitabilité : $S = {}^qD/{}^qE$.

L'inéquitabilité $1/{}^qE$ (Figure 3.9 [131]) est un nombre effectif d'espèces : c'est la richesse d'une communauté du même niveau d'équitabilité que la communauté réelle dans laquelle tous les individus (sauf un nombre négligeable) appartiendraient à une seule espèce.

[129] Sheldon (1969). « Equitability indices : dependence on the species count », cf. note 118, p. 77.

[130] Alatalo (1981). « Problems in the Measurement of Evenness in Ecology », cf. note 120, p. 77.

Synthèse

L'indice de Pielou est le plus utilisé dans la littérature[132] mais n'est pas invariant par réplication parce qu'il ne normalise pas une diversité mais une entropie. Il est rejeté pour cette raison[133] par une partie de la littérature. Jost[134] argumente au contraire en sa faveur : ${}^qE$ dépend de la richesse et ne peut donc pas être comparé entre communautés de richesse différente. Le logarithme de ${}^qE$ normalisé est une meilleure généralisation de l'indice de Pielou que celle de Mendes *et al.* :

$$ {}^qJ = \frac{\ln({}^qD)}{\ln S} = \frac{{}^qR}{\ln S} \tag{3.43} $$

${}^qJ$ varie entre 0 et 1 et est une transformation monotone de ${}^qE$ contrairement au ${}^qJ^*$ de Mendes *et al.* Il est égal au rapport de l'entropie de Rényi sur le logarithme du nombre d'espèces.

L'estimation de l'équitabilité repose lourdement sur l'estimation du nombre d'espèces, quelle que soit la valeur du paramètre q. La solution la plus évidente consiste à corriger le mieux possible le biais d'estimation du nombre d'espèces et celui de l'entropie d'ordre q :

$$ {}^q\tilde{J} = \frac{\ln\left(e_q^{{}^q\tilde{H}}\right)}{\ln \tilde{S}} \tag{3.44} $$

[131] Code R :

```
# Estimation du nombre d'espèces
library(SPECIES)
Ns <- colSums(BCI)
UBNs <- jackknife(AbdFreqCount(Ns), k=1)$Nhat
# Profil
DivPrf <- CommunityProfile(Diversity, Ns)
plot(y=UBNs/DivPrf$y, x=q.seq, type="l", col="black",
lwd=2, xlab="q", ylab="Inéquitabilité", main="")
```

[132] Tuomisto (2012). « An updated consumer's guide to evenness and related indices », cf. note *a*, p. 76.

[133] Smith et Wilson (1996). « A consumer's guide to evenness indices », cf. note 108, p. 76 ; Tuomisto (2012). « An updated consumer's guide to evenness and related indices », cf. note *a*, p. 76.

[134] Jost (2010b). « The Relation between Evenness and Diversity », cf. note 22, p. 4.

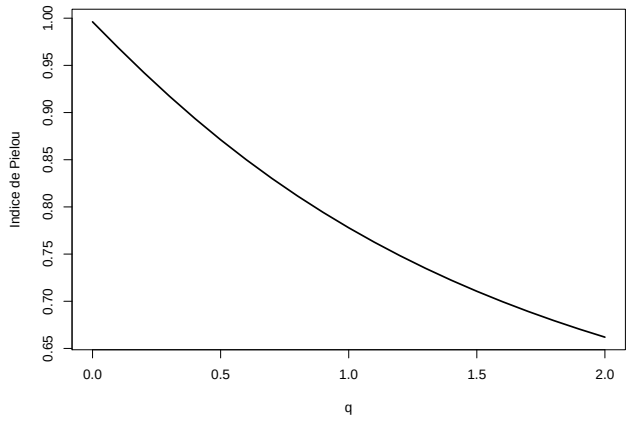


FIGURE 3.10 – Profil d'équitabilité de BCI : représentation alternative de la Figure 3.9. L'équitabilité est ici normalisée selon Jost.

Un profil d'équitabilité calculé de cette façon est présenté Figure 3.10 [135].

[135] Code R supplémentaire :

```
plot(y=log(DivPrf$y)/log(UBNs), x=q.seq, type="l", col="black",
lwd=2, xlab="q", ylab="Indice de Pielou", main="")
```

Chapitre 4

Diversité fonctionnelle et phylogénétique

Les mesures neutres de la diversité considèrent que toutes les classes auxquelles les objets appartiennent sont différentes, sans que certaines soient plus différentes que d'autres. Par exemple, toutes les espèces sont équidistantes les unes des autres, qu'elles appartiennent au même genre ou à des familles différentes. Intuitivement, l'idée qu'une communauté de S espèces toutes de genres différents est plus diverse qu'une communauté de S espèces du même genre est satisfaisante. Walker[1] argumente en faveur de la protection de groupes fonctionnels plutôt que de celle de chacune des espèces qui les constitue pour maintenir le bon état des écosystèmes.

Il s'agit donc de caractériser la différence entre deux classes d'objets, puis de construire des mesures de diversité en rapport.[2] En écologie, ces différences sont fonctionnelles ou phylogénétiques, définissant la diversité fonctionnelle[3] ou la diversité phylogénétique (*phylodiversity*).[4] Les premières propositions de ce type d'indices sont dues à Rao[5] (voir page 97) puis, avec nettement moins de succès, Vane-Wright *et al.*[6] et Warwick et Clarke.[7] Chave *et al.*[8] montrent que la diversité neutre prédit mal la diversité phylogénétique (calculée par l'entropie quadratique de Rao).

De nombreuses mesures de diversité ont été créées et plusieurs revues permettent d'en faire le tour.[9] Les mesures présentées ici sont les plus utilisées, et notamment celles qui peuvent être ramenées aux mesures classiques en fixant une distance égale entre toutes les espèces.

Des exemples montrent comment calculer cette diversité, principalement à l'aide du package *entropart*. Le package contient les données d'inventaire de deux hectares du dispositif de Paracou et la taxonomie des espèces concernées :

```
library(entropart)
# Chargement du jeu de données
data(Paracou618)
```

[1] B. H. Walker (1992). « Biodiversity and Ecological Redundancy ». In : *Conservation biology* 6.1, p. 18–23.

[2] Pielou (1975). *Ecological Diversity*, cf. note 1, p. 57 ; R. M. May (1990). « Taxonomy as Destiny ». In : *Nature* 347, p. 129–130 ; S. H. Cousins (1991). « Species diversity measurement : Choosing the right index ». In : *Trends in Ecology and Evolution* 6.6, p. 190–192.

[3] D. Tilman et al. (1997). « The Influence of Functional Diversity and Composition on Ecosystem Processes ». In : *Science* 277.5330, p. 1300–1302.

[4] C. O. Webb et al. (2006). « Integrating Phylogenies into Community Ecology ». In : *Ecology* 87.sp7, S1–S2.

[5] C. R. Rao (1982). « Diversity and dissimilarity coefficients : a unified approach ». In : *Theoretical Population Biology* 21, p. 24–43.

[6] R. Vane-Wright et al. (1991). « What to protect ?—Systematics and the agony of choice ». In : *Biological Conservation* 55.3, p. 235–254.

[7] R. M. Warwick et K. R. Clarke (1995). « New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress ». In : *Marine Ecology Progress Series* 129, p. 301–305.

[8] J. Chave et al. (2007). « The importance of phylogenetic structure in biodiversity studies ». In : *Scaling biodiversity*. Sous la dir. de D. Storch et al. Santa Fe : Institute Editions, p. 150–167.

[9] Ricotta (2007). « A semantic taxonomy for diversity measures », cf. note 18, p. 4 ; M. Vellend et al. (2010). « Measuring phylogenetic biodiversity ». In : *Biological diversity : frontiers in measurement and assessment*. Sous la dir. d'A. E. Magurran et B. J. McGill. Oxford : Oxford University Press, p. 194–207 ; Pavoine et Bonsall (2011). « Measuring biodiversity to explain community assembly : a unified approach », cf. note 19, p. 4.

4.1 Cadre

4.1.1 Dissimilarité et distance

Une similarité ou dissimilarité est toute application à valeurs numériques qui permet de mesurer le lien entre les individus d'un même ensemble ou entre les variables. Pour une similarité le lien est d'autant plus fort que sa valeur est grande.

Un indice de dissimilarité d (on peut dire une dissimilarité) vérifie (k et l sont deux individus) :

- La dissimilarité d'un individu avec lui-même est nulle : $d(k,k) = 0$;
- La dissimilarité entre deux individus différents est positive : $d(k,l) \geq 0$;
- La dissimilarité est symétrique : $d(k,l) = d(l,k)$.

Une distance vérifie en plus :

- La distance entre deux individus différents est strictement positive : $d(k,l) = 0 \Rightarrow k = l$;
- L'inégalité triangulaire : $d(k,m) \leq d(k,l) + d(l,m)$. De nombreux indices de dissimilarité ne vérifient pas cette propriété.

Une distance est euclidienne si elle peut être représentée par des figures géométriques. On peut rendre toute distance euclidienne par ajout d'une constante.[10] Dans R, utiliser `is.euclid()` pour vérifier qu'une distance est euclidienne, et `cailliez()` ou `lingoes()` pour la transformation.

Enfin, une distance est ultramétrique si $d(k,m) \leq \max(d(k,l), d(l,m))$. Les distances obtenues en mesurant les longueurs des branches d'un dendrogramme (arbre) résultant d'une classification hiérarchique sont ultramétriques.

La façon exacte de mesurer les longueurs de branche est illustrée par la Figure 4.1 : la distance entre les espèces 1 et 2 est $T_1 + T_2$, elle est égale à T_1 entre les espèces 4 et 5. La distance est la hauteur du premier nœud commun.

4.1.2 Distance phylogénétique

La façon la plus évidente de définir une distance entre espèces est d'utiliser la taxonomie,[11] en attribuant une distance arbitraire (par exemple 1) à deux espèces du même genre, une autre (par exemple 2) à deux espèces de la même famille, etc. La distance définie est ultramétrique.

La taxonomie peut être remplacée avantageusement par une phylogénie. La phylogénie idéale contiendrait l'histoire évolutive de toutes les espèces et les distances seraient les temps de divergence depuis le premier ancêtre commun. En pratique, les phylogénies sont établies à partir d'un nombre de marqueurs génétiques limités, sans datation précise (mais avec des calages partiels à partir de

[10] J. C. Lingoes (1971). « Some boundary conditions for a monotone analysis of symmetric matrices ». In : *Psychometrika* 36.2, p. 195–203 ; F. Cailliez (1983). « The analytical solution of the additive constant problem ». In : *Psychometrika* 48, p. 305–310.

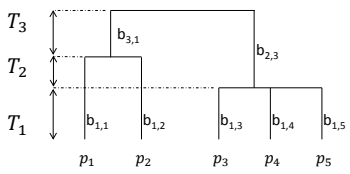


Figure 4.1 – Arbre phylogénétique ou fonctionnel hypothétique. Rappel de la figure 4.7(a). 5 espèces sont présentes ($S = 5$), leurs probabilités notées p_1 à p_5. Les noms des branches sont affichés.

[11] Clarke et Warwick (2001). « A further biodiversity index applicable to species lists : variation in taxonomic distinctness », cf. note 3, p. 9 ; R. M. Warwick et K. R. Clarke (2001). « Practical Measures of Marine Biodiversity Based on Relatedness of Species ». In : *Oceanography and Marine Biology. An Annual Review*. Sous la dir. de R. Gibson et al. T. 39. New York : Taylor & Francis, p. 207–232.

fossiles), et ne sont pas toujours ultramétriques. Elles peuvent prendre en compte chaque individu, sans regroupement par espèce. Une méthode pour dater une phylogénie est fournie par Chave *et al.*.[12] Zanne *et al.*[13] fournissent une phylogénie datée de plus de 32 000 espèces. Ricotta[14] montre sur des exemples que la diversité calculée à partir de phylogénies datées est très corrélée à celle calculée à partir de simples taxonomies.

Dans tous les cas, la distance est une mesure de la divergence évolutive. Du point de vue de la biologie de la conservation, chaque espèce accumule une quantité d'évolution, interprétée comme une quantité d'information[15] dont le maximum doit être préservé.

L'entropie phylogénétique (voir page 102) utilise un arbre phylogénétique pour mesurer la diversité.

[12] Chave et al. (2007). « The importance of phylogenetic structure in biodiversity studies », cf. note 8, p. 83.

[13] A. E. Zanne et al. (2014). « Three keys to the radiation of angiosperms into freezing environments ». In : *Nature* 506.7486, p. 89–92.

[14] C. Ricotta et al. (2012). « Computing diversity from dated phylogenies and taxonomic hierarchies : Does it make a difference to the conclusions ? » In : *Oecologia* 170.2, p. 501–506.

[15] R. H. Crozier (1997). « Preserving the Information Content of Species : Genetic Diversity , Phylogeny , and Conservation Worth Interaction with Reasons Justifying the Preservation ». In : *Annual Review of Ecology and Systematics* 28, p. 243–268.

4.1.3 Distance fonctionnelle

L'approche fonctionnelle est différente. Chaque espèce ou individu est représenté par ses valeurs de traits dans un espace multidimensionnel. Le vecteur de traits est considéré comme un proxy de la niche écologique. Les individus proches dans l'espace des traits sont donc considérés comme proche écologiquement. Les distances entre les points peuvent être calculées directement dans l'espace des traits ou, fréquemment, un arbre est construit par classification automatique hiérarchique.

La première étape consiste donc à choisir un ensemble de traits pertinents et à les mesurer de façon standardisée.[16] Toute la stratégie relative à la photosynthèse peut être par exemple assez bien résumée par la masse surfacique des feuilles,[17] mais, en forêt tropicale, ce trait est décorrélé de la densité du bois.[18] Les valeurs manquantes peuvent être complétées en utilisant toute l'information disponible par MICE (*multiple imputation by chained equations*),[19] disponible sous R dans le package `mice`.[20]

La prise en compte de variables qualitatives ou de rang et la possibilité de données manquantes pose un problème pratique de construction de la matrice de dissimilarité, traité par Gower.[21] La formule de Gower, étendue par Podani[22] puis Pavoine *et al.*[23] à d'autres types de variables, calcule la dissimilarité entre deux espèces par la moyenne des dissimilarités calculées pour chaque trait, dont la valeur est comprise entre 0 et 1 :

- Pour une variable quantitative, la différence de valeur entre deux espèces est normalisée par l'étendue des valeurs de la variable ;
- Les variables ordonnées sont remplacées par leur rang et traitées comme les variables quantitatives ;
- Pour des variables qualitatives, la dissimilarité vaut 0 ou 1 ;
- Les valeurs manquantes sont simplement ignorées et n'entrent pas dans la moyenne.

Une matrice de dissimilarités est construite de cette façon.

[16] J. H. C. Cornelissen et al. (2003). « A handbook of protocols for standardised and easy measurement of plant functional traits worldwide ». In : *Australian Journal of Botany* 51.4, p. 335–380.

[17] I. J. Wright et al. (2004). « The worldwide leaf economics spectrum ». In : *Nature* 428, p. 821–827.

[18] C. Baraloto et al. (2010). « Functional trait variation and sampling strategies in species rich plant communities ». In : *Functional Ecology* 24, p. 208–216.

[19] S. van Buuren et al. (2006). « Fully conditional specification in multivariate imputation ». In : *Journal of Statistical Computation and Simulation* 76.12, p. 1049–1064.

[20] S. van Buuren et K. Groothuis-Oudshoorn (2011). « mice : Multivariate Imputation by Chained Equations in R ». In : *Journal of Statistical Software* 45.3, p. 1–67.

[21] J. C. Gower (1971). « A General Coefficient of Similarity and Some of Its Properties ». In : *Biometrics* 27.4, p. 857–871.

[22] J. Podani (1999). « Extending Gower's General Coefficient of Similarity to Ordinal Characters ». In : *Taxon* 48.2, p. 331–340.

[23] S. Pavoine et al. (2009b). « On the challenge of treating various types of variables : application for improving the measurement of functional diversity ». In : *Oikos* 118.3, p. 391–402.

Podani et Schmera[24] suggèrent d'utiliser ensuite une classification hiérarchique par UPGMA[25] qu'ils montrent être la plus robuste (pour le calcul de FD, voir page 96) à l'ajout ou au retrait d'un trait ou d'une espèce. Mouchet *et al.*[26] suggèrent plutôt d'appliquer toutes les méthodes de classification et de retenir à la fin l'arbre dont la distribution des distances entre espèces dans l'arbre est la plus proche de la distribution des distances dans la matrice de dissimilarités : cette proximité est mesurée par la corrélation cophénétique, c'est-à-dire le coefficient de corrélation entre les valeurs de distances.[27] Un arbre consensus[28] est souvent plus proche de la matrice de distance.

Un dendrogramme fonctionnel n'a pas d'interprétation aussi claire qu'un arbre phylogénétique qui représente le processus de l'évolution. Il peut être interprété comme la représentation à des échelles de plus en plus grossières en allant vers le haut de l'arbre de regroupements fonctionnels dans des niches de plus en plus vastes.

La transformation d'une matrice (non ultramétrique) en dendrogramme déforme la topologie des espèces[29] : une mesure de diversité qui utilise directement la matrice est préférable, c'est un intérêt de la diversité de Leinster et Cobbold (voir page 105).

4.1.4 Équivalence des deux diversités

L'approche fonctionnelle étant particulièrement complexe et lourde à mettre en œuvre (notamment pour la mesure des traits sur chaque individu), la tentation a été grande de considérer que la phylogénie contenait plus d'information fonctionnelle que ce qui pouvait être mesuré, et donc de considérer la diversité phylogénétique comme proxy de la diversité fonctionnelle.

Du point de vue théorique, le modèle le plus simple de l'évolution de la valeur d'un trait hypothétique au cours du temps est le mouvement brownien : à chaque génération, la valeur du trait varie un peu, sans mémoire. Dans ce cadre,[30] la variance de la valeur actuelle du trait pour une espèce donnée est proportionnelle à la durée de l'évolution et la covariance entre deux espèces à celle de l'âge de leur ancêtre commun. Deux espèces proches dans l'arbre phylogénétique décrivant l'évolution doivent donc avoir des valeurs de trait corrélées.

Webb[31] a montré que des communautés d'arbres tropicaux avaient une moins grande diversité phylogénétique locale qu'attendue sous l'hypothèse nulle d'une distribution aléatoire des espèces, et a supposé que la cause en était le filtrage environnemental local, agissant sur les traits et observables par la phylogénie, sous l'hypothèse de conservation phylogénétique des traits fonctionnels. La discipline appelée *écologie phylogénétique des communautés* cherche encore à comprendre quels traits sont conservés et lesquels sont convergents.[32]

[24] J. Podani et D. Schmera (2006). « On dendrogram-based measures of functional diversity ». In : *Oikos* 115.1, p. 179–185.

[25] R. R. Sokal et C. D. Michener (1958). « A statistical method for evaluating systematic relationships ». In : *The University of Kansas Science Bulletin* 38.22, p. 1409–1438.

[26] M. Mouchet et al. (2008). « Towards a consensus for calculating dendrogram-based functional diversity indices ». In : *Oikos* 117, p. 794–800.

[27] R. R. Sokal et F. J. Rohlf (1962). « The Comparison of Dendrograms by Objective Methods ». In : *Taxon* 11.2, p. 33–40 ; P. Legendre et L. Legendre (2012). *Numerical Ecology*. 3rd Ed. Elsevier.

[28] J. Felsenstein (2004). *Inferring Phylogenies*. 2nd Ed. Sinauer Associates.

[29] S. Pavoine et al. (2005a). « Is the originality of a species measurable ? » In : *Ecology Letters* 8, p. 579–586 ; J. Podani et D. Schmera (2007). « How should a dendrogram-based measure of functional diversity function ? A rejoinder to Petchey and Gaston ». In : *Oikos* 116.8, p. 1427–1430.

[30] J. Felsenstein (1985). « Phylogenies and the Comparative Method ». In : *The American Naturalist* 25.1, p. 1–15.

[31] C. O. Webb (2000). « Exploring the phylogenetic structure of ecological communities : An example for rain forest trees ». In : *American Naturalist* 156.2, p. 145–155.

[32] J. Cavender-Bares et al. (2009). « The merging of community ecology and phylogenetic biology ». In : *Ecology Letters* 12.7, p. 693–715.

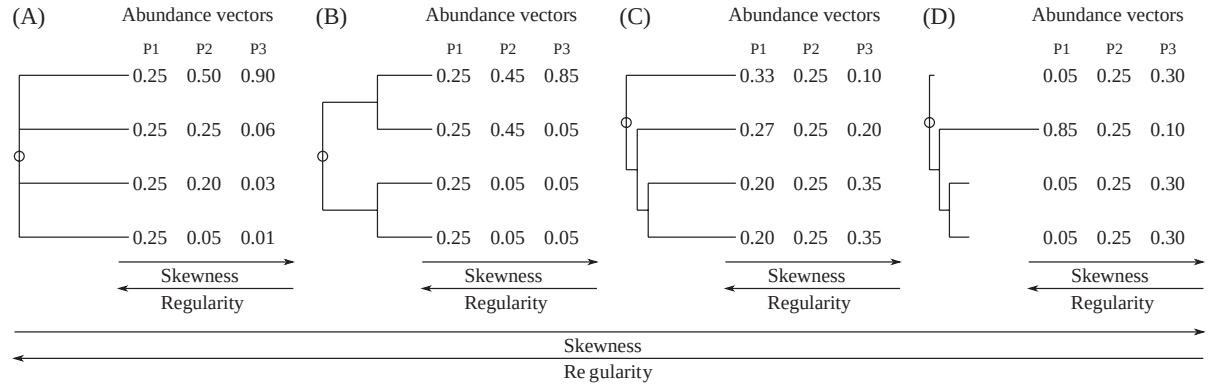


FIGURE 4.2 – Régularité contre irrégularité. Les arbres de A à D sont de plus en plus irréguliers. L'arbre A, parfaitement régulier, est le cadre des mesures classiques de la diversité. La phylogénie étant donnée, trois vecteurs d'abondance (P1 à P3) sont de moins en moins réguliers : dans les cas C et D, la régularité maximale n'est pas obtenue pour des effectifs identiques, mais en augmentant les effectifs des espèces originales (voir Maximum théorique, page 99). La figure est issue de S. PAVOINE et M. B. BONSALL (2011). « Measuring biodiversity to explain community assembly : a unified approach ». In : *Biological Reviews* 86.4, p. 792–812, fig. 1.

Swenson et Enquist[33] ont montré que la relation entre les deux diversités était faible. L'utilisation de la diversité phylogénétique comme proxy de la diversité fonctionnelle n'est satisfaisante.[34] Pour optimiser la conservation, les deux aspects de la diversité, souvent divergents, doivent être pris en compte.[35]

4.1.5 Typologie des mesures

À partir de la littérature,[36] une typologie des mesures de diversité émerge. Elle étend les notions classiques de richesse et équitabilité.

La richesse est l'accumulation de classes différentes dans les mesures classiques. Dans un arbre phylogénétique, la longueur des branches représente un temps d'évolution : la richesse en est la somme. FD et PD sont des mesures de richesse.

La régularité est une notion plus complexe (Figure 4.2). Elle mesure la façon dont les espèces occupent uniformément l'espace des niches.[37] Cette notion est simple dans un espace multidimensionnel (par exemple, l'espace des traits fonctionnels). Dans un arbre phylogénétique, la régularité de l'arbre[38] est un premier critère, complété éventuellement par les abondances. Dans un arbre parfaitement régulier, la régularité se réduit à l'équitabilité.

Les mesures de divergence sont des fonctions croissantes de la dissimilarité entre les espèces, généralement considérées par paire. Certaines sont pondérées par les abondances, d'autres non. Dans un arbre parfaitement régulier, l'indice de Simpson est une mesure de divergence pondérée. Ces mesures sont influencées par la richesse et la régularité.

Face à la profusion des mesures de diversité fonctionnelle, Ricotta,[39] en complément de Solow et Polaski[40] établit un certain nombre d'axiomes :

[33] N. G. SWENSON et B. J. ENQUIST (2009). « Opposing assembly mechanisms in a Neotropical dry forest : Implications for phylogenetic and functional community ecology ». In : *Ecology* 90.8, p. 2161–2170.

[34] PAVOINE et BONSALL (2011). « Measuring biodiversity to explain community assembly : a unified approach », cf. note 19, p. 4.

[35] DEVICTOR et al. (2010). « Spatial mismatch and congruence between taxonomic, phylogenetic and functional diversity : the need for integrative conservation strategies in a changing world. », cf. note 2, p. 9.

[36] RICOTTA (2007). « A semantic taxonomy for diversity measures », cf. note 18, p. 4 ; PAVOINE et BONSALL (2011). « Measuring biodiversity to explain community assembly : a unified approach », cf. note 19, p. 4.

[37] PAVOINE et BONSALL (2011). « Measuring biodiversity to explain community assembly : a unified approach », cf. note 19, p. 4.

[38] A. Ø. MOOERS et S. B. HEARD (1997). « Inferring Evolutionary Process from Phylogenetic Tree Shape ». In : *The Quarterly Review of Biology* 72.1, p. 31–54.

[39] C. RICOTTA (2005a). « A note on functional diversity measures ». In : *Basic and Applied Ecology* 6.5, p. 479–486.

[40] A. R. SOLOW et S. POLASKY (1994). « Measuring biological diversity ». In : *Environmental and Ecological Statistics* 1.2, p. 95–103.

1. Monotonicité d'ensemble : la diversité ne doit pas diminuer quand une espèce est ajoutée avec une faible probabilité (qui ne modifie pas la structure de la communauté existante), quelles que soient ses caractéristiques fonctionnelles ;
2. Jumelage :[41] l'introduction d'une espèce identique à une espèces existante ne doit pas augmenter la diversité. De façon moins triviale, une espèce infiniment proche ne doit pas augmenter la diversité : il s'agit donc d'un axiome de continuité dans l'espace des niches.
3. Monotonicité de distance : la diversité ne doit pas diminuer quand la distance entre espèces est augmentée.
4. Décomposabilité : les mesures de divergence doivent être décomposables en diversité α, β et γ, ce qui implique leur concavité par rapport aux probabilités.

L'entropie phylogénétique permet d'unifier ces notions.

[41] M. L. Weitzman (1992). « On Diversity ». In : *The Quarterly Journal of Economics* 107.2, p. 363–405.

4.2 Mesures particulières

Un certain nombre de mesures de diversité phylogénétiques a émergé dans la littérature. Elles sont passées en revue ici, en commençant par la diversité fonctionnelle envisagée dans l'espace multidimensionnel des traits. Les sections suivantes envisagent les espèces dans un arbre phylogénétique.

4.2.1 Richesse, équitabilité et divergence fonctionnelle

Mason *et al.*[42] postulent que la diversité fonctionnelle peut être abordée dans trois dimensions indépendantes (Jost[43] montrera que l'indépendance n'est pas assurée) :

- la richesse fonctionnelle, qui indique l'étendue de l'espace des niches fonctionnelles occupé par la communauté ;
- l'équitabilité de la distribution des espèces dans ces niches ;
- la divergence fonctionnelle, qui mesure comment la distribution des espèces dans l'espace des niches maximise la variabilité des caractéristiques fonctionnelles dans la communauté.

[42] N. W. H. Mason et al. (2005). « Functional richness, functional evenness and functional divergence : the primary components of functional diversity ». In : *Oikos* 111.1, p. 112–118.

[43] Jost (2010b). « The Relation between Evenness and Diversity », cf. note 22, p. 4.

Schleuter *et al.*[44] font une revue des mesures utilisées dans ce cadre. Les notations des indices utilisées ici sont les leurs.

[44] D. Schleuter et al. (2010). « A user's guide to functional diversity indices ». In : *Ecological Monographs* 80.3, p. 469–484.

Différents traits numériques sont connus pour toutes les espèces de la communauté. La matrice $\boldsymbol{X}$ les contient : l'élément $x_{t,s}$ est la valeur moyenne du trait t pour l'espèce s. On note $\boldsymbol{T}_s$ le vecteur des valeurs moyennes de tous les traits de l'espèce s.

Richesse fonctionnelle

L'étendue fonctionnelle[45] mesure l'étendue des valeurs d'un trait occupée par une communauté, normalisée par l'étendue maximale

[45] Mason et al. (2005). Cf. note 42.

possible.

$$FR_R = \frac{\max_s(x_{t,s}) - \min_s(x_{t,s})}{X_{t,max} - X_{t,min}} \tag{4.1}$$

$\max_s(x_{t,s})$ est la valeur maximale pour toutes les espèces de la matrice de la valeur du trait t, $X_{t,max}$ est sa valeur maximale absolue (les notations sont identiques pour les minima). Les extrêmes absolus peuvent être ceux de l'ensemble des communautés comparées. Ils sont toujours sous-estimés : il est toujours possible théoriquement de trouver des valeurs plus extrêmes en augmentant l'effort d'échantillonnage. L'étendue fonctionnelle peut être moyennée sur plusieurs traits.

Schleuter *et al.* développent l'indice FR_{Is} pour prendre en compte la variabilité intraspécifique et les valeurs de traits non occupées par des espèces dans l'étendue fonctionnelle. Les valeurs de traits individuelles sont nécessaires : l'étendue des valeurs d'un trait est définie comme l'union des étendues des valeurs de chaque espèce.

La définition de la niche écologique de Hutchinson[46] est l'hypervolume, dans l'espace des ressources environnementales, qu'une espèce peut occuper. L'espace des traits fonctionnels peut être considéré comme une approximation de l'espace des ressources. Le volume de l'enveloppe convexe de l'espace des traits occupé par la communauté (*convex hull volume*) est donc une mesure de richesse fonctionnelle multidimensionnelle.

De même que pour l'étendue fonctionnelle, Schleuter *et al.* développent une mesure proche, FR_{Im}, prenant en compte la variabilité intraspécifique et les espaces non occupés. Chaque espèce est supposée occuper un espace autour de sa position moyenne dans l'espace des traits, avec une fonction d'appartenance[47] gaussienne multidimensionnelle (une représentation unidimensionnelle se trouve en figure 4.3) : la fonction d'appartenance, issue de la logique floue, peut être vue comme une densité de probabilité non normalisée.

Les variances et covariances intraspécifiques des traits (rassemblés dans la matrice carrée $\boldsymbol{\Sigma}_s$ de dimension t pour chaque espèce s) sont nécessaires. La fonction d'appartenance de l'espèce s dans l'espace des traits est, pour le vecteur $\boldsymbol{T}$ de valeur de l'ensemble des traits :

$$f_s(\boldsymbol{T}) = e^{-\frac{1}{2}(\boldsymbol{T}-\boldsymbol{T}_s)^\top \boldsymbol{\Sigma}_s^{-1}(\boldsymbol{T}-\boldsymbol{T}_s)} \tag{4.2}$$

La richesse fonctionnelle est l'intégrale des valeurs maximales de f_s dans l'ensemble de l'espace des traits :

$$FR_{Im} = \int \max_s \left(f_s(\boldsymbol{T})\right) d\boldsymbol{T} \tag{4.3}$$

Les indices FAD et MFAD sont des mesures de divergence non pondérées.

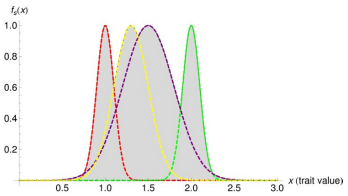


FIGURE 4.3 – Fonction d'appartenance de l'espace des niches de Schleuter *et al.*, en une dimension.[a] x, en ordonnée, est la valeur du trait considéré. Chaque espèce est considérée comme un ensemble flou dans l'espace des traits. Quatre espèces sont représentées, avec leur fonction d'appartenance $f_s(x)$. Le volume de l'espace des traits occupé, FR_{Im}, est obtenu en intégrant les fonctions d'appartenance : c'est la zone grisée de la figure.

[a] Figure issue de SCHLEUTER et al. (2010). Cf. note 44, Appendix B.

[46] G. E. HUTCHINSON (1957). « Concluding remarks ». In : *Cold Spring Harbor Symposia on Quantitative Biology* 2.2, p. 415–427.

[47] L. ZADEH (1965). « Fuzzy sets ». In : *Information and Control* 8.3, p. 338–353.

FAD a été proposé par Walker *et al.*,[48] à partir d'une matrice de distances dans l'espace des traits. $d_{s,t}$ est la distance entre deux espèces indicées par s et t, alors :

$$FAD = \sum_s \sum_t d_{s,t} \tag{4.4}$$

FAD est très sensible au nombre d'espèces. Sa version normalisée, MFAD[49] est :

$$MFAD = \frac{\sum_s \sum_t d_{s,t}}{S} \tag{4.5}$$

Les deux indices violent l'axiome de jumelage. À un facteur de normalisation près, ce sont des cas particuliers de l'indice de Rao pour des effectifs égaux.

Le calcul sous R est simple avec un arbre au format `phylog` d'*ADE4* :

```
Tree <- Paracou618.Taxonomy
(FAD <- sum(Tree$Wdist^2/2))

## [1] 264159

(MFAD <- FAD/length(Tree$leaves))

## [1] 621.5506
```

Après transformation de la matrice de distance en dendrogramme fonctionnel, éventuellement sous la forme d'un arbre-consensus,[50] la longueur totale des branches (FD)[51] est une autre mesure de richesse multidimensionnelle.

Equitabilité fonctionnelle

L'équitabilité fonctionnelle, ou régularité,[52] rend compte de l'homogénéité de l'occupation des niches.

L'indice d'équitabilité de Mouillot *et al.*,[53] que les auteurs nomment « indice de régularité fonctionnelle », est inspiré de l'indice d'équitabilité de Bulla (page 77). C'est un indice unidimensionnel : un seul trait est pris en compte. Les espèces sont classées par valeur croissante du trait. L'équitabilité maximale est obtenue si l'écart entre deux valeurs de traits est proportionnel aux abondances cumulées des deux espèces. La statistique fondamentale est appelée équitabilité pondérée (*weighted evenness*). Pour l'intervalle entre l'espèce s et l'espèce $s+1$:

$$EW_s = \frac{T_{s+1} - T_s}{N_{s+1} + N_s} \tag{4.6}$$

Cette valeur est normalisée : $PEW_s = {}^{EW_s}/_{\sum_s EW_s}$. Sa valeur de PEW_s attendue pour le maximum d'équitabilité est ${}^1/_{(S-1)}$. L'indice est celui de Bulla, appliqué aux $S-1$ intervalles entre

[48] B. Walker et al. (1999). « Plant Attribute Diversity, Resilience, and Ecosystem Function : The Nature and Significance of Dominant and Minor Species ». In : *Ecosystems* 2.2, p. 95–113.

[49] D. Schmera et al. (2009). « A measure for assessing functional diversity in ecological communities ». In : *Aquatic Ecology* 43.1, p. 157–167.

[50] Mouchet et al. (2008). « Towards a consensus for calculating dendrogram-based functional diversity indices », cf. note 26, p. 86.

[51] O. L. Petchey et K. J. Gaston (2002). « Functional diversity (FD), species richness and community composition ». In : *Ecology Letters* 5, p. 402–411.

[52] Pavoine et Bonsall (2011). « Measuring biodiversity to explain community assembly : a unified approach », cf. note 19, p. 4.

[53] D. Mouillot et al. (2005a). « Functional regularity : a neglected aspect of functional diversity ». In : *Oecologia* 142.3, p. 353–359.

espèces :

$$FE_s = \sum_{s=1}^{S-1} \min(PEW_s, \frac{1}{S-1}) \tag{4.7}$$

L'indice FE_s a été étendu pour être multidimensionnel par Villéger *et al.*.[54] L'arbre recouvrant de longueur minimum (*minimum spanning tree*) est d'abord calculé à partir des distances euclidiennes entre les espèces dans l'espace des traits : il s'agit de l'arbre de longueur totale minimale reliant tous les points. La longueur des branches est ensuite traitée de la même façon que δT_s précédemment.

[54] S. Villéger et al. (2008). « New multidimensional functional diversity indices for a multifaceted framework in functional ecology ». In : *Ecology* 89.8, p. 2290–2301.

L'indice Λ^+[55] mesure la variance des distances entre paires d'espèces :

[55] Clarke et Warwick (2001). « A further biodiversity index applicable to species lists : variation in taxonomic distinctness », cf. note 3, p. 9.

$$\Lambda^+ = \frac{\sum_s \sum_t \left(d_{s,t} - \hat{d}\right)^2}{S\,(S-1)} \tag{4.8}$$

Comme le montrent Mérigot et Gaertner,[56] les mesures de régularité n'ont absolument pas les mêmes propriétés que les mesures de diversité : elles peuvent par exemple augmenter quand on retire des espèces originales.

[56] B. Mérigot et J.-C. Gaertner (2011). « Incorporation of phylogeny in biological diversity measurement : Drawbacks of extensively used indices, and advantages of quadratic entropy ». In : *BioEssays* 33.11, p. 819–822.

Divergence fonctionnelle

Les mesures de divergence fonctionnelle décrivent la variabilité de position des espèces dans l'espace des traits. Ce sont tout simplement des mesures de diversité au sens classique du terme.

Quand un seul trait est considéré, la mesure la plus simple est sa variance. Mason *et al.*[57] effectuent une transformation logarithmique de la valeur du trait et pondèrent le calcul de la variance par les probabilités p_s. Ils transforment finalement le résultat en son arc-tangente pour qu'il soit compris entre 0 et 1.

[57] N. W. H. Mason et al. (2003). « An index of functional diversity ». In : *Journal of Vegetation Science* 14.4, p. 571–578.

Schleuter *et al.* proposent d'utiliser la différence entre le troisième et le premier quartile de la valeur du trait, normalisée par son étendue maximale possible.

Kader et Perry[58] regroupent les valeurs de trait par catégories et calculent l'entropie de Simpson des catégories.

[58] G. D. Kader et M. Perry (2007). « Variability for Categorical Variables ». In : *Journal of Statistics Education* 15.2.

Enfin, Villéger *et al.*[59] utilisent la distance euclidienne moyenne des espèces au centre de gravité de la communauté plutôt que la variance. Précisément, le centre de gravité de la communauté (sans pondération par les fréquences) est calculé. La distance euclidienne de l'espèce s au centre de gravité est dG_s ; la moyenne pour toutes les espèces $\bar{dG}$. L'écart moyen des individus de la communauté à la distance moyenne est $\Delta d = \sum_s p_s(dG_s - \bar{dG})$. L'écart moyen absolu est $\Delta|d| = \sum_s p_s|dG_s - \bar{dG}|$

[59] Villéger et al. (2008). Cf. note 54.

L'indice est :

$$FD_m = \frac{\Delta d + \bar{dG}}{\Delta|d| + \bar{dG}} \tag{4.9}$$

Sa forme lui permet d'être compris entre 0 et 1.

4.2.2 Originalité, richesse et équitabilité phylogénétique

La littérature de la diversité phylogénétique s'est intéressée tôt à l'originalité taxonomique des espèces parce que les questions traitées concernaient la conservation (intéressée par la valeur de l'héritage évolutif[60]).

[60] D. P. Faith (2008). « Threatened species and the potential loss of phylogenetic diversity : conservation scenarios based on estimated extinction probabilities and phylogenetic risk analysis. » In : *Conservation biology : the journal of the Society for Conservation Biology* 22.6, p. 1461–70.

Mesures spécifiques d'originalité

Vane-Wright *et al.*[61] définissent la distinction taxonomique (*taxonomic distinctness*) TD_s de chaque espèce comme l'inverse du nombre de nœuds entre elle et la racine de l'arbre phylogénétique, normalisé pour que $\sum_s TD_s = 1$. Dans l'arbre de la figure 4.4, les valeurs de TD non normalisées sont 1/2 pour les deux premières espèces et 1/3 pour les trois autres : la racine de l'arbre est comptabilisée dans le nombre de nœuds. Les valeurs de TD_s sont respectivement 1/4 et 1/6 après normalisation.

[61] Vane-Wright et al. (1991). « What to protect ?—Systematics and the agony of choice », cf. note 6, p. 83.

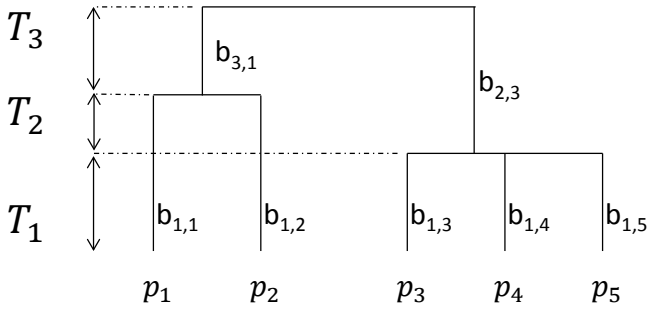


Figure 4.4 – Arbre phylogénétique ou fonctionnel hypothétique. L'arbre comprend 5 espèces dont les probabilités sont notées p_s, 3 périodes de durées T_k délimitées par les nœuds. Les branches sont notées b et indicées par la période à laquelle elles se terminent et un numéro d'ordre. La légende complète se trouve en figure 4.7.

La particularité évolutive[62] (*evolutive distinctiveness*, ED_s) de l'espèce s est la somme de la longueur des branches qui la relient à la racine de l'arbre, partagées entre tous les descendants de chaque branche. Pour l'espèce 3 de la figure 4.4, ED_3 est égal à $l(b_{1,3})$, la longueur de la branche terminale propre à l'espèce 3, plus un tiers de la longueur de la branche qui relie la racine de l'arbre à la polytomie dont l'espèce 3 est issue. Clairement, la diversité phylogénétique PD (section 4.2.3) est la somme des particularités évolutives de toutes les espèces de l'arbre.

[62] N. J. Isaac et al. (2007). « Mammals on the EDGE : Conservation Priorities Based on Threat and Phylogeny ». In : *PLoS ONE* 2.3, e296.

Originalité taxonomique de Ricotta

Ricotta[63] construit un indice paramétrique (permettant de donner plus ou moins d'importance aux espèces rares) à partir de l'espérance du nombre d'espèces tirées dans un échantillon de taille n fixée (l'indice de Hurlbert) :[64]

$$\mathbb{E}\left(S^n\right) = \sum_s \left[1 - (1 - p_s)^n\right] \tag{4.10}$$

[63] C. Ricotta (2004a). « A parametric diversity measure combining the relative abundances and taxonomic distinctiveness of species ». In : *Diversity and Distributions* 10, p. 143–146.

[64] Hurlbert (1971). « The Nonconcept of Species Diversity : A Critique and Alternative Parameters », cf. note 188, p. 53.

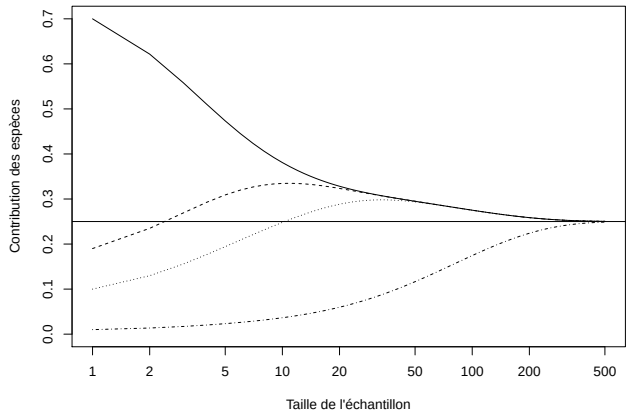


FIGURE 4.5 – Contribution des espèces à l'indice de Hurlbert. Les contributions de 4 espèces d'une communauté à l'indice sont représentées en fonction de la taille de l'échantillon n. Les fréquences des espèces sont lisibles pour $n = 1$: une espèce fréquente (probabilité égale à 0,7, supérieure à $1/S$), deux espèces peu fréquentes (0,19 et 0,1), et une espèce rare (0,01). Quand n est assez grand, toutes les contributions tendent vers $1/S = 1/4$ (ligne horizontale).

Ricotta pondère cette espérance par l'originalité taxonomique de chaque espèce, notée w_s et normalise la mesure :

$$ {}^{n}T = \frac{\sum_s w_s \left[1 - (1 - p_s)^n\right]}{\mathbb{E}\left(S^n\right)} \tag{4.11} $$

L'originalité taxonomique est définie comme la distance phylogénétique moyenne entre l'espèce s et les autres : $w_s = \sum_s d_{s,t}/S-1$.

Weikard *et al.*[65] montrent que cette définition de w_s ne permet pas de satisfaire l'axiome de monotonicité d'ensemble. La définition correcte de $w_s = \sum_s d_{s,t}$, validée par Ricotta.[66]

L'interprétation de ${}^{n}T$ est plus intuitive en inversant la logique de sa construction.

w_s est l'originalité de l'espèce s. $[1-(1-p_s)^n]/\mathbb{E}(S^n)$ est la contribution de l'espèce s à l'espérance du nombre d'espèces, comprise entre 0 et 1. ${}^{n}T$ est donc l'originalité moyenne des espèces de la communauté, pondérée par la contribution de chaque espèce à l'espérance du nombre d'espèces observé, dans un échantillon de taille n. Cette contribution est présentée en figure 4.5 [67]. Quand $n = 1$, c'est simplement la fréquence des espèces. Quand n augmente, le poids des espèces fréquentes (dont la proabilité est supérieure à

[65] H.-P. WEIKARD et al. (2006). « Diversity measurement combining relative abundances and taxonomic distinctiveness of species ». In : *Diversity & Distributions* 12.2, p. 215–217.

[66] C. RICOTTA (2006). « Strong requirements for weak diversities ». In : *Diversity and Distributions* 12.2, p. 218–219.

[67] Code R pour la figure :

```
Ps <- c(0.7, 0.19, 0.1, 0.01)
S <- length(Ps)
nRange <- 1:500
# Indice de Hurlbert
ESn <- c(1, sapply(nRange[-1], function(n) Hurlbert(Ps, n)))
# Préparation du graphique
plot(x=nRange, y=c(rep(0,length(nRange)-1),max(Ps)), type = "n",
     xlab="Taille de l'échantillon", ylab="Contribution des espèces",
     log="x")
# Tracé des courbes
sapply(1:S, function(s) lines(nRange, sapply(nRange,
              function(n) (1-(1-Ps[s])^n))/ESn, lty=s))
abline(h=1/(S))
```

$1/S$) diminue alors que celui des espèces intermédiaires augmente. Ce dernier atteint $1/\mathbb{E}(S^n)$ (une espèce est échantillonnée à coup sûr dès que l'échantillon est assez grand) d'autant plus rapidement que p_s est grand. Il baisse ensuite alors que les espèces rares atteignent à leur tour progressivement leur poids maximal.

Quand $n \to +\infty$, le numérateur tend vers FAD, et le dénominateur tend vers le nombre d'espèces : ${}^nT$ tend vers MFAD.

Richesse phylogénétique

La particularité taxonomique moyenne[68] (*Average Taxonomic Distinctiveness, AvTD*) est la distance moyenne dans l'arbre entre deux espèces choisies au hasard. C'est donc l'équivalent taxonomique de MFAD. La variabilité phylogénétique des espèces[69] (*Phylogenetic Species Variability, PSV*) est la même mesure, obtenue à partir du modèle d'évolution suivant un mouvement brownien.[70] La distance moyenne entre deux espèces est dans ce cadre proportionnelle à la covariance de leurs traits fonctionnels. La richesse phylogénétique des espèces, PSR, est PSV multipliée par le nombre d'espèces : c'est l'équivalent taxonomique de FAD.

La diversité fonctionnelle FD de Faith (section 4.2.3) semble avoir fait consensus, d'où un moindre développement des mesures de richesse que dans la littérature fonctionnelle.

Indices de Cadotte

Cadotte *et al.*[71] proposent un ensemble de mesure phylogénétiques : équitabilité, déséquilibre d'abondance et diversité de la particularité evolutive.

Equitabilité phylogénétique Les branches terminales de l'arbre phylogénétique sont ici à la base de la définition de l'équitabilité. Figure 4.6, les branches $b_{1,s}$ sont les segments terminés par une feuille (au bas de l'arbre), c'est-à-dire que leur longueur est la partie de l'arbre que les espèces ne partagent pas. La mesure d'équitabilité est :

$$PAE = \frac{PD + \sum_s l(b_{1,s})(n_s - 1)}{PD + (\frac{n}{S} - 1)\sum_s l(b_{1,s})} \tag{4.12}$$

PAE vaut 1 quand les espèces sont sont distribuées équitablement pour la longueur des branches. Une valeur plus grande est obtenue quand les espèces sont concentrées au bout des longues branches, et inversement si $0 < PAE < 1$.

Déséquilibre d'abondance L'équilibre d'abondance est défini par une distribution aléatoire des espèces par division de l'effectif total à partir de la racine de l'arbre phylogénétique. En figure 4.6, en partant de la racine, la moitié des individus est supposée se

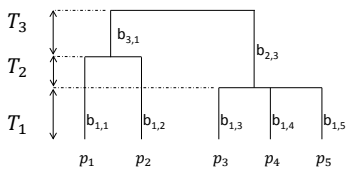


FIGURE 4.6 – Arbre phylogénétique ou fonctionnel hypothétique. Rappel de la figure 4.7(a).

[68] WARWICK et CLARKE (1995). « New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress », cf. note 7, p. 83.

[69] M. R. HELMUS et al. (2007). « Phylogenetic measures of biodiversity ». In : *The American naturalist* 169.3, E68–83.

[70] FELSENSTEIN (1985). « Phylogenies and the Comparative Method », cf. note 30, p. 86.

[71] M. W. CADOTTE et al. (2010). « Phylogenetic diversity metrics for ecological communities : integrating species richness, abundance and evolutionary history ». In : *Ecology Letters* 13.1, p. 96–105.

répartir sur chaque branche. À la période 2, les effectifs de la branche de gauche se partagent en deux parties égales. À partir du nombre total d'individus de la communauté, n, le nombre attendu d'individus n_s^0 de l'espèce s est une fraction de n correspondant au nombre de nœuds et au nombre de branches partant de chaque nœud. On note $y_{k,s}$ le nombre de branches partant du nœud ancestral de l'espèce s à la période k, s'il existe, $y_{k,s} = 1$ sinon :

$$n_s^0 = \frac{n}{\prod_{k=2}^{K}} \tag{4.13}$$

En figure 4.6, pour l'espèce 3, $y_{1,3} = 3$ (l'espèce est issue d'une polytomie), $y_{2,3} = 1$ (absence de nœud) et $y_{3,3} = 2$ (la racine de l'arbre est dichotomique). On s'attend donc à ce que l'espèce 3 soit représentée par un sixième des individus.

L'indice de déséquilibre d'abondance mesure l'écart à cette distribution théorique :

$$IAC = \frac{\sum_s |n_s - n_s^0|}{\nu} \tag{4.14}$$

ν est le nombre de nœuds de l'arbre (3 dans l'exemple de la figure 4.6).

Diversité de la particularité évolutive L'entropie de Shannon peut être appliquée pour mesurer la diversité des particularités évolutives :

$$H_{ED} = -\sum_s \frac{ED_s}{PD} \ln \frac{ED_s}{PD} \tag{4.15}$$

L'équitabilité des particularités évolutives est celle de Pielou :

$$E_{ED} = \frac{H_{ED}}{\ln S} \tag{4.16}$$

Ces mesures ne prennent pas en compte les abondances des espèces. Pour y remédier, il suffit d'ajouter une période de durée nulle à chaque feuille de l'arbre, correspondant à une polytomie entre les n_s individus de chaque espèce. La particularité évolutive AED_s des individus de l'espèce s, est calculée en partageant la longueur de chaque branche ancestrale entre le nombre d'individus qui en descendent (et non le nombre d'espèces). Alors $PD = \sum_s n_s AED_s$. L'entropie devient :

$$H_{AED} = -\sum_s \frac{n_s AED_s}{PD} \ln \frac{n_s AED_s}{PD} \tag{4.17}$$

Et l'équitabilité correspondante est :

$$E_{AED} = \frac{H_{AED}}{\ln N} \tag{4.18}$$

Ces indices ne mesurent pas la diversité phylogénétique au sens des autres mesures présentées dans ce chapitre. Ils mesurent la diversité de le particularité évolutive, autrement dit du temps d'évolution accumulé par chaque espèce ou chaque individu. Pour un nombre d'espèces fixé, E_{ED} atteint son maximum quand les valeurs de ED_s sont toutes identiques. Un arbre phylogénétique composé d'une seule branche depuis la racine terminé par une polytomie de longueur nulle portant toutes les espèces correspond à cette description. Dans cet exemple, les espèces ont une divergence évolutive nulle, mais $E_{ED} = S$, son maximum possible.

4.2.3 FD et PD

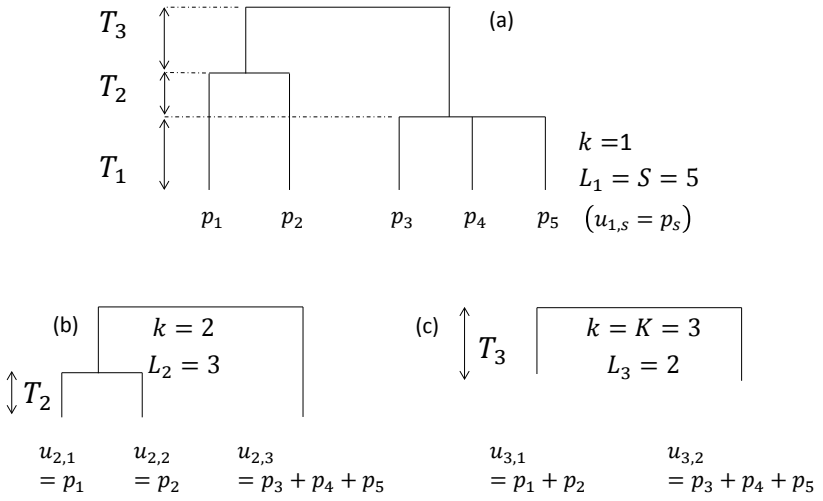


FIGURE 4.7 – Arbre phylogénétique ou fonctionnel hypothétique. (a) Arbre complet. 5 espèces sont présentes ($S = 5$). Une période de l'arbre est définie entre deux nœuds successifs : l'arbre contient $K = 3$ périodes. Les hauteurs des périodes sont notées T_k. À chaque période correspond un arbre plus simple (b pour la période 2, c pour la période 3) dans lequel les espèces originales sont regroupées. Le nombre de feuilles de ces arbres est noté L_k. Les probabilités pour un individu d'appartenir à une feuille sont notées $u_{k,l}$.

Si la dissimilarité entre les espèces est représentée par un dendrogramme, les indices de diversité les plus simples sont la diversité phylogénétique[72] et sa transposition, la diversité fonctionnelle.[73]

Étant donné un arbre contenant toutes les espèces ou tous les individus étudiés, PD ou FD sont égaux à la somme de la longueur des branches (Figure 4.7 : $PD = 5 \times T_1 + 3 \times T_2 + 2 \times T_3$).

Dans le cas particulier ou toutes les branches sont de longueur 1, c'est-à-dire que toutes les espèces sont liées à la même racine (on dira que l'arbre est parfaitement régulier), PD et FD sont égales à la richesse spécifique.

Le package *entropart* fournit la fonction `PDFD` :

```
# Vecteur des probabilités
Ps <- Paracou618.MC$Ps
PDFD(Ps, Paracou618.Taxonomy)

## None
##  395
```

Chao *et al.*[74] fournissent un estimateur de PD (ou FD) permettant de l'estimer sans biais pour un échantillon plus petit que celui observé et de façon fiable pour un échantillon de taille double

[72] D. P. FAITH (1992). « Conservation evaluation and phylogenetic diversity ». In : *Biological Conservation* 61.1, p. 1–10.

[73] PETCHEY et GASTON (2002). « Functional diversity (FD), species richness and community composition », cf. note 51, p. 90.

[74] A. CHAO et al. (2015a). « Rarefaction and extrapolation of phylogenetic diversity ». In : *Methods in Ecology and Evolution* 6.4, p. 380–388.

au maximum. Ces estimateurs permettent de tracer des courbes de raréfaction et d'extrapolation.

En même temps que Faith, Weitzman[75] a établi une fonction de diversité identique à PD. À partir d'une matrice de dissimilarités entre espèces, un arbre est construit par la méthode suivante. Les deux espèces les plus proches sont rassemblées. L'une d'elles, celle qui diminue le moins la longueur totale des branches de l'arbre final en la retirant, est l'espèce de *lien*, l'autre est l'espèce *représentative* du clade. L'espèce de lien est retirée et le regroupement poursuivi entre les deux nouvelles espèces les plus proches. À chaque étape, l'espèce de lien est retirée jusqu'au regroupement final entre les deux dernières espèces. La difficulté est que l'arbre final n'est pas connu aux premiers stades de regroupement donc tous les arbres doivent être testés pour trouver la solution unique. La diversité est la somme des distances entre les espèces de lien et leur espèce représentative, c'est-à-dire la longueur totale des branches de l'arbre obtenu. L'application à la conservation est que les espèces représentatives doivent être favorisées par rapport aux espèces de lien ; en d'autres termes, les espèces dont les branches sont les plus courtes dans l'arbre sont celles qui apportent le moins de diversité. Ce résultat est assez évident quand on dispose d'un arbre phylogénétique. L'originalité de la méthode de Weitzman est de fournir un algorithme pour créer l'arbre à partir d'une matrice de distances qui est cohérent avec la mesure de diversité appliquée.

[75] Weitzman (1992). « On Diversity », cf. note 41, p. 88.

La mesure de diversité peut être combinée avec une probabilité d'extinction de chaque espèce (à dire d'expert) pour calculer l'espérance de la diversité à une échéance donnée[76] : 2^S arbres peuvent être construit en faisant disparaître certaines des S espèces de la communauté, la probabilité de chaque arbre est calculable à partir des probabilités d'extinction de chaque espèce (supposées indépendantes) et l'espérance de diversité est simplement la moyenne des diversités de chaque arbre pondérée par sa probabilité. L'espérance de la perte de diversité peut être calculée de façon plus simple,[77] comme la moyenne de la longueur des branches pondérée par la probabilité de leur disparition, qui est le produit de la probabilité de disparition de toutes les espèces descendant de la branche.

[76] M. L. Weitzman (1993). « What to Preserve ? An Application of Diversity Theory to Crane Conservation ». In : *The Quarterly Journal of Economics* 108.1, p. 157–183.

[77] L. Witting et V. Loeschcke (1995). « The optimization of biodiversity conservation ». In : *Biological Conservation* 71.2, p. 205–207.

Diverses optimisations économiques sont possibles, dont le choix des espèces à protéger à partir de l'élasticité de la diversité, c'est-à-dire le gain de diversité entraîné par la diminution de la probabilité de disparition de chaque espèce : les espèces ayant la plus grande élasticité sont celles sur lesquelles les efforts auront les plus grands résultats.

4.2.4 Indice de Rao

L'indice de Rao est une mesure de divergence pondérée. Son utilisation s'est largement développée depuis le début des années

2000[78] en raison de ses propriétés particulièrement intéressantes.

Principe

À partir de relevés fournissant la fréquence de chaque espèce par communauté et d'une matrice de dissimilarité entre paires d'espèces, l'indice de Rao[79] donne la dissimilarité moyenne entre deux individus choisis au hasard.

L'indice de Rao est souvent appelé « entropie quadratique » en raison de sa forme mathématique.

Formalisation

Les espèces sont prises deux à deux et sont donc notées ici s' et s''.

On note Δ la matrice de dissimilarité dont les éléments sont $d_{s's''}$, la dissimilarité entre l'espèce s' et l'espèce s''. Il n'est pas nécessaire à ce stade que Δ soit une distance. $\mathcal{P}$ est le vecteur des probabilités dont $p_{s'}$ et $p_{s''}$ dont des éléments :

L'indice de Rao est :

$$H_{\Delta}(\mathcal{P}) = \sum_{s'} \sum_{s''} p_{s'} p_{s''} d_{s's''} \tag{4.19}$$

Discussion

La définition de la distance est essentielle :

- en fixant $d_{s's''} = 1$ si deux espèces sont différentes, on obtient l'indice de Simpson. Sa valeur peut être interprétée comme la probabilité qu'une paire d'individus choisie au hasard soit de deux espèces différentes ;
- Dans un espace unidimensionnel où la valeur y_s associée à l'espèce s est une variable quantitative Y, choisir $d_{s's''} = (y_k - y_l)^2/2$ rend l'indice de Rao égal à la variance de Y.

Pavoine *et al.*[80] ont montré que l'utilisation de distances ordinaires fait que la valeur maximale de l'entropie quadratique pour un effectif donné est obtenue en éliminant les espèces intermédiaires en ne retenant que les espèces extrêmes (le résultat est évident en une dimension : la variance est maximale en ne retenant que les valeurs extrêmes d'un échantillon). Ce résultat est contraire aux propriétés attendues d'un indice de diversité. Les auteurs ont établi que l'utilisation de distances ultramétriques corrige ce défaut. L'indice atteint alors son maximum pour des fréquences d'autant plus grandes que l'espèce est originale.[81]

L'estimation empirique de l'indice se fait simplement en estimant les probabilités par les fréquences. Le biais d'estimation est très faible[82] : par analogie avec l'estimateur de l'indice de Simpson, les espèces rares interviennent peu.

[78] J. Izsák et L. Papp (2000). « A link between ecological diversity indices and measures of biodiversity ». In : *Ecological Modelling* 130, p. 151–156 ; K. Shimatani (2001b). « On the measurement of species diversity incorporating species differences ». In : *Oikos* 93.1, p. 135–147 ; A. Escalas et al. (2013). « A unifying quantitative framework for exploring the multiple facets of microbial biodiversity across diverse scales ». In : *Environmental Microbiology* 15.10, p. 2642–2657.

[79] Rao (1982). « Diversity and dissimilarity coefficients : a unified approach », cf. note 5, p. 83.

[80] S. Pavoine et al. (2005b). « Measuring, diversity from dissimilarities with Rao's quadratic entropy : Are any dissimilarities suitable ? » In : *Theoretical Population Biology* 67.4, p. 231–239.

[81] Pavoine et al. (2005a). « Is the originality of a species measurable ? », cf. note 29, p. 86.

[82] Marcon et al. (2014a). « Generalization of the Partitioning of Shannon Diversity », cf. note 38, p. 64.

Calcul sous R

Le package *entropart* fournit la fonction `Rao` :

```
# Vecteur des probabilités
Ps <- Paracou618.MC$Ps
Rao(Ps, Paracou618.Taxonomy)

##     None
## 2.878133
```

Le package *ADE4* permet le calcul avec la fonction `divc` qui utilise un format différent pour les probabilités et surtout la matrice des racines carrées du double des distances de l'objet `phylog` (chaque élément de `$Wdist` vaut $\sqrt{2d_{s's''}}$, c'est-à-dire la distance euclidienne entre les espèces dans le cadre d'une représentation géométrique[83]) au lieu de l'arbre lui-même :

[83] S. Champely et D. Chessel (2002). « Measuring biological diversity using Euclidean metrics ». In : *Environmental and Ecological Statistics* 9.2, p. 167–177.

```
library(ade4)
divc(as.data.frame(Ps), Paracou618.Taxonomy$Wdist)

##    diversity
## Ps  2.878133
```

`divc` peut traiter plusieurs communautés simultanément ; `Rao` peut être utilisé avec la fonction `apply` :

```
divc(as.data.frame(Paracou618.MC$Psi), Paracou618.Taxonomy$Wdist)

##      diversity
## P006  2.820856
## P018  2.876405

apply(Paracou618.MC$Psi, 2, Rao, Tree = Paracou618.Taxonomy)

##     P006     P018
## 2.820856 2.876405
```

Maximum théorique

Pavoine *et al.*[84] ont défini l'originalité d'une espèce comme sa fréquence maximisant l'entropie quadratique, sachant la matrice de distances entre espèces. Les espèces les plus originales sont celles ayant le moins d'espèces proches dans la classification.

[84] Pavoine et al. (2005a). « Is the originality of a species measurable ? », cf. note 29, p. 86.

L'originalité n'est pas intéressante dans une taxonomie : une phylogénie doit être créée pour illustrer cette notion. Le fichier `rao.traits.csv` contient une espèce par ligne, identifiée par le champ `Code`, et un certain nombre de valeurs de traits en colonnes.

```
# Lecture des données : traits pour 34 espèces
traits <- read.csv2("http://www.ecofog.gf/IMG/csv/rao.traits.csv",
    row.names = 1, header = T)
```

Le résultat est un *data frame* nommé `traits` à 8 lignes (espèces) et 6 colonnes (traits) :

```
# Aperçu
traits[1:4, 1:3]

##          SLA       Am       Gm
## Ess 10.10341 97.88627 1.454455
## Me  14.27473 63.52255 1.493502
## S1  13.71211 73.57748 1.963457
```

```
## Sr  10.71081 68.03988 1.305796
```

La matrice de distances est créée par classification automatique hiérarchique.

```
# ACP sur les traits foliaires
pcaf <- dudi.pca(traits, scale = T, scannf = FALSE, nf = 2)
```

`pcaf` (Figure 4.8 [85]) est une liste qui contient les résultats de l'ACP, à utiliser pour la classification :

```
# CAH Ward des traits foliaires
hf <- hclust(dist(pcaf$tab), "ward.D")
```

Le résultat de la classification est un objet `hclust` (Figure 4.9 [86] qui doit être transformé en `phylog` pour la suite de l'analyse :

```
# Transformation de l'arbre du format hclust au format phylog
phyf <- hclust2phylog(hf)
```

Le résultat est en Figure 4.10 [87].

La limite des distances ultramétriques est leur tendance à déformer le jeu de points.[88] Dans cet exemple, les deux premiers axes de l'ACP rendent compte de presque toute l'inertie. Le nuage de points est pratiquement contenu dans un plan alors que sa représentation en distance ultramétrique est une hypersphère en 7 dimensions.

Le calcul de l'originalité des espèces utilise la fonction `originality` (Figure 4.11 [89]).

La fonction a pour paramètres l'objet `phylog` contenant la classification et le numéro de la méthode de calcul à utiliser, 5 pour l'entropie quadratique. Sa représentation graphique est faite par `dotchart.phylog` :

L'originalité ne repose que sur l'arbre, pas sur la fréquence des espèces.

Si la distance utilisée n'est pas ultramétrique, il existe plusieurs distributions possibles d'espèces qui maximisent la diversité,[90] le concept d'originalité n'a pas de sens dans ce cas.

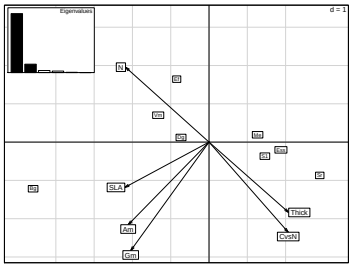


FIGURE 4.8 – Analyse en composante principale des traits foliaires

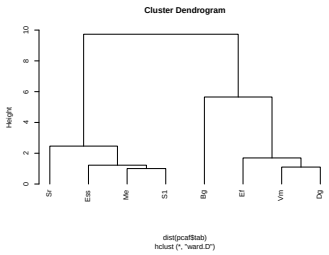


FIGURE 4.9 – Arbre phylogénétique issu de la classification automatique.

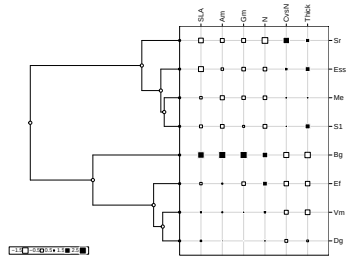


FIGURE 4.10 – Présentation de l'arbre phylogénétique avec la contribution de chacune des variables.

[88] PAVOINE et al. (2005a). « Is the originality of a species measurable ? », cf. note 29, p. 86.

[85] Code R :

```
scatter(pcaf)
```

[86] Code R :

```
plot(hf, h = -1)
```

[87] Code R :

```
table.phylog(pcaf$tab[names(phyf$leaves), ], phyf)
```

[89] Code R :

```
dotchart.phylog(phyf, originality(phyf, 5))
```

4.2.5 Diversité et moyenne

Garnier *et al.*[91] définissent la moyenne pondérée d'un trait à l'échelle de la communauté (CWM : *Community Weigthed Mean*), simplement égal à la moyenne de la valeur du trait pondérée par la fréquence des espèces :

$$CWM = \sum_s p_s y_s \tag{4.20}$$

CWM n'est pas une mesure de diversité fonctionnelle, bien qu'il ait été utilisé parfois en tant que tel,[92] mais une mesure de la composition de la communauté. CWM peut être étendu à plusieurs traits : le vecteur formé par les valeurs de CWM pour chaque trait correspond au centre de gravité de la communauté représentée dans l'espace des traits. L'entropie quadratique de Rao mesure la dispersion des espèces autour de ce point : les deux mesures se complètent donc.[93]

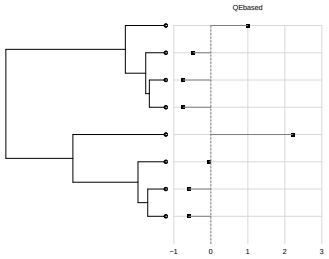


FIGURE 4.11 – Originalité des espèces.

[90] S. PAVOINE et M. B. BONSALL (2009). « Biological diversity : Distinct distributions can lead to the maximization of Rao's quadratic entropy ». In : *Theoretical Population Biology* 75.2-3, p. 153–163.

[91] E. GARNIER et al. (2004). « Plant functional markers capture ecosystem properties during secondary succession ». In : *Ecology* 85.9, p. 2630–2637.

[92] S. LAVOREL et al. (2008). « Assessing functional diversity in the field - Methodology matters ! » In : *Functional Ecology* 22, p. 134–147.

[93] C. RICOTTA et M. MORETTI (2011). « CWM and Rao's quadratic diversity : A unified framework for functional ecology ». In : *Oecologia* 167, p. 181–188.

4.2.6 H_p et I_1

Simultanément, Pavoine *et al.*[94] et Allen *et al.*[95] ont proposé la généralisation de l'indice de Shannon à la diversité phylogénétique. La présentation de Allen *et al.* est donnée ici, celle de Pavoine *et al.*, plus générale, sera détaillée dans le paragraphe suivant.

Étant donné une phylogénie, comme celle de la Figure 4.12, on définit une branche $b_{k,l}$ comme un segment terminé par une feuille (au bas de l'arbre) ou un nœud (dans l'arbre). k indique la période de l'arbre à laquelle la branche se termine, l le numéro d'ordre Une branche n'existe que si elle se termine effectivement par un nœud ou une feuille : il n'y a pas de branche $b_{2,1}$ par exemple. Sa probabilité $p(b_{k,l})$ est la somme des probabilités des feuilles de la branche, c'est-à-dire $u_{k,l}$, et $l(b)$ est sa longueur. Sur la figure, la branche commençant en haut de l'arbre et se terminant au nœud réunissant les espèces 3 à 5 a une valeur $p(b_{2,3})$ égale à la somme des probabilités d'occurrence des espèces 3 à 5 alors que $p(b_{1,1})$ est seulement la probabilité de l'espèce 1. Leurs longueurs respectives sont $T_2 + T_3$ et $T_1 + T_2$. L'arbre possède 7 branches.

L'indice d'entropie phylogénétique est :

$$H_p = -\sum_b l(b)p(b)\ln p(b) \tag{4.21}$$

Dans un arbre parfaitement régulier, toutes les branches sont de longueur 1 et H_p est l'indice de Shannon.

[94] S. PAVOINE et al. (2009a). « Hierarchical partitioning of evolutionary and ecological patterns in the organization of phylogenetically-structured species assemblages : Application to rockfish (genus : Sebastes) in the Southern California Bight ». In : *Ecology Letters* 12.9, p. 898–908.

[95] B. ALLEN et al. (2009). « A New Phylogenetic Diversity Measure Generalizing the Shannon Index and Its Application to Phyllostomid Bats ». In : *American Naturalist* 174.2, p. 236–243.

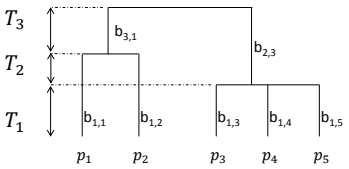


FIGURE 4.12 – Arbre phylogénétique ou fonctionnel hypothétique. Rappel de la figure 4.7(a).

4.3 Entropie phylogénétique

4.3.1 Généralisation de l'entropie HCDT

Pavoine *et al.*[96] découpent l'arbre phylogénétique en périodes. À partir de la racine de l'arbre, une nouvelle période est définie à chaque ramification d'une branche quelconque. Les débuts et fins de périodes sont notés t_k, la racine de l'arbre est fixée à $t_0 = 0$. L'arbre est ultramétrique.

Nous suivrons plutôt les notations de Chao *et al.*[97] en numérotant les périodes à partir du présent et en notant T_k leur durée. Figure 4.7, la première période se termine quand les branches des espèces 3 à 5 se rejoignent. L'arbre comprend $K = 3$ périodes.

L'entropie HCDT (^qH de l'équation (3.6)) est calculée à chaque période. Figure 4.7, à la deuxième période (T_2), l'arbre a trois feuilles, avec des probabilités égales à celle des espèces 1 et 2 et la somme de celles des espèces 3 à 5. ^qH peut être calculée avec ces valeurs de probabilités. On notera cette valeur d'entropie ${}^q_kH$ où k est le quantième de la période.

L'indice I_q de Pavoine *et al.* est la somme des ${}^q_kH$ pondérée par la durée de chaque période. Nous le normalisons par la hauteur totale de l'arbre (T) pour définir $^q\bar{H}(T)$:

$$^q\bar{H}\,(T) = \sum_{k=1}^{K} \frac{T_k}{T}\, {}^q_kH \tag{4.22}$$

Dans un arbre parfaitement régulier, toutes les branches sont de longueur 1, il n'y a qu'une seule période, et $I_q = {}^qH$.

Shimatani,[98] puis Ricotta[99] avaient montré que l'indice de Rao est la somme pondérée sur chaque période de l'indice de Simpson, c'est-à-dire l'égalité (4.22) pour le cas particulier $q = 2$.

Les indices $^q\bar{H}(T)$ généralisent les mesures d'entropie classique à la diversité phylogénétique : $T[^0\bar{H}(T) + 1]$ est égal à PD ou FD, $^1\bar{H}(T)$ est H_p et $^2\bar{H}(T)$ est l'indice de Rao. On peut les interpréter intuitivement comme une somme pondérée par la longueur des périodes des valeurs de l'entropie à chaque période. À la dernière période (près des feuilles), toutes les classes sont présentes, la diversité est donc maximale. En remontant dans l'arbre, les classes se confondent et la diversité diminue progressivement. Deux classes peu distantes, comme les espèces 3 à 5 de la Figure 4.7, apportent peu de diversité supplémentaire par rapport à une situation où les deux espèces seraient confondues (et leurs effectifs ajoutés), contrairement aux espèces 1 et 2.

4.3.2 Biais d'estimation

La correction du biais d'estimation applicable à l'entropie généralisée l'est aussi à chaque période de l'arbre. Les biais sont de

[96] Pavoine et Bonsall (2009). « Biological diversity : Distinct distributions can lead to the maximization of Rao's quadratic entropy », cf. note 90, p. 101.

[97] Chao et al. (2010). « Phylogenetic diversity measures based on Hill numbers », cf. note 82, p. 27.

[98] K. Shimatani (2001a). « Multivariate point processes and spatial variation of species diversity ». In : *Forest Ecology and Management* 142.1-3, p. 215–229.

[99] C. Ricotta (2005b). « Additive partitioning of Rao's quadratic diversity : a hierarchical approach ». In : *Ecological Modelling* 183.4, p. 365–371.

moins en moins importants quand on se rapproche de la racine de l'arbre (il est de moins en moins probable de ne pas observer une classe quand les classes sont de plus en plus vastes).

Le package *entropart* fournit la fonction `PhyloEntropy` pour calculer ${}^{q}\bar{H}(T)$ à partir d'un vecteur de probabilité ou d'abondances. Dans le dernier cas, la correction de `Tsallis` est appliquée à chaque période.

Exemple : le package contient les données d'inventaire de deux hectares du dispositif de Paracou et la taxonomie des espèces concernées. Le calcul de l'indice de Rao, corrigé du biais d'estimation utilise le code suivant :

```
data(Paracou618)
PhyloEntropy(Paracou618.MC$Ns, 2, Paracou618.Taxonomy)

## $Distribution
## [1] "NorP"
##
## $Function
## [1] "PhyloEntropy"
##
## $Tree
## [1] "Tree"
##
## $Normalized
## [1] TRUE
##
## $Cuts
##         1         2         3
## 0.9863260 0.9709804 0.9233893
##
## $Total
## [1] 0.9602319
##
## $Type
## [1] "alpha or gamma"
##
## $Order
## [1] 2
##
## $Correction
## [1] "Best"
##
## attr(,"class")
## [1] "PhyloEntropy" "PhyloValue"
```

La fonction retourne la valeur de ${}^{q}_{k}H$ dans chaque intervalle (l'arbre est ici une taxonomie espèce-genre-famille, correspondant aux limites des périodes (`Cuts`) 1, 2 et 3 dont l'entropie va en décroissant) et la valeur de ${}^{q}\bar{H}(T)$.

4.3.3 Entropie et diversité

La diversité phylogénétique est le nombre d'espèces équifréquentes et dont la distance à toutes les autres dans la phylogénie est maximale, dont l'entropie serait égale à l'entropie observée.

L'entropie ${}^{q}\bar{H}(T)$ peut être transformée en diversité[100] de la même façon que ${}^{q}D = e^{{}^{q}H}$:

[100] Marcon et al. (2014a). « Generalization of the Partitioning of Shannon Diversity », cf. note 38, p. 64.

$$ {}^{q}\bar{D}\,(T) = e_q^{{}^{q}\bar{H}\,(T)} \tag{4.23} $$

Le nombre effectif d'espèces de l'entropie de Rao, ${}^{2}\bar{D}(T) = {}^{1}/_{[1-{}^{2}\bar{H}\,(T)]}$ a été établi par Ricotta et Szeidl.[101]

Chao *et al.*[102] obtiennent ce résultat sans recourir explicitement à l'entropie, mais en faisant le même calcul :

$$ {}^{q}\bar{D}\,(T) = \left(\sum_{i \in \boldsymbol{B}_T} \frac{T_i}{T} p_i^q \right)^{\frac{1}{1-q}} \tag{4.24} $$

$\boldsymbol{B}_T$ est l'ensemble des branches de l'arbre, défini de façon identique à celle de Allen *et al.*

Les entropies ont un comportement linéaire : elles s'additionnent tout au long de l'arbre pour donner ${}^{q}\bar{H}\,(T)$. Les diversités ${}^{q}_{k}D$ calculées à chaque période ne peuvent pas être sommées sur le modèle de l'équation (4.22) : ${}^{q}\bar{D}(T)$ n'est *pas* la moyenne pondérée des diversités aux différentes périodes, sauf dans le cas particulier $q = 1$ où, comme pour la décomposition de l'indice de Shannon, il en est la moyenne géométrique pondérée.

La fonction `PhyloDiversity` de *entropart* permet de calculer ${}^{q}\bar{D}(T)$ avec ou sans correction de biais, selon qu'elle traite un vecteur d'abondances ou de probabilités. À la suite de l'exemple précédent, les résultats sont présentés sous forme de diversité au lieu d'entropie :

```
PhyloDiversity(Paracou618.MC$Ns, 2, Paracou618.Taxonomy)

## $Distribution
## [1] "NorP"
##
## $Function
## [1] "bcPhyloDiversity"
##
## $Tree
## [1] "Tree"
##
## $Normalized
## [1] TRUE
##
## $Cuts
##        1        2        3
## 73.13163 34.45951 13.05301
##
## $Total
## [1] 25.1458
##
## $Type
## [1] "alpha or gamma"
##
## $Order
## [1] 2
##
## $Correction
## [1] "Best"
##
## attr(,"class")
## [1] "PhyloDiversity" "PhyloValue"
```

[101] C. Ricotta et L. Szeidl (2009). « Diversity partitioning of Rao's quadratic entropy ». In : *Theoretical Population Biology* 76.4, p. 299–302.

[102] Chao et al. (2010). « Phylogenetic diversity measures based on Hill numbers », cf. note 82, p. 27.

4.3.4 Diversité individuelle

La construction de la diversité fonctionnelle ou phylogénétique n'implique pas de regrouper les individus par espèces : chaque catégorie peut se réduire à un individu si des données individuelles moléculaires ou de traits sont disponibles.

Le regroupement des individus en une espèce revient simplement à considérer leur distance comme nulle. Diviser une espèce en deux espèces infiniment proches revient seulement à créer une période supplémentaire dans l'arbre, de longueur infinitésimale ; en d'autres termes, la mesure de diversité est continue face au regroupement. Cette propriété permet de limiter les conséquences du problème de l'espèce : séparer les individus d'une espèce en deux espèces proches dans l'arbre n'a que peu d'effet sur la diversité.

4.3.5 Arbres non ultramétriques

Chao *et al.*[103] définissent la diversité phylogénétique selon l'équation (4.24) quelle que soit la forme de l'arbre, y compris s'il n'est pas ultramétrique. Dans ce cas, T est remplacé par $\bar{T}$, la longueur moyenne des branches pondérée par la fréquence des espèces. Cette généralisation est très discutable : son sens n'est pas clair sur le plan de la mesure de la diversité au-delà du parallélisme de la forme mathématique.

L'arbre de la Figure 4.13 peut être découpé en périodes mais les deux premières (T_1 : seule l'espèce 2 est présente ; T_2 : les espèces 1 et 2 sont présentes) sont incomplètes au sens où la somme des probabilités n'y est pas égale à 1 donc $(\sum p_i^q)^{\frac{1}{1-q}}$ ne définit pas une diversité à ces périodes.

Pavoine et Bonsall[104] traitent en détail les résultats aberrants que cause un arbre non ultramétrique dans le cas particulier de l'entropie de Rao. Leinster et Cobbold[105] montrent qu'un arbre non ultramétrique implique que la dissimilarité entre les espèces dépende de leur fréquence, ce qui est contradictoire avec le cadre dans lequel la diversité phylogénétique a été définie.

Dans l'état actuel des connaissances, aucune méthode n'est applicable de façon satisfaisante aux arbres non ultramétriques.

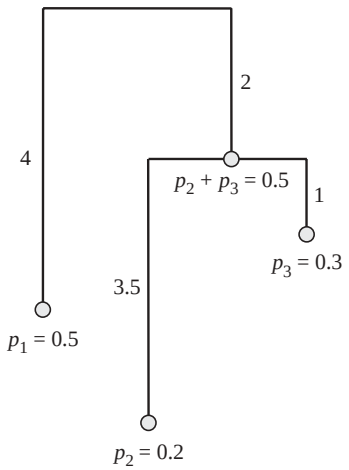


FIGURE 4.13 – Arbre phylogénétique hypothétique non ultramétrique.[a]

[a] Ibid., figure 1b.

4.4 Diversité de Leinster et Cobbold

4.4.1 Définitions

Leinster et Cobbold[106] proposent une unification des mesures de diversité à partir de la définition de la banalité des espèces. Une matrice carrée de dimension égale au nombre d'espèces, $\boldsymbol{Z}$, décrit par ses valeurs $z_{s,t}$ la similarité entre l'espèce s et l'espèce t comprises entre 0 et 1 ($z_{s,s} = 1$).

[103] CHAO et al. (2010). « Phylogenetic diversity measures based on Hill numbers », cf. note 82, p. 27.

[104] PAVOINE et BONSALL (2009). « Biological diversity : Distinct distributions can lead to the maximization of Rao's quadratic entropy », cf. note 90, p. 101.

[105] LEINSTER et COBBOLD (2012). « Measuring diversity : the importance of species similarity », cf. note 191, p. 54.

[106] Ibid.

La banalité d'une espèce s est définie par $\sum_t p_t z_{s,t}$, c'est-à-dire la moyenne pondérée de sa similarité avec toutes les autres espèces : au minimum p_s si elle est totalement différente des autres (approche de la diversité neutre, $\boldsymbol{Z}$ est la matrice identité $\boldsymbol{I}$), au maximum 1 si toutes les espèces sont totalement similaires ($\boldsymbol{Z}$ ne contient que des 1). Une espèce rare totalement différente des autres est peu banale.

La moyenne généralisée d'ordre r,[107] *generalized mean* ou *power mean* en anglais, est définie pour une distribution de valeurs de x_s dont la probabilité d'occurrence est p_s par :

$$\bar{x} = \left(\sum_s p_s x_s^r \right)^{\frac{1}{r}} \tag{4.25}$$

[107] G. H. Hardy et al. (1952). *Inequalities*. Cambridge University Press.

La moyenne généralisée se réduit à la moyenne arithmétique pondérée pour $r = 1$. Elle donne un fort poids aux petites valeurs de x_s dans la moyenne pour les faibles valeurs de r (qui peuvent être négatives), et un fort poids aux grandes valeurs quand r est grand (Figure 4.14 [108]).

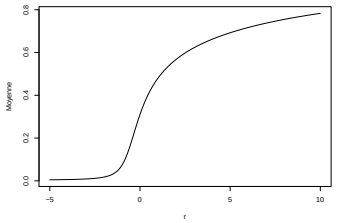


Figure 4.14 – Moyenne généralisée d'ordre r (en abscisse) d'une distribution uniforme de 100 valeurs tirées entre 0 et 1. La moyenne arithmétique 0,5 est obtenue pour $r = 1$. La moyenne généralisée tend vers la plus petite valeur (près de 0) quand $r \to -\infty$, et vers la plus grande valeur (près de 1) quand $r \to +\infty$.

La banalité peut s'écrire sous forme matricielle en notant $\boldsymbol{p}$ le vecteur des probabilités p_s, on note alors $(\boldsymbol{Zp})_s = \sum_t p_t z_{s,t}$.

La banalité moyenne des espèces de la communauté est calculée en prenant $r = q - 1$:

$$\left(\overline{\boldsymbol{Zp}}\right) = \left(\sum_s p_s (\boldsymbol{Zp})_s^{q-1} \right)^{\frac{1}{q-1}} \tag{4.26}$$

La diversité de la communauté est simplement l'inverse de la banalité moyenne des espèces :

$${}^qD^{\boldsymbol{Z}} = \left(\sum_s p_s (\boldsymbol{Zp})_s^{q-1} \right)^{\frac{1}{1-q}} \tag{4.27}$$

${}^qD^{\boldsymbol{Z}}$ converge vers ${}^1D^{\boldsymbol{Z}}$ quand q tend vers 1 :

$${}^1D^{\boldsymbol{Z}} = \frac{1}{\prod_s (\boldsymbol{Zp})_s^{p_s}} \tag{4.28}$$

4.4.2 Diversité neutre

Leinster et Cobbold montrent que la diversité HCDT est un cas particulier de ${}^qD^{\boldsymbol{Z}}$, pour la matrice identité :

[108] Le code R nécessaire pour réaliser la figure est :

```r
X <- runif(100)
# La moyenne n'est pas définie à r=0, décalage de 10^-8
r <- seq(-5, 10, 0.1) + 1e-08
plot(y = sapply(r, function(r) (mean(X^r))^(1/r)), x = r, type = "l",
    ylab = "Moyenne")
```

$$ {}^{q}D = {}^{q}D^{\boldsymbol{I}} \tag{4.29}$$

4.4.3 Diversité non neutre

Une mesure de diversité non neutre est obtenue en utilisant une matrice $\boldsymbol{Z}$ qui contient l'information sur la similarité entre les espèces. La dissimilarité entre espèces peut être obtenue à partir des arbres. L'exemple suivant utilise les données du calcul d'originalité, page 99. Pour un arbre de classe `hclust`, la fonction `cophenetic` fournit une demi-matrice de distances (classe `dist`).

```
cophenetic(hf)
```

```
##         Ess       Me       S1       Sr       Vm       Bg
## Me 1.223905
## S1 1.223905 1.003477
## Sr 2.463593 2.463593 2.463593
## Vm 9.730566 9.730566 9.730566 9.730566
## Bg 9.730566 9.730566 9.730566 9.730566 5.652955
## Ef 9.730566 9.730566 9.730566 9.730566 1.694406 5.652955
## Dg 9.730566 9.730566 9.730566 9.730566 1.101468 5.652955
##          Ef
## Me
## S1
## Sr
## Vm
## Bg
## Ef
## Dg 1.694406
```

La matrice doit être normalisée puis transformée en matrice de similarité. Une transformation simple est 1 moins la dissimilarité (préalablement normalisée pour être comprise entre 0 et 1) :

```
Dis <- cophenetic(hf)
(Z <- 1 - as.matrix(Dis/max(Dis)))
```

```
##           Ess        Me        S1        Sr        Vm
## Ess 1.0000000 0.8742206 0.8742206 0.7468192 0.0000000
## Me  0.8742206 1.0000000 0.8968737 0.7468192 0.0000000
## S1  0.8742206 0.8968737 1.0000000 0.7468192 0.0000000
## Sr  0.7468192 0.7468192 0.7468192 1.0000000 0.0000000
## Vm  0.0000000 0.0000000 0.0000000 0.0000000 1.0000000
## Bg  0.0000000 0.0000000 0.0000000 0.0000000 0.4190517
## Ef  0.0000000 0.0000000 0.0000000 0.0000000 0.8258677
## Dg  0.0000000 0.0000000 0.0000000 0.0000000 0.8868033
##           Bg        Ef        Dg
## Ess 0.0000000 0.0000000 0.0000000
## Me  0.0000000 0.0000000 0.0000000
## S1  0.0000000 0.0000000 0.0000000
## Sr  0.0000000 0.0000000 0.0000000
## Vm  0.4190517 0.8258677 0.8868033
## Bg  1.0000000 0.4190517 0.4190517
## Ef  0.4190517 1.0000000 0.8258677
## Dg  0.4190517 0.8258677 1.0000000
```

La diversité peut être calculée par la fonction `Dqz` de *entropart* :

```
# Vecteur contenant 8 espèces...
(effectifs <- read.csv2("http://www.ecofog.gf/IMG/csv/rao.effectifs.csv",
    row.names = 1, header = T))
```

```
##     P1 P2 P3 P4
## Ess 97 36 34 61
```

```
## Me  62 24 49  4
## S1   4  6  6 75
## Sr  71 78 99 36
## Vm  76 29 34  2
## Bg  58 49 47 19
## Ef   9 34 91 35
## Dg  98 14 95 14

Ps <- rowSums(effectifs)/sum(effectifs)
Dqz(Ps, q = 2, Z)

##     None
## 2.520938
```

Dans le cas particulier $q = 2$, ${}^2D^{\boldsymbol{Z}}$ est égal à ${}^2\bar{D}(T)$:

```
PhyloDiversity(Ps, q = 2, phyf)$Total

## [1] 2.520938
```

Mais ce n'est pas le cas en général.

4.4.4 Entropie de Ricotta et Szeidl

Ricotta et Szeidl[109] ont montré une similitude entre les entropies de Shannon et de Rao en généralisant l'entropie de Shannon en deux temps. Tout d'abord en remarquant que la probabilité d'occurrence d'une espèce est 1 moins la somme de celle des autres, d'où :

$$ {}^1H = -\sum_s p_s \ln \left(1 - \sum_{t \neq s} p_t\right) \tag{4.30} $$

$1 - \sum_{t \neq s} p_t$ est la banalité de l'espèce s si on mesure la diversité neutre. De façon plus générale, $(\boldsymbol{Zp})_s$ peut exprimer cette banalité. Enfin, l'entropie HCDT peut généraliser l'entropie de Shannon pour définir une mesure Q_α que nous noterons ${}^qH^{\boldsymbol{Z}}$ pour la cohérence avec les autres mesures $(q = \alpha)$:

$$ {}^qH^{\boldsymbol{Z}} = \frac{1 - \sum_s p_s (\boldsymbol{Zp})_s^{q-1}}{q-1} \tag{4.31} $$

$$ {}^1H^{\boldsymbol{Z}} = -\sum_s p_s \ln (\boldsymbol{Zp})_s \tag{4.32} $$

De façon plus rigoureuse, on peut remarquer que $(\boldsymbol{Zp})_s$ décroît quand p_s décroît parce que la similarité d'une espèce avec elle-même est maximale. Une entropie définie à partir d'une fonction d'information de p_s reste donc une entropie quand on utilise la même fonction d'information sur $(\boldsymbol{Zp})_s$: la fonction reste décroissante et vaut 0 pour $p_s = 1$ puisque $(\boldsymbol{Zp})_s = 1$ dans ce cas. ${}^qH^{\boldsymbol{Z}}$ est donc bien une entropie. La fonction d'information $\ln_q(1/p_s)$ de l'entropie HCDT (3.14) est simplement remplacée par $\ln_q(1/(\boldsymbol{Zp})_s)$.

[109] C. Ricotta et L. Szeidl (2006). « 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, p. 237–243.

L'entropie de Ricotta et Szeidl étend l'entropie HCDT en utilisant une fonction d'information plus générale, dépendant de la banalité de l'espèce plutôt que de sa seule probabilité, les deux étant égales dans le cas particulier de la diversité neutre. L'ordre de la diversité q permet de donner un plus ou moins grand poids aux espèces banales (et non aux espèces fréquentes : p_s est à la puissance 1). La fréquence et la banalité se confondent seulement pour la diversité neutre.

Le logarithme d'ordre q de ${}^qD^{\boldsymbol{Z}}$ est ${}^qH^{\boldsymbol{Z}}$:

$$\ln_q {}^qD^{\boldsymbol{Z}} = {}^qH^{\boldsymbol{Z}} \tag{4.33}$$

4.4.5 Définition de la similarité

La transformation d'une matrice de dissimilarité en une matrice de similarité nécessite une fonction strictement décroissante, dont le résultat est compris entre 0 et 1. La plus simple est $z_{s,t} = 1 - {}^{d_{s,t}}/_{\max(d_{s,t})}$. Leinster et Cobbold argumentent en faveur d'une transformation exponentielle négative, déjà utilisée par Nei.[110]

Le problème majeur de la diversité ${}^qD^{\boldsymbol{Z}}$ est qu'elle fournit souvent des valeurs presque indépendantes de q quand on l'applique à une matrice de similarité fonctionnelle.[111] De nombreux exemples, y compris dans l'article original de Leinster et Cobbold, montrent une très faible décroissance de la diversité de $q = 0$ à $q = 2$. Ce n'est pas un problème théorique mais numérique. Si la banalité de l'espèce s est proche de 1, sa puissance q peut être approchée par son développement limité au premier ordre : $(\boldsymbol{Zp})_s^{q-1} \approx 1 + (q-1)[(\boldsymbol{Zp})_s - 1]$. Cette approximation linéaire implique que $\sum_s p_s(\boldsymbol{Zp})_s^{q-1} \approx [\sum_s p_s(\boldsymbol{Zp})_s]^{q-1}$. La puissance $q-1$ disparaît donc dans le calcul de la diversité et l'équation 4.27 devient ${}^qD^{\boldsymbol{Z}} \approx \bar{z}$ où $\bar{z} = \sum_s p_s(\boldsymbol{Zp})_s$. Dans les faits, cette approximation vaut pour des valeurs de banalité assez éloignées de 1 (Figure 4.15 [112]). Si une majorité des espèces (au sens de la somme de leurs probabilités respectives) a une banalité supérieure à 0,5, l'approximation linéaire est assez bonne et la moyenne généralisée de la banalité est proche de sa moyenne arithmétique. Ce n'est pas un problème pour la comparaison de profils de diversité de communautés différentes calculés à partir de la même matrice de similarité mais l'interprétation de la diversité en termes de nombres effectifs d'espèces dépendant de q n'est pas très intuitive.

[110] M. Nei (1972). « Genetic Distance between Populations ». In : *The American Naturalist* 106.949, p. 283–292.

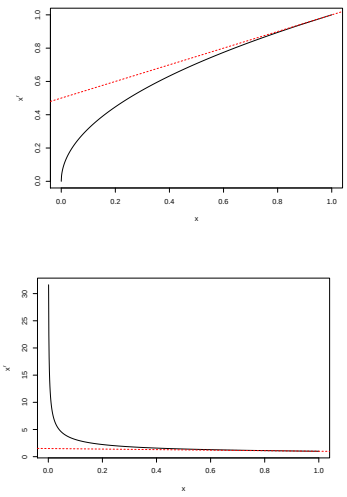


Figure 4.15 – Approximation de $(\boldsymbol{Zp})_s^{q-1}$ par son développement limité au premier ordre pour $q = 1.5$ en haut ($r = q - 1$) et $q = 0.5$ en bas.

[111] C.-H. Chiu et A. Chao (2014). « Distance-based functional diversity measures and their decomposition : a framework based on hill numbers. » In : *PloS one* 9.7, e100014.

[112] Le code R nécessaire pour réaliser la figure est :

```
r <- 0.5
seqZp <- seq(0, 1, 0.001)
plot(seqZp^r~seqZp, type="l", xlab="x", ylab=expression(x^r))
abline(a=1-r, b=r, col="red", lty=2)
```

```
r <- -0.5
plot(seqZp^r~seqZp, type="l", xlab="x", ylab=expression(x^r))
abline(a=1-r, b=r, col="red", lty=2)
```

La transformation exponentielle négative $z_{s,t} = e^{-u\frac{d_{s,t}}{\max(d_{s,t})}}$ est justifiée par Leinster.[113] u est une constante positive. Plus u est grand, plus la transformation est convexe : les similarités sont tassées vers 0 (Figure 4.16 [114]). Augmenter la valeur de u diminue les similarités et donc la banalité des espèces, ce qui règle (arbitrairement) le problème de la faible sensibilité de la diversité au paramètre q.

[113] T. Leinster (2013). « The Magnitude of Metric Spaces ». In : *Documenta Mathematica* 18, p. 857–905.

Du point de vue théorique, l'ordre de la matrice des distances est justifié (par les différences entre traits fonctionnels par exemple) mais leur distribution ne l'est pas : elle n'est que la conséquence de la méthode de calcul. Le choix de u ne change pas l'ordre des distances mais déforme la distribution des similarités. La distance étant normalisée, u fixe l'échelle des distances avant la transformation.

Leinster relie la valeur de ${}^qD^{\boldsymbol{Z}}$ à la magnitude de l'espace des espèces (espace dans lequel les espèces sont des points dont les distances deux à deux sont décrites par la matrice $\boldsymbol{\Delta}$, qui doit être euclidienne) : la magnitude de l'espace est la valeur maximale que peut atteindre la diversité. La magnitude est une propriété qui décrit la taille d'un espace dans le cadre de la théorie des catégories. Modifier u modifie la magnitude de l'espace selon une relation non triviale : u est un paramètre au même titre que q. Quand $u \to +\infty$, $\boldsymbol{Z} \to \boldsymbol{I}$: la diversité tend vers la diversité neutre. Leinster et Cobbold suggèrent de comparer les profils de diversité en fonction à la fois de q et de u mais cette voie n'a pas été explorée.

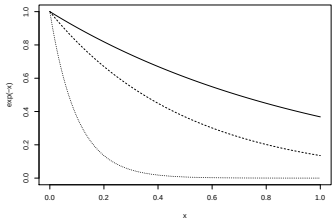


Figure 4.16 – Transformation des distances (en abscisse) en similarités (en ordonnées) selon la valeur de u. Trait plein : $u = 1$; pointillés longs : $u = 2$; pointillés : $u = 10$. Plus u est grand, plus les similarités sont tassées vers 0.

4.4.6 Biais d'estimation

Comme les autres mesures de diversité, ${}^qD^{\boldsymbol{Z}}$ est sujette au biais d'estimation. Deux estimateurs réduisant le biais existent.[115] Le premier utilise la technique de Chao et Shen[116] déjà vue pour la correction du biais de l'entropie de Shannon (page 49). L'estimateur de ${}^qH^{\boldsymbol{Z}}$ est :

[115] É. Marcon et al. (2014b). « The Decomposition of Similarity-Based Diversity and its Bias Correction ». In : *HAL* 00989454.version 3.

[116] Chao et Shen (2003). « Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample », cf. note 166, p. 49.

$$ {}^q\tilde{H}^{\boldsymbol{Z}} = \sum_s \frac{\hat{C}\hat{p}_s \ln_q \frac{1}{\left(\widetilde{\boldsymbol{Zp}}\right)_s}}{1 - \left(1 - \hat{C}\hat{p}_s\right)^n} \tag{4.34} $$

où :

$$ \left(\widetilde{\boldsymbol{Zp}}\right)_s = \sum_t \hat{C}\hat{p}_t z_{s,t} + \left(1 - \hat{C}\right)\bar{z} \tag{4.35} $$

[114] Le code R nécessaire pour réaliser la figure est :

```
curve(exp(-x), ylim=c(0,1))
curve(exp(-2*x), lty=2, add=TRUE)
curve(exp(-10*x), lty=3, add=TRUE)
```

L'estimateur de la diversité est ${}^{q}\tilde{D}^{\boldsymbol{Z}} = e_q^{{}^{q}\tilde{H}^{\boldsymbol{Z}}}$.

L'autre estimateur suppose que la similarité des espèces non observées avec toutes les autres est la similarité moyenne observée. L'estimation utilise la technique de Zhang et Grabchak,[117] qui consiste à développer l'estimateur en série entière au voisinage de $p_s = 1$. L'estimateur de ${}^{q}D^{\boldsymbol{Z}}$ est, pour $q \neq 1$:

[117] Zhang et Grabchak (2014a). « Entropic Representation and Estimation of Diversity Indices », cf. note 61, p. 67.

$$ {}^{q}\tilde{D}^{\boldsymbol{Z}} = \left\{ \hat{K} + \hat{V} - \sum_{s \leq s_{\neq 0}} \hat{C}\hat{p}_s \left[\bar{z} \left(1 - \hat{C}\hat{p}_s \right) + \hat{C}\hat{p}_s \right]^{q-1} \right\}^{\frac{1}{1-q}} \tag{4.36} $$

où :

$$ \hat{K} = \sum_{s \leq s_{\neq 0}} \hat{C}\hat{p}_s \left[\left(\sum_{t \leq s_{\neq 0}} \hat{C}\hat{p}_t z_{s,t} \right) + \left(1 - \hat{C} \right) \bar{z} \right]^{q-1} \tag{4.37} $$

et :

$$ \hat{V} = 1 + \sum_{s \leq s_{\neq 0}} \frac{n_s}{n} \sum_{v=1}^{n-n_s} (1 - \bar{z})^v \left[\prod_{i=1}^{v} \frac{i-q}{i} \right] \left[\prod_{j=1}^{v} \left(1 - \frac{n_s - 1}{n - j} \right) \right] \tag{4.38} $$

L'estimateur de ${}^{q}H^{\boldsymbol{Z}}$ est :

$$ {}^{q}\tilde{H}^{\boldsymbol{Z}} = \frac{\hat{K} + \hat{V} - \sum_{s \leq s_{\neq 0}} \hat{C}\hat{p}_s \left[\bar{z} \left(1 - \hat{C}\hat{p}_s \right) + \hat{C}\hat{p}_s \right]^{q-1} - 1}{1 - q} \tag{4.39} $$

Pour $q = 1$:

$$ {}^{1}\tilde{H}^{\boldsymbol{Z}} = \hat{L} + \hat{W} + \sum_{s \leq s_{\neq 0}} \hat{C}\hat{p}_s \ln \left[\bar{z} \left(1 - \hat{C}\hat{p}_s \right) + \hat{C}\hat{p}_s \right] \tag{4.40} $$

$$ {}^{1}\tilde{D}^{\boldsymbol{Z}} = e^{{}^{1}\tilde{H}^{\boldsymbol{Z}}} \tag{4.41} $$

où :

$$ \hat{L} = - \sum_{s \leq s_{\neq 0}} \hat{C}\hat{p}_s \ln \left[\left(\sum_{t \leq s_{\neq 0}} \hat{C}\hat{p}_j z_{s,t} \right) + \left(1 - \hat{C} \right) \bar{z} \right] \tag{4.42} $$

$$ \hat{W} = \sum_{s \leq s_{\neq 0}} \frac{n_s}{n} \sum_{v=1}^{n-n_s} \frac{(1 - \bar{z})^v}{v} \left[\prod_{j=1}^{v} \left(1 - \frac{n_s - 1}{n - j} \right) \right] \tag{4.43} $$

Le package *entropart* fournit les fonctions `Dqz` et `HqZ` pour calculer ces estimateurs.

```
data(Paracou618)
# Matrice de distances fonctionnelles
DistanceMatrix <- as.matrix(Paracou618.dist)
# Matrice de similarité
Z <- 1 - DistanceMatrix/max(DistanceMatrix)
# Calcul de la diversité
Dqz(Paracou618.MC$Ns, q = 1, Z)

##     Best
## 1.488151
```

4.4.7 Diversité phylogénétique

La diversité phylogénétique ${}^q\bar{D}(T)$ est un cas particulier de ${}^q D^{\boldsymbol{Z}}$ pour une matrice $\boldsymbol{Z}$ dont les lignes et colonnes sont la similarité des « espèces historiques », c'est-à-dire les ancêtres des espèces actuelles dans l'arbre phylogénétique. La matrice est construite en prenant en compte les paires (espèce actuelle ; branche ancestrale), par exemple, Figure 4.7, page 96, pour l'espèce 1, les branches $b_{1,1}$ et $b_{3,1}$, pour l'espèce 5 les branches $b_{1,5}$ et $b_{2,3}$. La matrice $\boldsymbol{Z}$ est de dimension 10 dans cet exemple. La diversité phylogénétique ${}^q\hat{D}(T)$ est obtenue quand les éléments de la matrice valent 1 ou 0 selon les règles suivantes : les éléments d'une ligne (par exemple les deux lignes correspondant à la branche $b_{3,1}$) valent 1 pour toutes les colonnes correspondant à une espèce descendant de la branche (les 4 colonnes correspondants aux espèces 1 et 2), 0 pour les autres colonnes. La matrice n'est donc pas symétrique.

La probabilité de chaque espèce historique est celle de l'espèce actuelle multipliée par la longueur de la branche normalisée par la hauteur de l'arbre.

4.5 Synthèse

${}^q D(T)$ est l'exponentielle d'ordre q de la moyenne pondérée de l'entropie HCDT calculée à chaque période de l'arbre phylogénétique.

${}^q D^{\boldsymbol{Z}}$ est l'exponentielle d'ordre q de l'entropie de Ricotta et Szeidl, mais aussi l'inverse de la banalité moyenne des espèces de la communauté.

${}^q D^{\boldsymbol{Z}}$ et ${}^q D$ sont des nombres effectifs d'espèces, respectent toutes les propriétés demandées à une mesure de diversité dont le principe de réplication. Leurs valeurs sont identiques pour $q = 2$ (diversité de Rao). ${}^q D^{\boldsymbol{Z}}$ est moins sensible au paramètre q que ${}^q D(T)$: la Figure 4.17 [118] compare les deux mesures.

[118] Le code R nécessaire pour réaliser la figure est :

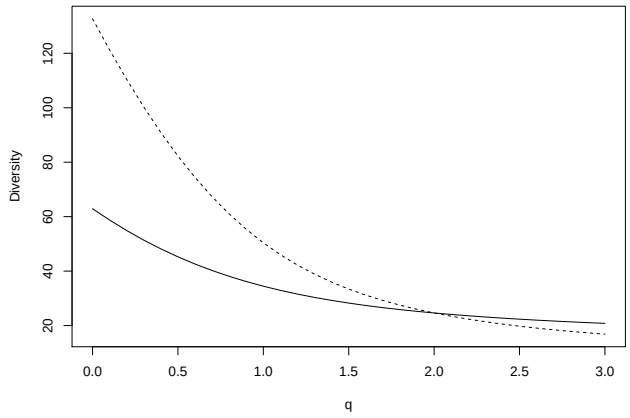


FIGURE 4.17 – Diversité phylogénétique (l'arbre est une taxonomie) d'ordre q de deux hectares de forêt de Paracou : ${}^{q}D^{\boldsymbol{Z}}$ (trait plein) et ${}^{q}\hat{D}(T)$ (pointillé).

${}^{q}D^{\boldsymbol{Z}}$ est particulièrement intéressante pour mesurer la diversité fonctionnelle, souvent définie à partir d'une matrice de distances entre les espèces. La transformation d'une matrice en un arbre phylogénétique déforme les données.[119] La diversité de Leinster et Cobbold est calculée directement à partir de la matrice.

Si l'arbre phylogénétique n'est pas ultramétrique, le calcul de la diversité est possible, mais la matrice $\boldsymbol{Z}$ contient alors des valeurs comprises entre 0 et 1 qui dépendent des effectifs des espèces actuelles (ils interviennent dans le calcul de la hauteur de l'arbre qui est la moyenne de la longueur des branches). La dépendance entre similarité et fréquence des espèces constitue un problème théorique qui empêche d'interpréter la diversité calculée à partir d'un arbre non ultramétrique.

[119] PAVOINE et al. (2005a). « Is the originality of a species measurable ? », cf. note 29, p. 86 ; PODANI et SCHMERA (2006). « On dendrogram-based measures of functional diversity », cf. note 24, p. 86 ; PODANI et SCHMERA (2007). « How should a dendrogram-based measure of functional diversity function ? A rejoinder to Petchey and Gaston », cf. note 29, p. 86.

4.6 Diversité individuelle

La diversité ${}^{q}D^{\boldsymbol{Z}}$ et tous ses cas particuliers (phylodiversité et diversité neutre) peuvent être envisagés au niveau individuel plutôt qu'au niveau de l'espèce.

Le calcul de la diversité n'est pas affecté par le remplacement d'une espèce par deux espèces identiques. La ligne et la colonne de la matrice $\boldsymbol{Z}$, correspondant à l'espèce s sont remplacées par deux lignes et colonnes identiques correspondant aux espèces s' et s'', dont la similarité avec les autres espèces est la même que celle de l'espèce s et la similarité entre elles égale à 1, les

```
q <- c(seq(0, 3, 0.1))
DistanceMatrix <- as.matrix(Paracou618.Taxonomy$Wdist^2/2)
Z <- 1 - DistanceMatrix/max(DistanceMatrix)
Divqz <- sapply(q, function(q) Dqz(Paracou618.MC$Ps, q, Z))
Divq <- sapply(q,
     function(q) ChaoPD(Paracou618.MC$Ps, q, Paracou618.Taxonomy))
plot(y=Divq, x=q, type="l", ylab="Diversity", lty=2)
lines(y=Divqz, x=q, lty=1)
```

probabilités vérifiant $p_s = p_{s'} + p_{s''}$. L'opération peut être répétée jusqu'à la désagrégation complète de la matrice où chaque ligne correspondrait à un individu et les probabilités seraient égales à $1/N$. Il n'y donc pas de différence entre la mesure de la diversité individuelle et celle de la diversité spécifique qui n'est qu'une façon pratique de regrouper des individus ayant la même banalité. Dit autrement, l'entropie d'une communauté est la somme des entropie de ses espèces, qui n'est que la somme pondérée des entropies individuelles si tous les individus d'une espèce ont la même entropie.

La variabilité intraspécifique peut être traitée par un raisonnement similaire. L'idée d'intégrer aux mesures de diversité la possibilité que tous les individus d'une espèce ne soient pas semblables émerge dans la littérature.[120] Mais si des individus de la même espèce ne sont pas totalement similaires, ils ne peuvent pas être regroupés en prenant pour similarité de l'espèce avec elle-même une valeur unique qui résumerait la similarité entre les individus (au lieu de 1) pour tous les ordres de diversité.

[120] S. Pavoine et C. Ricotta (2014). « Functional and phylogenetic similarity among communities ». In : *Methods in Ecology and Evolution* 5.7, p. 666–675 ; Chiu et Chao (2014). « Distance-based functional diversity measures and their decomposition : a framework based on hill numbers. », cf. note 111, p. 109.

Formellement, si l'espèce s est composée de deux groupes s' et s'', la banalité du groupe s' est $(\boldsymbol{Zp})_{s'} = \sum_{t \neq s',s''} p_t z_{s',t} + p_{s'} + p_{s''} z_{s',s''}$, celle du groupe s'' est $(\boldsymbol{Zp})_{s''} = \sum_{t \neq s',s''} p_t z_{s'',t} + p_{s'} z_{s',s''} + p_{s''}$. Les similarités avec les autres espèces sont identiques : $z_{s,t} = z_{s',t} = z_{s'',t}$. La contribution à la diversité des deux groupes $\sum_{s'} p_{s'} (\boldsymbol{Zp})_{s'}^{q-1} + \sum_{s''} p_{s''} (\boldsymbol{Zp})_{s''}^{q-1}$ est remplacée après regroupement par $\sum_s p_s (\boldsymbol{Zp})_s^{q-1}$. Puisque $p_s = p_{s'} + p_{s''}$, le regroupement n'est possible quel que soit q que si les banalités des deux groupes sont identiques, ce qui interdit toute variabilité intraspécifique : $z_{s',s''}$ est obligatoirement égal à 1. La non-linéarité de la moyenne généralisée ne permet pas de définir une matrice de similarité intégrant la variabilité intraspécifique.

Dans le cas de la diversité de Rao ($q = 2$), on peut chercher la valeur de $z_{s,s}$, différente de 1, définissant la similarité de l'espèce avec elle même pour prendre en compte sa variabilité. La résolution de l'équation $\sum_{s'} p_{s'} (\boldsymbol{Zp})_{s'} + \sum_{s''} p_{s''} (\boldsymbol{Zp})_{s''} = \sum_s p_s (\boldsymbol{Zp})_s$ permet de trouver $z_{s,s} = 1 + 2p_{s'}p_{s''}(z_{s',s''} - 1)/p_s^2$. Dans le cas le plus simple où $p_{s'} = p_{s''}$, la similarité intraspécifique est $(1 + z_{s',s''})/2$. La valeur de $z_{s,s}$ peut être cherchée pour n'importe quelle valeur de q, mais la résolution de l'équation est en général impossible analytiquement, et $z_{s,s}$ varie avec q.

L'exemple suivant considère deux espèces, dont une comprenant deux groupes :

```
# Similarité entre les deux groupes
zs12 <- 0.9
# Probabilités des deux groupes
ps1 <- 1/5
ps2 <- 1/5
# Matrice de similarité
(Z2s <- matrix(c(1, zs12, 0.5, zs12, 1, 0.5, 0.5, 0.5, 1), nrow = 3))

##      [,1] [,2] [,3]
```

```
## [1,]  1.0  0.9  0.5
## [2,]  0.9  1.0  0.5
## [3,]  0.5  0.5  1.0

# Probabilités
(P2s <- c(ps1, ps2, 1 - ps1 - ps2))

## [1] 0.2 0.2 0.6

# Diversité
Dqz(P2s, 2, Z2s)

##     None
## 1.329787

# Similarité intraspécifique après regroupement
(zss <- 1 + 2 * ps1 * ps2/(ps1 + ps2)^2 * (zs12 - 1))

## [1] 0.95

# Matrice de similarité après regroupement
(Z1s <- matrix(c(zss, 0.5, 0.5, 1), nrow = 2))

##      [,1] [,2]
## [1,] 0.95  0.5
## [2,] 0.50  1.0

# Probabilité après regroupement
(P1s <- c(ps1 + ps2, 1 - ps1 - ps2))

## [1] 0.4 0.6

# Diversité
Dqz(P1s, 2, Z1s)

##     None
## 1.329787
```

En pratique, si les similarités entre individus ou groupes sont connues, le regroupement n'a aucun intérêt. Mais en absence de données individuelles, il est possible de choisir arbitrairement une similarité intraspécifique différente de 1 (ou de façon équivalente une distance différente de 0) pour calculer une diversité d'ordre fixé, de préférence 2.

4.7 Diversité des valeurs propres

L'approche de Pavoine et Izsák[121] permet de prendre en compte la variabilité intraspécifique de façon rigoureuse. Pour mesurer la diversité neutre, les espèces sont placées dans un espace multidimensionnel, de dimension S. Chaque espèce est représentée par un vecteur de longueur p_s sur son axe. Cette représentation a été utilisée par Campos et Isaza[122] qui ont relié la diversité de Simpson au volume de la sphère sur laquelle se trouve le point définissant la communauté (Figure 4.18).

La matrice de similarité entre les espèces $\boldsymbol{Z}$ permet d'affiner la représentation. Soit la matrice $\boldsymbol{C} = \sqrt{\boldsymbol{Z}}$. Chaque espèce s est maintenant représentée par le vecteur dont la coordonnée sur l'axe t est $\sqrt{p_s p_t} c_{s,t}$. Les axes correspondent aux espèces « pures », totalement dissimilaires, les vecteurs des espèces réelles prennent

[121] S. Pavoine et J. Izsák (2014). « New biodiversity measure that includes consistent interspecific and intraspecific components ». In : *Methods in Ecology and Evolution* 5.2, p. 165–172.

[122] D. Campos et J. F. Isaza (2009). « A geometrical index for measuring species diversity ». In : *Ecological Indicators* 9.4, p. 651–658.

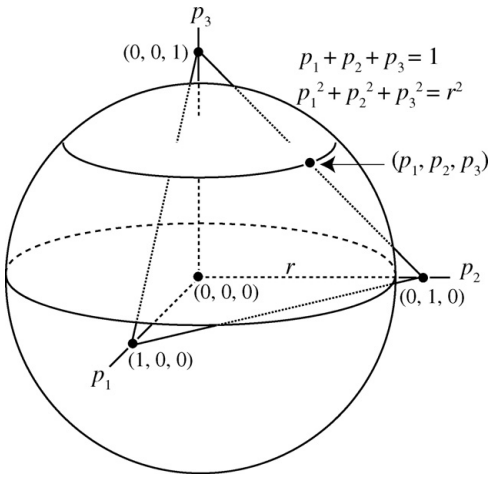


FIGURE 4.18 – Représentation d'une communauté dans un espace multidimensionnel dont chaque axe correspond à une espèce. La communauté est composée de trois espèces. Le point représentant la communauté se trouve sur la sphère de rayon $r = \sqrt{\sum_s p_s^2}$. Le plan (qui est en réalité un hyperplan de dimension $S - 1$) d'équation $\sum_s p_s = 1$ est représenté par un triangle. (*in* Campos et Isaza, 2009)

en compte la similarité. Dans le cas extrême de la diversité neutre, $\boldsymbol{C} = \boldsymbol{I}$, chaque espèce est située uniquement sur son axe. Dans l'autre cas extrême de totale similarité, où $\boldsymbol{C}$ ne contient que des 1, toutes les espèces sont colinéaires.

Il est possible de calculer les S valeurs propres notées λ_s de la matrice des coordonnées des espèces et de les normaliser : $\mu_s = \lambda_s / \sum_s \lambda_s$. La communauté peut maintenant être représentée par ses valeurs propres μ_s correspondant à la proportion d'une « espèce composite » pure sur chaque axe généré par les vecteurs propres : cette transformation est une ACP non centrée, non réduite. La diversité HCDT des valeurs propres est la diversité de la communauté, qui sera notée ici ${}^qD^{\boldsymbol{Z}}(\Lambda)$.

La définition de cette diversité autorise toute matrice $\boldsymbol{C}$ symétrique, dont les valeurs de similarité sont positives ou nulles, strictement positives sur la diagonale. Si $\boldsymbol{C}$ est la racine carrée d'une matrice de similarité au sens strict $\boldsymbol{Z}$, c'est-à-dire contenant des valeurs comprises entre 0 et 1 et dont les éléments de la diagonale sont tous égaux à 1, alors ${}^2D^{\boldsymbol{Z}}(\Lambda)$ est égale à la diversité de Rao appliquée à une matrice de distances $\boldsymbol{D} = 1 - \boldsymbol{Z}$. La définition de $\boldsymbol{C}$ comme racine carrée de $\boldsymbol{Z}$ se justifie par des raisons géométriques : la norme de chaque vecteur représentant une espèce est la racine carrée de la somme des carrés de ses coordonnées. Le carré de la norme du vecteur de l'espèce s est donc p_s multiplié par la banalité de l'espèce.

Si $\boldsymbol{Z} = \boldsymbol{I}$, l'ACP ne modifie pas le nuage de points original et ${}^qD^{\boldsymbol{Z}}(\Lambda) = {}^qD$.

L'exemple suivant calcule la diversité de la méta communauté méta-communauté `Paracou618` selon la matrice de distances `Paracou618.dist`.

```
# Probabilités
Ps <- Paracou618.MC$Ps[Paracou618.MC$Ps > 0]
# Matrice de distances fonctionnelles
DistanceMatrix <- as.matrix(Paracou618.dist)
# Transformation en matrice de similarité
Z <- 1 - DistanceMatrix/max(DistanceMatrix)
```

```
# Mise en correspondance de la matrice et du vecteur de
# probabilités
Z <- Z[names(Ps), names(Ps)]
# Matrice diagonale contenant les probabilités
Q <- diag(Ps)
# Matrice des espèces (les lignes de Sp sont les coordonnées
# des vecteurs)
Sp <- sqrt(Q %*% Z %*% Q)
# Valeurs propres normalisées
Lambda <- svd(Sp)$d
Mu <- Lambda/sum(Lambda)
# Comparaison des valeurs à q=2
Diversity(Mu, 2)

##     None
## 1.516347

Dqz(Ps, 2, Z)

##     None
## 1.472353

# La valeur à q=0 est le nombre d'espèces
Diversity(Mu, 0)

## None
##  229

length(Ps)

## [1] 229
```

La valeur de la diversité des valeurs propres est légèrement différente de celle de la diversité de Leinster et Cobbold à cause des approximations numériques liées au calcul des valeurs propres (qui nécessitent l'inversion d'une matrice carrée de dimension S).

Le profil de diversité se trouve en Figure 4.19 [123].

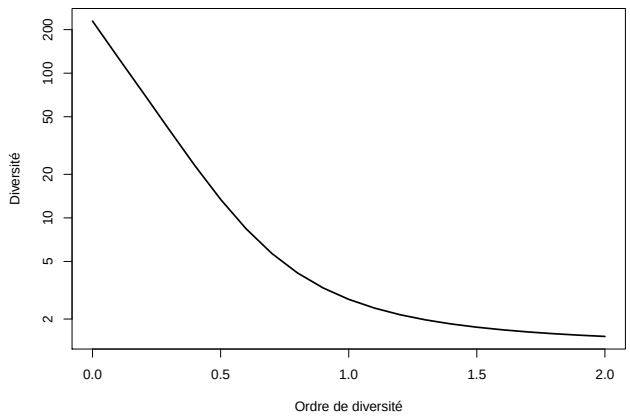


FIGURE 4.19 – Profil de diversité des valeurs propres de la méta-communauté `Paracou618`.

La diversité des valeurs propres permet de ramener le pro-

[123] Code R :

```
DqZLambda <- CommunityProfile(Diversity, Mu)
plot(DqZLambda, log = "y", xlab = "Ordre de diversité", ylab = "Diversité")
```

blème de la diversité d'espèces partiellement similaires au calcul de la diversité neutre d'une communauté d'espèces composites totalement dissimilaires.

Les éléments de la diagonale de la matrice de similarité ne sont pas forcément égaux à 1 : des valeurs inférieures correspondent à la variabilité intraspécifique, selon le mécanisme de regroupement traité au paragraphe précédent. Les raisons qui empêchent d'utiliser des valeurs de similarité intraspécifique différentes de 1 dans le calcul de la diversité de Leinster et Cobbold ne s'appliquent pas ici : quelle que soit la valeur de q, les valeurs propres sont les mêmes.

Les limites de la diversité des valeurs propres sont cependant nombreuses. La première est numérique : son calcul est imprécis dans des communautés très riches. L'inversion d'une matrice de dimension 200 ou plus est toujours problématique.

Les espèces non échantillonnées entraînent un biais d'estimation : le nombre de dimensions est sous-estimé. Comme les espèces manquantes ont une faible probabilité, leur vecteur est petit et n'influe pas beaucoup sur la diagonalisation de la matrice. Les valeurs propres manquantes sont donc petites, elles influent sur la diversité pour les faibles valeurs de q, notamment la diversité d'ordre 0 qui est le nombre de dimensions de la matrice des espèces (en général, le nombre d'espèces, ou une valeur inférieure si des espèces sont colinéaires). Il n'existe pas de technique pour corriger ce biais d'estimation.

La composition de la communauté en espèces composites dépend à la fois de la matrice de similarité et des probabilités des espèces réelles. Elle est donc unique pour chaque communauté, ce qui empêche toute décomposition de la diversité selon les méthodes présentées plus loin.

4.8 Diversité de Scheiner

[124] S. M. Scheiner (2012). « A metric of biodiversity that integrates abundance, phylogeny, and function ». In : *Oikos* 121.8, p. 1191–1202.

Scheiner[124] développe un cadre unifié pour mesurer la diversité spécifique, phylogénétique ou fonctionnelle, séparément ou simultanément. L'idée générale est que toute quantité partagée par les espèces (le nombre d'individus, le temps d'évolution accumulé, la taille des niches écologique) peut être traduite en nombre de Hill.

La diversité spécifique est simplement ${}^{q}D$. Scheiner la note ${}^{q}D(A)$, pour diversité d'abondance.

La diversité phylogénétique ${}^{q}D(P)$ ne prend pas en compte les abondances mais mesure la diversité de la divergence évolutive des espèces. La divergence totale dans un arbre phylogénétique (pas forcément ultramétrique) est la longueur totale des branches (c'est-à-dire FD). Elle est répartie entre toutes les espèces : la longueur de chaque branche (représentant une quantité d'évolution) est partagée à parts égales entre les espèces qui en descendent. La

divergence de l'espèce 1 de la Figure 4.7, page 96, est $T1 + T2$, la longueur de la branche terminale, plus ${}^{T3}/_{2}$, parce que la branche est partagée par les espèces 1 et 2. La divergence de l'espèce 1, est donc $L_1 = T1 + T2 + {}^{T3}/_{2}$. La part de la divergence de l'espèce s est $l_s = {}^{L_s}/_{FD}$. La diversité phylogénétique est le nombre de Hill des divergences :

$$ {}^{q}D = \left(\sum_{s=1}^{S} l_s^q \right)^{\frac{1}{1-q}} \tag{4.44} $$

La diversité fonctionnelle ${}^{q}D(F)$ est la diversité des tailles des niches. La taille de la niche est définie par Scheiner comme le volume de l'hypersphère (dans l'espace des traits fonctionnels de dimension m) centrée sur chaque espèce dont le rayon est la moitié de la distance à l'espèce la plus proche pour que les sphères ne se superposent pas. Une meilleure définition[125] de la taille de la niche est la somme des distances aux autres espèces : $t_s = \sum_t d_{s,t}$. La part de chaque espèce est $f_s = {}^{t_s}/_{\sum_t t_t}$ et la définition de la diversité fonctionnelle est :

[125] S. J. Presley et al. (2014). « Evaluation of an integrated framework for biodiversity with a new metric for functional dispersion. » In : *PloS one* 9.8, e105818.

$$ {}^{q}D(T) = \left(\sum_{s=1}^{S} f_s^q \right)^{\frac{1}{1-q}} \tag{4.45} $$

La diversité peut prendre en compte plusieurs composantes, pas exemple l'abondance et la phylogénie pour définir :

$$ {}^{q}D(AP) = \left(\sum_{s=1}^{S} \left(\frac{n_s L_s}{\sum_{t=1}^{S} n_t L_t} \right)^q \right)^{\frac{1}{1-q}} \tag{4.46} $$

La diversité d'abondance et phylogénétique est la diversité des divergences pondérées par les effectifs des espèces. La mesure de biodiversité de Scheiner, incluant les trois composantes, est :

$$ {}^{q}D(APF) = \left(\sum_{s=1}^{S} \left(\frac{n_s L_s t_s}{\sum_{t=1}^{S} n_t L_t t_t} \right)^q \right)^{\frac{1}{1-q}} \tag{4.47} $$

Son interprétation est moins immédiate. Chaque espèce est associée à une fraction d'une des trois dimensions de la diversité (abondance, temps d'évolution cumulé, taille des niches). Elle occupe un parallélépipède de dimensions (p_s, l_s, f_s) dans le cube de dimension 1 qui les contient toutes. Cette représentation ne donne pas d'importance particulière à l'abondance, et peut être appliquée à un nombre quelconque de dimensions. La mesure de biodiversité est la diversité des volumes occupés par les espèces, normalisés par leur somme (qui n'est pas égale à 1).

En se limitant à deux dimensions pour la lisibilité, la figure 4.20[126] présente les rectangles correspondant à la diversité d'abondance et fonctionnelle occupés par les espèces de la méta-communauté `Paracou618`, classées par fréquences décroissantes.

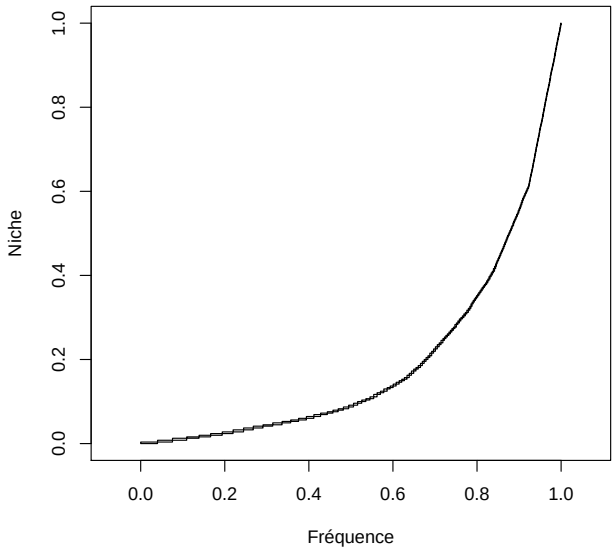


FIGURE 4.20 – Rectangles de surface fréquence × taille de niche des espèces de la méta-communauté `Paracou618`. La diversité de Scheiner ${}^{q}D(AF)$ est la diversité de leur surface.

La diversité d'abondance et fonctionnelle est la diversité de la surface des rectangles :

```
Surface <- Ps * fs
Rs <- Surface/sum(Surface)
Diversity(Rs, q = 2)

##     None
## 71.72482
```

[126] Code R :

```
# Probabilités
Ps <- Paracou618.MC$Ps[Paracou618.MC$Ps > 0]
Ps <- sort(Ps, decreasing = TRUE)
# Fréquences cumulées
PsCum <- cumsum(Ps)
Xgauche <- c(0, PsCum)
Xdroite <- c(PsCum, 1)
# Matrice de distances fonctionnelles
DistanceMatrix <- as.matrix(Paracou618.dist)
# Mise en correspondance de la matrice et du vecteur de
# probabilités
DistanceMatrix <- DistanceMatrix[names(Ps), names(Ps)]
# Taille des niches
ts <- rowSums(DistanceMatrix)
fs <- ts/sum(ts)
# Fréquences cumulées
fsCum <- cumsum(fs)
Ybas <- c(0, fsCum)
Yhaut <- c(fsCum, 1)
# Rectangles occupés par chaque espèce
plot(c(0, 1), c(0, 1), asp = 1, type = "n", xlab = "Fréquence",
    ylab = "Niche")
rect(Xgauche, Ybas, Xdroite, Yhaut)
```

Le profil de diversité d'abondance et fonctionnelle est en figure 4.21 [127].

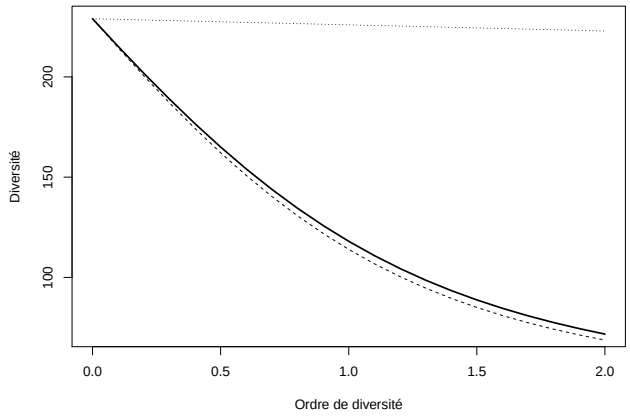


FIGURE 4.21 – Profil de diversité d'abondance et fonctionnelle de Scheiner de la méta-communauté `Paracou618`. Pointillés longs : diversité d'abondance ; pointillés courts : diversité fonctionnelle ; trait plein : diversité d'abondance et fonctionnelle.

La notion de diversité phylogénétique ou fonctionnelle traitée par Scheiner est assez différente de celles vues précédemment parce qu'elle considère toutes les dimensions de façon symétrique : l'ordre de diversité, q, s'applique à toutes les dimensions. Si q est grand, les espèces dont le produit des dimensions de la diversité est petit (que l'espèce soit rare, ait une divergence phylogénétique faible ou une niche étroite) sont négligées, alors que q n'affecte que la fréquence dans le cadre de la phylodiversité ${}^{q}\bar{D}$ ou la banalité pour ${}^{q}D^{Z}$.

[127] Code R :

```
plot(CommunityProfile(Diversity, Rs), xlab = "Ordre de diversité",
    ylab = "Diversité")
lines(CommunityProfile(Diversity, Ps), lty = 2)
lines(CommunityProfile(Diversity, fs), lty = 3)
```

Chapitre 5

Diversité β et décomposition

La notion de diversité β a été introduite par Whittaker[1] comme le niveau de changement dans la composition des communautés, ou le degré de différenciation des communautés, en relation avec les changements de milieu. La traduction de cette notion intuitive en une définition sans ambiguïté est encore une question de recherche et de débats. Anderson *et al.*[2] fournissent une revue des analyses utiles de la diversité β en forme de guide à destination des écologues reprise ici en introduction.

Dans la littérature, la diversité β est généralement une mesure *dérivée*,[3] c'est-à-dire calculée à partir des diversités α et γ, d'où un abondant débat sur l'indépendance souhaitée mais pas toujours observée entre les valeurs de diversité α et β.[4] Un forum a été consacré à la question dans la revue *Ecology*, conclu par Chao *et al.*,[5] et une revue a été faite par Tuomisto.[6]

Une approche alternative consiste à prendre en compte la distance entre paires de communautés, et utiliser les méthodes d'analyse multivariée pour représenter la diversité.[7] Elle est traitée dans le chapitre 6.

Pour simplifier l'exposé, les individus seront échantillonnés dans des communautés, appartenant à une méta-communauté. Le Tableau 5.1 résume les notations.

Enfin, toute ambiguïté sera évitée en appelant « entropie » les indices de diversité qui sont des mesures d'entropie ${}^qH$ (indices et Shannon et de Simpson par exemple) et « diversité » leur nombre équivalent ${}^qD$.

Des exemples montrent comment calculer cette diversité, principalement à l'aide du package *entropart*. Le package contient les données d'inventaire de deux hectares du dispositif de Paracou :

```
library(entropart)
# Chargement du jeu de données
data(Paracou618)
```

Les données sont organisées dans le package sous la forme d'un objet `MetaCommunity` qui contient notamment la matrice des $n_{s,i}$.

[1] Whittaker (1960). « Vegetation of the Siskiyou Mountains, Oregon and California », cf. note 23, p. 4, page 320.

[2] M. J. Anderson et al. (2011). « Navigating the multiple meanings of β diversity : a roadmap for the practicing ecologist ». In : *Ecology Letters* 14.1, p. 19–28.

[3] A. M. Ellison (2010). « Partitioning diversity ». In : *Ecology* 91.7, p. 1962–1963.

[4] A. Baselga (2010a). « Multiplicative partition of true diversity yields independent alpha and beta components ; additive partition does not ». In : *Ecology* 91.7, p. 1974–1981 ; L. Jost (2010a). « Independence of alpha and beta diversities ». In : *Ecology* 91.7, p. 1969–1994 ; J. A. Veech et T. O. Crist (2010). « Diversity partitioning without statistical independence of alpha and beta ». In : *Ecology* 91.7, p. 1964–1969.

[5] A. Chao et al. (2012). « Proposing a resolution to debates on diversity partitioning ». In : *Ecology* 93.9, p. 2037–2051.

[6] H. Tuomisto (2010a). « A diversity of beta diversities : straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity ». In : *Ecography* 33.1, p. 2–22 ; H. Tuomisto (2010b). « A diversity of beta diversities : straightening up a concept gone awry. Part 2. Quantifying beta diversity and related phenomena ». In : *Ecography* 33.1, p. 23–45.

[7] C. J. F. ter Braak (1983). « Principal components biplots and alpha and beta diversity ». In : *Ecology* 64.3, p. 454–462.

TABLE 5.1 – Notations des effectifs, tableau espèces-communautés.

	Communauté i	...	Total : méta-communauté
Espèce s	$n_{s,i}$: nombre d'individus de l'espèce s dans la communauté i. $\hat{p}_{s,i} = n_{s,i}/n_{+i}$ est l'estimateur de la probabilité $p_{s,i}$ qu'un individu de la communauté i soit de l'espèce s.		$n_{s+} = \sum_i n_{s,i}$ $p_s = \sum_i w_i p_{s,i}$
...			
Total	n_{+i} : nombre d'individus de la communauté. w_i : poids de la communauté		n : nombre total d'individus échantillonnés

Les parcelles de Paracou sont dans l'objet `Paracou618.MC`.

```
summary(Paracou618.MC)

## Meta-community (class 'MetaCommunity') made of 1124
## individuals in 2 communities and 425 species.
##
## Its sample coverage is 0.92266748426447
##
## Community weights are:
## [1] 0.5720641 0.4279359
## Community sample numbers of individuals are:
## P006 P018
##  643  481
## Community sample coverages are:
##      P006      P018
## 0.8943859 0.8463782
```

5.1 Cadre

5.1.1 Définition de la diversité β par objectifs

Anderson *et al.*[8] proposent une revue des définitions de la diversité β selon les analyses écologiques auquelles elles peuvent contribuer plutôt que selon leurs propriétés mathématiques. Deux types de diversité β peuvent être définies[9] en accord avec les notions originales de Whittaker : le *remplacement* (*turnover*) des espèces le long d'un gradient environnemental, spatial ou temporel et la *variation* de la composition spécifique entre communautés, qui diffèrent de la méta-communauté et entre elles. Ces deux définitions sont présentées sur les figures 5.1 et 5.2.

Les mesures de remplacement ne sont pas traitées en détail dans ce chapitre mais simplement présentées ici.

La plus simple (Figure 5.1, T1) est la dissimilarité entre deux communautés notée Δy. De nombreuses mesures de dissimilarité existent, elles sont détaillées dans le chapitre 6, à partir de la page 156. Whittaker n'envisageait que la présence ou l'absence des espèces, mais leur abondance peut être prise en compte.

Cette dissimilarité peut être analysée en fonction du gradient de référence. La dissimilarité entre des paires de communautés présentant la même différence (par exemple ayant subi ou non un

[8] ANDERSON et al. (2011). « Navigating the multiple meanings of β diversity : a roadmap for the practicing ecologist », cf. note 2, p. 123.

[9] M. VELLEND (2001). « Do commonly used indices of f-diversity measure species turnover ? » In : *Journal of Vegetation Science* 12, p. 545–552.

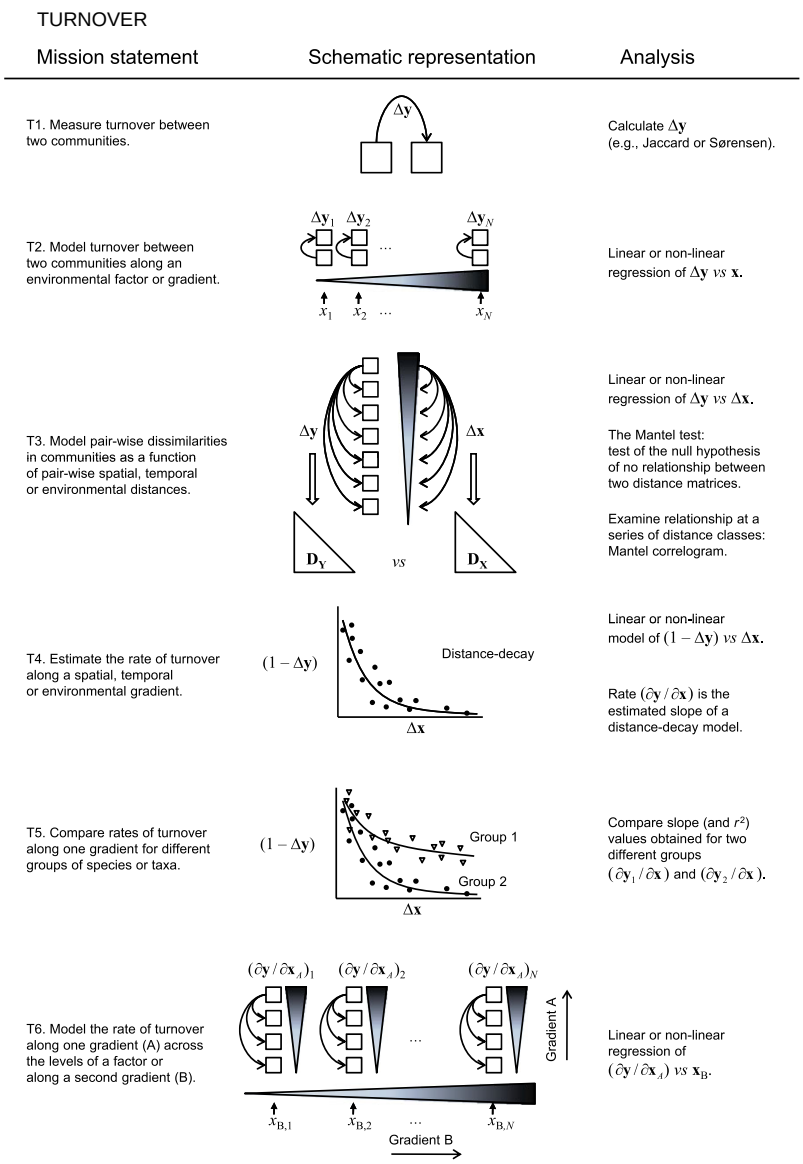


FIGURE 5.1 – Diversité β liée au remplacement des espèces le long d'un gradient. Δy est la mesure de la dissimilarité entre deux communautés, voir page 156. $\boldsymbol{x}$ est un gradient environnemental, spatial ou temporel. Δx est l'écart entre deux points du gradient. (*in* Anderson *et al.*[a])

[a] M. J. ANDERSON et al. (2011). « Navigating the multiple meanings of β diversity : a roadmap for the practicing ecologist ». In : *Ecology Letters* 14.1, p. 19–28.

traitement) situées le long d'un gradient (par exemple la fertilité du sol) peut être analysée en fonction de ce gradient (Figure 5.1, T2). Plus classiquement, la diversité β peut être régressée contre la valeur du gradient (T3). La dissimilarité de chaque paire de communautés est confrontée à son écart sur le gradient. Les deux analyses précédentes peuvent être combinées pour mesurer le remplacement des espèces selon deux gradients emboîtés (T6).

Une variation sur le modèle T2 consiste à le représenter sous la forme d'un modèle de décroissance en fonction de la distance,[10] éventuellement pour plusieurs groupes de communautés (Figure 5.1, T4 et T5). L'indice de Simpson spatialement explicite (page 147) entre dans ce cadre bien qu'il soit construit différemment.

[10] J. C. NEKOLA et P. S. WHITE (1999). « The Distance Decay of Similarity in Biogeography and Ecology ». In : *Journal of Biogeography* 26, p. 867–878.

Les mesures de variation (Figure 5.1, V1) concernent la dissimilarité entre plusieurs communautés ou entre chaque communauté

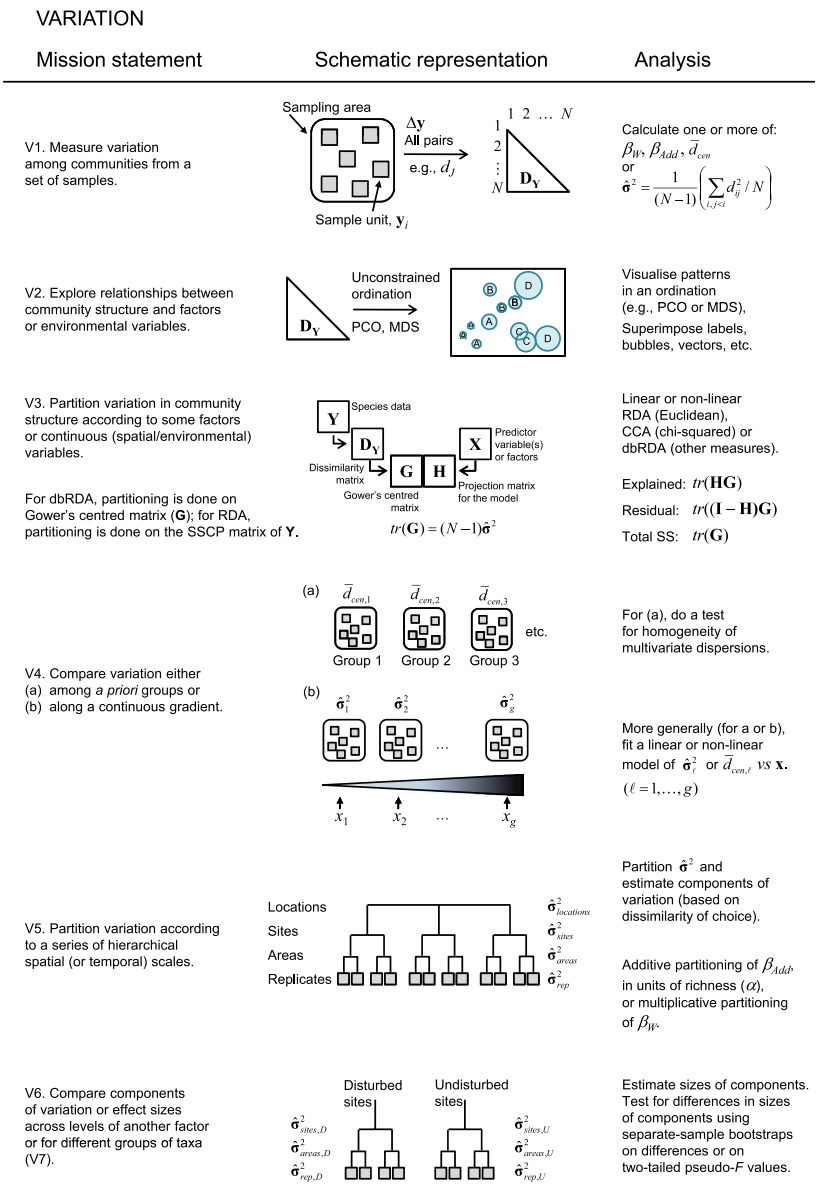


FIGURE 5.2 – Diversité β liée à la variation de la composition spécifique. β_W est le rapport entre la diversité γ et la diversité α, β_{Add} leur différence. d_{ij} est la dissimilarité entre les communautés i et j. $\hat{\sigma}^2$ est la variance de cette dissimilarité, $\bar{d}$ sa moyenne. Si d est une distance euclidienne, $\bar{d}_{cen}$ est la distance moyenne des communautés à leur centre de gravité. (*in* Anderson *et al.*[a])

[a] M. J. ANDERSON et al. (2011). « Navigating the multiple meanings of β diversity : a roadmap for the practicing ecologist ». In : *Ecology Letters* 14.1, p. 19–28.

et la méta-communauté. L'équivalence de ces deux concepts n'est pas évidente : elle a été montrée pour les mesures de dissimilarité entre paires de communautés par Legendre,[11] pour l'entropie de Rao par Pavoine[12] et pour l'entropie HCDT par Marcon *et al.*[13] La diversité β peut être mesurée pour un assemblage de communautés au choix comme la moyenne des dissimilarités entre les paires de communautés (voir chapitre 6, page 156) ou à partir des diversités γ et α, c'est-à-dire en tant que mesure dérivée. Cette dernière approche est traitée dans ce chapitre, où il sera montré que la mesure dérivée est aussi une dissimilarité moyenne entre communautés.

Si la diversité est mesurée comme moyenne des dissimilarités, elle peut être analysée par des méthodes d'ordination indirectes (Figure 5.2, V2) ou directes (V3). Elle peut aussi être comparée

[11] P. LEGENDRE et M. J. ANDERSON (1999). « Distance-based redundancy analysis : testing multispecies responses in multifactorial ecological experiments ». In : *Ecological Monographs* 69.1, p. 1–24 ; P. LEGENDRE et E. GALLAGHER (2001). « Ecologically meaningful transformations for ordination of species data ». In : *Oecologia* 129.2, p. 271–280 ; P. LEGENDRE et al. (2005). « Analyzing Beta Diversity : Partitioning the Spatial Variation of Community Composition Data ». In : *Ecological Monographs* 75.4, p. 435–450.

[12] S. PAVOINE et al. (2004). « From dissimilarities among species to dissimilarities among communities : a double principal coordinate analysis ». In : *Journal of Theoretical Biology* 228.4, p. 523–537.

entre méta-communautés ou modélisée en fonction d'un gradient environnemental (V4), de façon similaire au modèle T6.

La diversité β peut être décomposée en niveaux hiérarchiques emboîtés (V5). Des tests statistiques sont nécessaires pour comparer les niveaux : par exemple, montrer que la diversité (inter) entre des communautés supposées différentes est significativement plus grande que celle (intra) des placettes d'échantillonnage de chacune d'entre elles, à la manière d'une analyse de variance (V6).

[13] Marcon et al. (2014a). « Generalization of the Partitioning of Shannon Diversity », cf. note 38, p. 64.

5.1.2 Définitions de la diversité β, mesure dérivée

Tuomisto[14] passe en revue l'ensemble des définitions de la diversité β dérivée des diversités γ et α dans un article ardu mais très complet. Toutes ont en commun :

- Une définition de la mesure de diversité, appliquée à la diversité γ, qui est généralement une des mesures vues dans le chapitre 2 ;
- Une définition de la diversité α, qui peut être par exemple :
  - La diversité locale mesurée dans chaque communauté, indépendamment de toute référence hors de la communauté ;
  - De façon équivalente, le nombre d'espèces effectives dans les communautés.
- Une façon de combiner les diversités γ et α pour obtenir la diversité β, par exemple :
  - $\beta = \gamma/\alpha$;
  - $\beta = \gamma - \alpha$.

[14] Tuomisto (2010a). « A diversity of beta diversities : straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity », cf. note 6, p. 123.

L'utilisation des nombres de Hill, la mesure locale de la diversité α et la définition de la diversité β comme rapport des diversités γ et α permet de définir la « vraie diversité »[15] β qui est un nombre de communautés équivalentes (*compositionnal units*) similaire au nombre d'espèces équivalentes des diversités α et γ. Tuomisto[16] milite pour que le terme diversité soit réservé à la vraie diversité (homogène à un nombre d'espèces) et que les autres mesures soient appelées différemment : « entropie » de Shannon ou « probabilité » de Gini-Simpson notamment.

[15] Jost (2006). « Entropy and diversity », cf. note 35, p. 62 ; Jost (2007). « Partitioning diversity into independent alpha and beta components », cf. note 42, p. 17.

[16] H. Tuomisto (2011). « Commentary : do we have a consistent terminology for species diversity ? Yes, if we choose to use it ». In : *Oecologia* 167.4, p. 903–911.

5.1.3 Le débat sur la décomposition

L'objectif est de décomposer la diversité totale, notée γ en une composante inter-groupes, notée β et une composante intra-groupes notée α.

Whittaker[17] est l'auteur de ce concept. Il a posé le principe que la diversité γ devait être le produit des diversités α et β.

Lande[18] a une approche additive et postule que les mesures de diversité doivent être concaves : la diversité d'un jeu de données regroupant plusieurs communautés doit être supérieure ou égale à la somme pondérée des diversités dans chaque communauté. De

[17] Whittaker (1960). « Vegetation of the Siskiyou Mountains, Oregon and California », cf. note 23, p. 4 ; Whittaker (1972). « Evolution and Measurement of Species Diversity », cf. note 37, p. 8.

[18] Lande (1996). « Statistics and partitioning of species diversity, and similarity among multiple communities », cf. note 156, p. 46.

cette façon, il est possible de définir une diversité totale égale à la somme pondérée des diversités α (intra-communautés) et β (inter-communautés), toutes les diversités étant positives ou nulles. Il note que « la partition serait plus facilement interprétable si les différentes composantes de la diversité pouvaient être exprimés au moyen de la même formule » (ce qui n'est en fait jamais le cas). Une revue sur les avantages de la décomposition additive est proposée par Veech *et al.*.[19]

Un débat assez stérile a découlé de l'opposition entre les deux approches, principalement dû à la transformation logarithmique,[20] à des définitions imprécises et des démonstrations empiriques remises en question.[21] Il reste que la décomposition multiplicative permet seule la définition de la diversité β en tant que diversité au sens strict.[22]

Jurasinski *et al.*[23] distinguent plusieurs types de mesures de diversité :

- La diversité d'inventaire (*inventory diversity*), qui traite des données récoltées sur une unité spatiale, ce qui correspond à la définition des diversités α et γ ;
- La diversité de différenciation (*differentiation diversity*), qui mesure à quel point les unités spatiales sont différentes, ce qui correspond à la définition de la diversité β donnée plus haut ;
- La diversité proportionnelle (*proportional diversity*), diversité β qui se construit par différence ou rapport des diversités γ et α.

Marcon *et al.*[24] montrent que l'entropie de Shannon H_β est une mesure de diversité de différenciation en donnant sa définition indépendamment de H_α et H_γ, comme l'avaient fait Ricotta et Marignani pour l'entropie quadratique de Rao.[25] C'est également une diversité proportionnelle, comme toutes les mesures passées en revue par Tuomisto : la diversité de Shannon permet d'unifier les deux approches.

L'indice de Shannon, couplé à son expression sous forme de nombre de Hill, respecte finalement tous les critères imposés ou souhaités. Sa décomposition est détaillée ci-dessous. Enfin, Marcon *et al.*[26] généralisent ce résultat à toutes les mesures de diversités dérivées de l'entropie généralisée de Tsallis.

5.1.4 Décomposition multiplicative de la diversité

Jost[27] et Chao *et al.*[28] ont montré que la décomposition des nombres de Hill en éléments indépendants est multiplicative :

$$ {}^qD_\gamma = {}^qD_\alpha \, {}^qD_\beta \tag{5.1} $$

${}^qD_\alpha$ et ${}^qD_\gamma$ sont les nombres de Hill d'ordre q égaux aux diversités α et γ. Ce sont des nombres équivalents d'espèces. ${}^qD_\beta$ est

[19] J. A. Veech et al. (2002). « The additive partitioning of species diversity : recent revival of an old idea ». In : *Oikos* 99.1, p. 3–9.

[20] C. Ricotta (2005c). « On hierarchical diversity decomposition ». In : *Journal of Vegetation Science* 16.2, p. 223–226.

[21] Baselga (2010a). « Multiplicative partition of true diversity yields independent alpha and beta components ; additive partition does not », cf. note 4, p. 123 ; Veech et Crist (2010). « Diversity partitioning without statistical independence of alpha and beta », cf. note 4, p. 123.

[22] Chao et al. (2012). « Proposing a resolution to debates on diversity partitioning », cf. note 5, p. 123.

[23] G. Jurasinski et al. (2009). « Inventory, differentiation, and proportional diversity : a consistent terminology for quantifying species diversity ». English. In : *Oecologia* 159.1, p. 15–26.

[24] Marcon et al. (2012). « The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity », cf. note 13, p. 60.

[25] C. Ricotta et M. Marignani (2007). « Computing β-diversity with Rao's quadratic entropy : A change of perspective ». In : *Diversity and Distributions* 13.2, p. 237–241.

[26] Marcon et al. (2014a). « Generalization of the Partitioning of Shannon Diversity », cf. note 38, p. 64.

[27] Jost (2007). « Partitioning diversity into independent alpha and beta components », cf. note 42, p. 17.

[28] Chao et al. (2012). « Proposing a resolution to debates on diversity partitioning », cf. note 5, p. 123.

le « nombre de communautés effectives » ou « nombre équivalent de communautés », c'est-à-dire le nombre de communautés de poids égal ne possédant aucune espèce en commun (et dont la dissimilarité entre les espèces de communautés différentes est maximale quand la diversité est phylogénétique ou fonctionnelle),[29] qui fourniraient la même valeur de diversité β.

La diversité β est indépendante de la diversité α si les poids des communautés sont égaux. L'*indépendance* signifie que la valeur de ${}^{q}D_{\beta}$ n'est pas contrainte par celle de ${}^{q}D_{\alpha}$. Cette propriété est souvent considérée comme importante,[30] et sera discutée largement ici.

[29] S. Pavoine et al. (2015). « "Equivalent numbers" for species, phylogenetic, or functional diversity in a nested hierarchy of multiple scales ». In : *Methods in Ecology and Evolution* in rev.

[30] M. V. Wilson et a. Shmida (1984). « Measuring Beta Diversity with Presence-Absence Data ». In : *The Journal of Ecology* 72.3, p. 1055 ; Gregorius (2010). « Linking Diversity and Differentiation », cf. note 44, p. 64.

5.1.5 Définitions de la diversité α

La diversité α est calculée pour chaque communauté, et notée ${}^{q}_{i}D_{\alpha}$ (l'entropie correspondante est notée ${}^{q}_{i}H_{\alpha}$). La diversité α de la méta-communauté est intuitivement la moyenne de celle des communautés. En raison de leurs propriétés mathématiques, il est plus simple de considérer la moyenne des entropies α.

Le poids de chaque groupe est w_i, souvent choisi égal au nombre d'individus de la communauté divisé par le nombre total ou à la surface de chaque groupe, mais qui peut être arbitraire, tant que $p_s = \sum_i w_i p_{s,i}$. Deux façons de pondérer la somme émergent de la littérature.[31] La définition classique est selon Routledge[32] :

$$ {}^{q}H_{\alpha} = \sum_i w_i \, {}^{q}_{i}H_{\alpha} \tag{5.2} $$

$$ {}^{q}D_{\alpha} = \left(\sum_s \sum_i w_i p_{s,i}^{q} \right)^{1/(1-q)} \tag{5.3} $$

[31] Chao et al. (2012). « Proposing a resolution to debates on diversity partitioning », cf. note 5, p. 123.

[32] R. D. Routledge (1979). « Diversity indices : Which ones are admissible ? » In : *Journal of Theoretical Biology* 76.4, p. 503–515.

La pondération proposée par Jost[33] est :

$$ {}^{q}H_{\alpha} = \sum_i \frac{w_i^{q}}{\sum_i w_i^{q}} \, {}^{q}_{i}H_{\alpha} \tag{5.4} $$

$$ {}^{q}D_{\alpha} = \left(\sum_s \sum_i \frac{w_i^{q}}{\sum_i w_i^{q}} p_{s,i}^{q} \right)^{1/(1-q)} \tag{5.5} $$

[33] Jost (2007). « Partitioning diversity into independent alpha and beta components », cf. note 42, p. 17.

La première est la pondération naturelle. La seconde, qui utilise les poids à la puissance q, donne donc d'autant moins d'importance que q est grand aux communautés dont le poids est faible. Les deux définitions se confondent pour $q = 1$.

Si les poids sont différents, Jost[34] a montré que l'indépendance n'est possible que si la diversité α est définie selon sa pondération. En revanche, cette pondération ne garantit pas que la diversité γ soit supérieure à la diversité α si q n'est égal ni à 0 ni à 1. Jost conclut donc que la décomposition n'est possible que

[34] Ibid.

TABLE 5.2 – Définition explicite de la diversité de Shannon en tant que divergence entre une distribution observée et une distribution attendue. Les indices H_α et H_γ sont les compléments au logarithme du nombre d'espèces de cette divergence.

Indice	Distribution observée	Distribution attendue	Formule
H_α	Fréquence des espèces dans la communauté	Fréquences égales, hors formule	${}_iH_\alpha = \sum_s p_{s,i} \ln p_{s,i}$ $H_\alpha = \sum_i w_i \, {}_iH_\alpha$
H_β	Fréquence des espèces dans la communauté	Fréquence totale des espèces	${}_iH_\beta = \sum_s \hat{p}_{s,i} \ln p_{s,i}/p_s$ $H_\beta = \sum_i w_i \, {}_iH_\beta$
H_γ	Fréquence des espèces dans la méta-communauté	Fréquences égales, hors formule	$H_\gamma = \sum_s p_s \ln p_s$

pour des communautés de même poids ou pour $q = 0$ ou $q = 1$. Ricotta[35] contourne cette difficulté en proposant une méthode de ré-échantillonnage qui permet d'égaliser les poids.

Chiu *et al.*[36] établissent une nouvelle définition de la diversité α qui permet d'assurer l'indépendance et garantit que la diversité β est positive quels que soient les poids des communautés :

$$ {}^qD_\alpha = \frac{1}{I}\left(\sum_s \sum_i (w_i p_{s,i})^q\right)^{1/(1-q)} \tag{5.6}$$

Cette définition pose problème : alors que le choix de q a pour but de donner une importance plus ou moins grande aux espèces rares, son effet est le même sur la taille des communautés La diversité dépendra essentiellement des espèces rares dans les communautés de faible poids si q est petit, définissant une diversité bi-dimensionnelle. Cette question est traitée en détail page 172.

Marcon *et al.*[37] préfèrent la définition de Routledge qui est plus intuitive que garantit que la diversité γ est supérieure à la diversité α. La dépendance entre ${}^qD_\beta$ et ${}^qD_\alpha$ est le prix à payer.

5.2 Décomposition de l'entropie HCDT

5.2.1 Décomposition de l'indice de Shannon

Définition

La décomposition explicite (Tableau 5.2) est due à Marcon *et al.*[38] mais avait été établie plus ou moins clairement plusieurs fois dans la littérature, notamment par Rao et Nayak[39] : la diversité β est la divergence de Kullback-Leibler entre la distribution des espèces dans chaque parcelle et la distribution moyenne, appelée divergence de Jensen-Shannon.[40] Les diversités α et γ sont les compléments à $\ln S$ des indices de dissimilarité de Theil (3.1), divergences de Kullback-Leibler entre les distributions des espèces et leur équiprobabilité.

Lewontin,[41] dans son célèbre article sur la diversité génétique humaine, définit la diversité de Shannon inter-groupe comme la

[35] C. RICOTTA (2009). « Computing parametric beta diversity with unequal plot weights : A solution based on resampling methods ». In : *Theoretical Ecology* 2.1, p. 13–17.

[36] C.-H. CHIU et al. (2014b). « Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers ». In : *Ecological Monographs* 84.1, p. 21–44.

[37] MARCON et al. (2014a). « Generalization of the Partitioning of Shannon Diversity », cf. note 38, p. 64.

[38] MARCON et al. (2012). « The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity », cf. note 13, p. 60.

[39] C. R. RAO et T. K. NAYAK (1985). « Cross entropy, dissimilarity measures, and characterizations of quadratic entropy ». In : *Information Theory, IEEE Transactions on* 31.5, p. 589–593.

[40] J. LIN (1991). « Divergence Measures Based on the Shannon Entropy ». In : *IEEE Transactions on Information Theory* 37.1, p. 145–151.

[41] R. LEWONTIN (1972). « The apportionment of human diversity ». In : *Evolutionary biology* 6, p. 381–398.

différence entre la diversité totale et la diversité moyenne intra-groupe, mais ne cherche pas à explorer ses propriétés.

La forme de H_β a été établie par Ricotta et Avena,[42] sans la relier à celle de H_α et H_γ. Enfin, l'idée de la décomposition de la divergence de Kullback-Leibler, mais avec une approche différente, sans rapprochement avec l'indice de Shannon, a été publiée par Ludovisi et Taticchi.[43]

[42] C. Ricotta et G. C. Avena (2003a). « An information-theoretical measure of beta-diversity ». In : *Plant Biosystems* 137.1, p. 57–61.

[43] A. Ludovisi et M. I. Taticchi (2006). « Investigating beta diversity by Kullback-Leibler information measures ». In : *Ecological Modelling* 192.1-2, p. 299–313.

Les méta-communautés peuvent à leur tour être regroupées à un niveau supérieur, la diversité γ du niveau inférieur devenant diversité α pour le niveau supérieur. La décomposition ou le regroupement peuvent être effectués sur un nombre quelconque de niveaux.

Calcul sous R : La fonction `BetaEntropy` calcule l'entropie β de la méta-communauté (l'argument $q = 1$ permet de choisir l'entropie de Shannon).

```
# Pas de correction du biais d'estimation
summary(BetaEntropy(Paracou618.MC, q = 1, Correction = "None"))

## Neutral beta entropy of order 1 of metaCommunity
## Paracou618.MC with correction: None
##
## Entropy of communities:
##      P006      P018
## 0.3499358 0.5771645
## Average entropy of the communities:
## [1] 0.4471751
```

Biais d'estimation

L'estimateur du taux de couverture de la communauté i est noté $\hat{C}_i$, le taux de couverture total $\hat{C}$. L'estimateur corrigé de ${}_iH_\beta$ est :

$$ {}_i\tilde{H}_\beta = \sum_{i=1}^{s_{\neq 0}} \frac{\hat{C}_i \hat{p}_{s,i} \ln \frac{\hat{C}_i \hat{p}_{s,i}}{\hat{C} \hat{p}_s}}{1 - \left(1 - \hat{C}_i \hat{p}_{s,i}\right)^n} \tag{5.7} $$

Calcul sous R : `BetaEntropy` applique cette correction :

```
summary(BetaEntropy(Paracou618.MC, q = 1, Correction = "ChaoShen"))

## Neutral beta entropy of order 1 of metaCommunity
## Paracou618.MC with correction: ChaoShen
##
## Entropy of communities:
##      P006      P018
## 0.2955826 0.4520726
## Average entropy of the communities:
## [1] 0.3625503
```

Chao *et al.*[44] utilisent une approche différente pour corriger le biais d'estimation : l'information mutuelle. Les données sont organisées comme dans le tableau 5.1, mais les probabilités qu'un individu appartienne à la communauté i et à l'espèce s sont

[44] Chao et al. (2013). « Entropy and the species accumulation curve : a novel entropy estimator via discovery rates of new species », cf. note 29, p. 13.

calculées pour que leur somme soit égale à 1 sur l'ensemble du tableau : $\sum_s \sum_i p^*_{s,i} = 1$ (au lieu de $\sum_i p_{s,i} = 1$).

L'information mutuelle est la divergence de Kullback-Leibler entre la distribution de $\left\{p^*_{s,i}\right\}$ et celle du produit des distributions marginales $\{p_s\}$ et $\{p_i\}$. Si les distributions marginales sont indépendantes, la probabilité $p^*_{s,i}$ qu'un individu appartienne à l'espèce s et se trouve dans la communauté i est donnée par le produit des probabilités marginales p_s et p_i, et l'information mutuelle est nulle.

L'information mutuelle se calcule comme la différence entre l'entropie de Shannon calculée sur l'ensemble du tableau $H(\{p^*_{s,i}\})$ et les entropies marginales $H(\{p_s\})$ et $H(\{p_i\})$. $H(\{p_s\})$ est l'entropie γ de la méta-communauté. Si on choisit de pondérer les communautés par leurs effectifs ($w_i = p_i$), alors $H(\{p^*_{s,i}\}) - H(\{p_i\})$ est l'entropie α de la méta-communauté, c'est-à-dire la moyenne pondérée des entropies α des communautés. Dans ce cas, l'entropie β est égale à l'opposée de l'information mutuelle. Comme chaque terme du calcul est une entropie α, il peut être corrigé selon l'équation (2.61) pour obtenir un estimateur moins biaisé de H_β.

Calcul sous R :

```
Shannon(Paracou618.MC$Ns) + Shannon(colSums(Paracou618.MC$Nsi)) -
    Shannon(Paracou618.MC$Nsi)

## ChaoWangJost
##    0.3701823
```

Cet estimateur est le moins biaisé à ce jour, mais n'est applicable qu'à l'entropie de Shannon, sous la condition que la pondération des communautés soit donnée par les effectifs.

[45] Z. Zhang et M. Grabchak (2014b). « Nonparametric Estimation of Kullback-Leibler Divergence ». In : *Neural computation* 26.11, p. 2570–2593.

Zhang et Grabchak[45] proposent un estimateur de la divergence de Kullback-Leibler avec un biais réduit selon la même méthode que pour l'estimation de l'entropie HCDT ou de l'entropie de Shannon :

$$
{}_i\tilde{H}_\beta = \sum_{s=1}^{s\neq 0} \frac{n_{s,i}}{n}
\left\{ \sum_{v=1}^{n-n_s} \frac{1}{v} \prod_{j=1}^{v} \left(1 - \frac{n_s}{n-j+1}\right) - \sum_{v=1}^{n_i-n_{s,i}} \frac{1}{v} \prod_{j=1}^{v} \left(1 - \frac{n_{s,i}-1}{n_i-j}\right) \right\} \tag{5.8}
$$

Cette correction est incluse dans la fonction `ShannonBeta` de *entropart* :

```
# Divergence de Kullback-Leibler entre chaque communauté de
# Paracou618.MC et la métacommunauté
(Beta <- apply(Paracou618.MC$Nsi, MARGIN = 2, ShannonBeta, Nexp = Paracou618.MC$Ns,
    Correction = "ZhangGrabchak", Check = FALSE))

##      P006      P018
```

```
## 0.1442318 0.2994995

# Entropie beta
sum(Paracou618.MC$Wi * Beta)

## [1] 0.2106764
```

La correction de Zhang et Grabchak réduit beaucoup plus fortement l'estimation que celle de Chao et Shen. Le meilleur estimateur peut être évalué à partir de la différence entre les entropies γ et α, corrigées par le meilleur estimateur disponible (celui de Chao, Wang et Jost, voir page 48) :

```
# Entropie alpha de chaque communauté de Paracou618.MC
(Alpha <- apply(Paracou618.MC$Nsi, 2, Shannon, CheckArguments = FALSE))

##     P006     P018
## 4.384539 4.706510

# Entropie beta estimée par différence
bcShannon(Paracou618.MC$Ns) - sum(Paracou618.MC$Wi * Alpha)

## ChaoWangJost
##    0.3698367
```

L'estimateur de Chao et Shen est apparemment le meilleur, la correction de celui de Zhang et Grabchak est trop forte.

5.2.2 Décomposition de l'entropie généralisée

La décomposition additive de l'entropie selon Lande[46] doit être la suivante :

$$ {}^{q}H_{\gamma} = {}^{q}H_{\alpha} + {}^{q}H_{\beta} \tag{5.9} $$

Bourguignon[47] comme Lande[48] définissent une mesure d'inégalité décomposable comme respectant les propriétés suivantes :

- La population totale étant partitionnée, chaque partition recevant un poids w_i, la composante intra-groupe de la mesure H_{α} est égale à la somme pondérée des mesures dans chaque-groupe $H_{\alpha} = \sum_i w_i \, {}_iH_{\alpha}$;
- La composante intergroupe H_{β} est la mesure d'inégalité entre les groupes ;
- La mesure totale H_{γ} est la somme des mesures intra et intergroupes.

En passant par les nombres de Hill, Jost[49] montre que l'indice de Shannon est le seul pouvant être décomposé de cette façon.

La démonstration de Jost repose sur le postulat que la transformation de l'entropie en nombre de Hill, définie pour la diversité α, doit avoir la même forme pour la diversité β. La remise en cause de ce postulat[50] permet de décomposer la diversité d'ordre quelconque quel que soit le poids des communautés.

La décomposition de l'entropie est faite en écrivant le logarithme d'ordre q de la décomposition de la diversité, équation (5.1) :

[46] Lande (1996). « Statistics and partitioning of species diversity, and similarity among multiple communities », cf. note 156, p. 46.

[47] F. Bourguignon (1979). « Decomposable Income Inequality Measures ». In : *Econometrica* 47.4, p. 901–920.

[48] Lande (1996). « Statistics and partitioning of species diversity, and similarity among multiple communities », cf. note 156, p. 46.

[49] Jost (2007). « Partitioning diversity into independent alpha and beta components », cf. note 42, p. 17.

[50] Marcon et al. (2014a). « Generalization of the Partitioning of Shannon Diversity », cf. note 38, p. 64.

$$ {}^{q}H_{\gamma} = {}^{q}H_{\alpha} + \ln_q {}^{q}D_{\beta} - (q-1)\,{}^{q}H_{\alpha} \ln_q {}^{q}D_{\beta} \tag{5.10} $$

Ceci est la décomposition de Jost.[51] Jost désigne sous le nom de « composante β de l'entropie » le logarithme d'ordre q de la diversité, qu'il note H_B. La décomposition n'est pas additive, mais H_B est indépendant de l'entropie α.

[51] Jost (2007). « Partitioning diversity into independent alpha and beta components », cf. note 42, p. 17.

Les deux derniers termes peuvent être regroupés et réarrangés pour obtenir ${}^{q}H_{\beta}$ conformément à la décomposition additive de l'équation (5.9) :

$$ {}^{q}H_{\beta} = \sum_i w_i \sum_s p_{s,i}^{q} \ln_q \frac{p_{s,i}}{p_s} \tag{5.11} $$

L'entropie β est la somme pondérée par w_i (et pas autrement) des contributions de chaque communauté :

$$ {}^{q}_{i}H_{\beta} = \sum_s p_{s,i}^{q} \ln_q \frac{p_{s,i}}{p_s} \tag{5.12} $$

${}^{q}_{i}H_{\beta}$ est une divergence de Kullback-Leibler généralisée.[52] ${}^{q}H_{\beta}$ en est la moyenne sur l'ensemble des communautés, c'est donc une généralisation de la divergence de Jensen-Shannon.[53] ${}^{q}H_{\gamma}$ peut s'écrire sous la forme ${}^{q}H_{\gamma} = -\sum_s p_s^q \ln_q p_s$: la formalisation de la décomposition est une généralisation de la décomposition de l'entropie de Shannon, résumée dans le Tableau 5.3.

[52] L. Borland et al. (1998). « Information gain within nonextensive thermostatistics ». In : *Journal of Mathematical Physics* 39.12, p. 6490–6501 ; C. Tsallis et al. (1998). « The role of constraints within generalized nonextensive statistics ». In : *Physica A* 261.3, p. 534–554.

[53] Marcon et al. (2014a). « Generalization of the Partitioning of Shannon Diversity », cf. note 38, p. 64.

Il est intéressant d'écrire ${}^{q}_{i}H_{\beta}$ et ${}^{q}H_{\gamma}$ sous la forme d'entropies, pour faire apparaître leur fonction d'information. Pour q différent de 1, l'entropie γ est :

$$ {}^{q}H_{\gamma} = \sum_s p_s \frac{1 - p_s^{q-1}}{q-1} = \sum_s p_s \ln_q \frac{1}{p_s} \tag{5.13} $$

Sa fonction d'information a été tracée Figure 3.1, page 58.

L'entropie β est :

$$ {}^{q}_{i}H_{\beta} = \sum_s p_{s,i} \frac{p_{s,i}^{q-1} - p_s^{q-1}}{q-1} = \sum_s p_{s,i} \left(\ln_q \frac{1}{p_s} - \ln_q \frac{1}{p_{s,i}} \right) \tag{5.14} $$

La forme de l'entropie β évoque la divergence de puissance[54] qui est une statistique qui mesure la divergence entre deux distributions (déjà mentionnée page 75), dont les propriétés asymptotiques sont connues. La divergence de puissance peut être écrite sous la forme $I_q = {}^{-2}/_{q} \sum_s p_{s,i} \ln_q {}^{p_s}/_{p_{s,i}}$. Elle est égale à (deux fois) la divergence de Kullback-Leibler pour $q = 1$, mais diffère de l'entropie β pour les autres valeurs de q. La divergence non-logarithmique de Jensen-Shannon[55] est identique à une normalisation près : $K_q = -\sum_s p_{s,i} \ln_q {}^{p_s}/_{p_{s,i}}$. Elle mesure l'écart entre la diversité α et la valeur de $\sum_s w_i p_{s,i}^{q} \ln_q {p_s}^{q}$, qui n'est pas la diversité γ.[56]

[54] Cressie et Read (1984). « Multinomial Goodness-of-Fit Tests », cf. note 99, p. 75.

[55] P. W. Lamberti et A. P. Majtey (2003). « Non-logarithmic Jensen–Shannon divergence ». In : *Physica A : Statistical Mechanics and its Applications* 329.1–2, p. 81–90.

Calcul sous R : La fonction `DivPart` calcule entropie et diversité en même temps :

```
DivPart(q = 1, Paracou618.MC, Biased = FALSE)

## $MetaCommunity
## [1] "Paracou618.MC"
##
## $Order
## [1] 1
##
## $Biased
## [1] FALSE
##
## $Correction
## [1] "Best"
##
## $Normalized
## [1] TRUE
##
## $TotalAlphaDiversity
## [1] 92.04908
##
## $TotalBetaDiversity
## ChaoWangJost
##     1.447498
##
## $GammaDiversity
## ChaoWangJost
##     133.2409
##
## $CommunityAlphaDiversities
##      P006      P018
##  80.20125 110.66525
##
## $TotalAlphaEntropy
## [1] 4.522322
##
## $TotalBetaEntropy
## ChaoWangJost
##    0.3698367
##
## $GammaEntropy
## ChaoWangJost
##     4.892159
##
## $CommunityAlphaEntropies
##     P006     P018
## 4.384539 4.706510
##
## $CommunityBetaEntropies
## [1] NA
##
## $Method
## [1] "HCDT"
##
## attr(,"class")
## [1] "DivPart"
```

5.2.3 Décomposition du nombre d'espèces

Les résultats généraux se simplifient pour $q = 0$. L'entropie γ est le nombre d'espèces de la méta-communauté moins un (${}^0H_\gamma = S - 1$), l'entropie α est la moyenne pondérée du nombre d'espèces des communautés moins un (${}^0H_\alpha = \bar{S} - 1$), l'entropie β est la différence : ${}^0H_\beta = S - \bar{S}$.

En termes de diversité :

$$ {}^0D_\beta = \frac{S}{\bar{s}} \tag{5.15} $$

${}^0D_\beta$ peut être supérieur au nombre de communautés[57] si les poids sont très inégaux, ce qui complique son interprétation. La pondération de Jost donne pour le nombre d'espèces le même poids à toutes les communautés et garantit ${}^0D_\beta \leq I$, ce qui constitue pour Chao *et al.* un argument en faveur de son utilisation. Un contre-argumentaire est fourni par Marcon *et al.*.[58] Quel que soit la distribution des poids des communautés, il est toujours possible de ramener les données à un ensemble de communautés de poids identiques, dont certaines sont répétées. La communauté dont le poids w_{min} est le plus faible est représentée une seule fois, les autres ${}^{w_i}/\!/_{w_{min}}$ fois. L'indépendance entre les diversités α et β est donc bien vérifiée. Le diversité β maximale théorique est ${}^1/\!/_{w_{min}}$ (et non I). Elle n'est pas atteinte parce que plusieurs communautés sont identiques.

[57] Chao et al. (2012). « Proposing a resolution to debates on diversity partitioning », cf. note 5, p. 123.

[58] Marcon et al. (2014a). « Generalization of the Partitioning of Shannon Diversity », cf. note 38, p. 64.

La définition de la diversité β par Whittaker[59] s'appliquait aux données de présence-absence. La diversité β était le rapport entre le nombre d'espèces total de l'assemblage des communautés (diversité γ) et le nombre moyen d'espèces dans chaque communauté (diversité α). C'est précisément la définition de ${}^0D_\beta$ (équation 5.15).

[59] Whittaker (1960). « Vegetation of the Siskiyou Mountains, Oregon and California », cf. note 23, p. 4.

5.2.4 Décomposition de l'indice de Gini-Simpson

L'entropie de Simpson peut aussi être décomposée comme une variance. La probabilité qu'un individu appartienne à l'espèce s est p_s. Elle suit une loi de Bernoulli, dont la variance est $p_s\left(1-p_s\right)$. Cette variance peut être décomposée entre les communautés, où la probabilité est $p_{s,i}$. La décomposition entre variance intra et inter-communautés est :

$$ p_s\left(1-p_s\right) = \sum_i w_i \left[p_{s,i}\left(1-p_{s,i}\right) + \left(p_{s,i}-p_s\right)^2\right] \tag{5.16} $$

Cette égalité peut être sommée sur toutes les espèces pour donner :

$$ E_\gamma = \sum_i w_i E_{\alpha,i} + \sum_i w_i \sum_s \left(p_{s,i}-p_s\right)^2 = E_\alpha + E_\beta \tag{5.17} $$

L'entropie γ de Simpson peut être décomposée en sa diversité α, somme pondérée des entropies α des communautés, et son entropie β, égale à la somme pondérée des distances l^2 entre la distribution des fréquences dans chaque communauté et celle de l'ensemble de la communauté.

TABLE 5.3 – Entropie, diversité et décomposition. L'entropie α de chaque communauté se calcule comme l'entropie γ

Mesure de diversité	Entropie généralisée	Shannon
Entropie γ	${}^qH_\gamma = -\sum_s p_s^q \ln_q p_s$	${}^1H_\gamma = -\sum_s p_s \ln p_s$
Entropie β	${}^qH_\beta = \sum_i w_i \sum_s p_{s,i} \left(\ln_q \frac{1}{p_s} - \ln_q \frac{1}{p_{s,i}}\right)$	${}^1H_\beta = \sum_i w_i \sum_s p_{s,i} \ln \frac{p_{s,i}}{p_s}$
Diversité γ (Nombre de Hill)	${}^qD_\gamma = e_q^{{}^qH_\gamma}$	${}^1D_\gamma = e^{{}^1H_\gamma}$
Diversité β (nombre équivalent)	${}^qD_\beta = e_q^{\frac{{}^qH_\beta}{1-(q-1){}^qH_\alpha}}$	${}^1D_\beta = e^{{}^1H_\beta}$

En génétique, l'égalité s'écrit[60] $H_t = H_s + D_{s,t}$. H_t est l'hétérozygotie totale, décomposée en H_s, l'hétérozygotie moyenne des populations et $D_{s,t}$ la différenciation absolue entre populations. $G_{s,t} = D_{s,t}/H_t$ est la différenciation relative.

L'entropie E_β calculée de cette manière est égale à ${}^2H_\beta$ calculée à l'équation (5.14) pour le cas général, mais la contribution de chaque communauté est différente : la décomposition de la variance est une façon alternative de décomposer l'entropie, valable uniquement pour l'entropie de Simpson.

Cette décomposition est remise en cause par Jost,[61] qui a généré le débat.[62] E_γ étant inférieure ou égale à 1, E_β n'est pas indépendante de E_α : seule la décomposition multiplicative permet l'indépendance, et Jost propose d'utiliser une transformation de ${}^2D_\beta$ comme mesure de différenciation (voir page 143). Gregorius[63] montre que ce problème n'est pas limité à l'entropie de Simpson mais s'applique à toutes les entropies HCDT d'ordre supérieur à 1.

Jost postule que la diversité β de Simpson a la même relation à E_β que les diversités α et γ :

$$ {}^qD_\gamma = {}^qD_\alpha {}^qD_\beta \Leftrightarrow \frac{1}{1-E_\gamma} = \frac{1}{1-E_\alpha}\frac{1}{1-E_\beta} \tag{5.18} $$

D'où une décomposition additive différente, qui correspond à sa décomposition générale (5.10) :

$$ E_\gamma = E_\alpha + E_\beta - E_\alpha E_\beta \tag{5.19} $$

Il n'y a en réalité aucune raison pour que le nombre équivalent de la diversité β ait la même forme que celle de la diversité α : les deux diversités sont par nature très différentes, et Hill ne traitait pas la diversité β. La diversité de Shannon est exceptionnelle parce que, pour des raisons différentes, son nombre de Hill est l'exponentielle de l'entropie, et l'exponentielle de la décomposition additive (5.9) est la décomposition multiplicative (5.1).

[60] M. NEI (1973). « Analysis of Gene Diversity in Subdivided Populations ». In : *Proceedings of the National Academy of Sciences of the United States of America* 70.12, p. 3321–3323.

[61] JOST (2007). « Partitioning diversity into independent alpha and beta components », cf. note 42, p. 17 ; L. JOST (2008). « GST and its relatives do not measure differentiation ». In : *Molecular Ecology* 17.18, p. 4015–4026.

[62] R. HELLER et H. R. SIEGISMUND (2009). « Relationship between three measures of genetic differentiation G(ST), D-EST and G'(ST) : how wrong have we been ? » In : *Molecular Ecology* 18.10, p. 2080–2083 ; N RYMAN et O LEIMAR (2009). « G(ST) is still a useful measure of genetic differentiation - a comment on Jost's D ». English. In : *Molecular Ecology* 18.10, p. 2084–2087 ; L. JOST (2009a). « D vs. GST : Response to Heller and Siegismund (2009) and Ryman and Leimar (2009) ». In : *Molecular Ecology* 18.10, p. 2088–2091 ; M. C. WHITLOCK (2011). « G'ST and D do not replace FST ». In : *Molecular Ecology* 20.6, p. 1083–1091.

[63] GREGORIUS (2014). « Partitioning of diversity : the "within communities" component », cf. note 27, p. 61.

Le nombre équivalent de communautés a une forme légèrement différente d'un nombre de Hill (Tableau 5.3).

5.2.5 Décomposition de l'indice de Rao

Les notations sont celles de la présentation de l'indice de Rao, page 97.

La diversité de la méta-communauté, peut être décomposée en une somme (pondérée) de diversités intra et une diversité inter[64] en définissant une dissimilarité entre les communautés i et j :

$$D_{H_\Delta}(\mathcal{P}_i, \mathcal{P}_j) = 2H_\Delta\left(\frac{\mathcal{P}_i + \mathcal{P}_j}{2}\right) - H_\Delta(\mathcal{P}_j) - H_\Delta(\mathcal{P}_j) \quad (5.20)$$

[64] Pavoine et al. (2004). « From dissimilarities among species to dissimilarities among communities : a double principal coordinate analysis », cf. note 12, p. 126.

Elle est définie par Rao comme la différence entre l'entropie quadratique du mélange des deux communautés et celles des deux communautés prises individuellement. Elle peut être utilisée pour calculer un indice de diversité entre communautés, de la même façon qu'on calcule la diversité entre espèces d'une communauté :

$$H_\Delta(\mathcal{P}) = \sum_i w_i H_\Delta(\mathcal{P}_i) + \sum_i \sum_j w_i w_j D_{H_\Delta}(\mathcal{P}_i, \mathcal{P}_j) \quad (5.21)$$

Le deuxième terme de la somme est la diversité inter-communautés. C'est un indice de Rao : les poids des communautés sont équivalents aux tableaux de fréquences et les dissimilarités entre communautés sont équivalentes aux dissimilarités entre espèces de l'équation (4.19). La divergence entre communautés est construite à partir des entropies de chacune : l'entropie β est définie comme la différence entre les entropies γ et α.

L'exemple suivant utilise les communautés du fichier `rao.effectifs.csv` :

```
# Vecteur contenant 8 espèces...
(effectifs <- read.csv2("http://www.ecofog.gf/IMG/csv/rao.effectifs.csv",
    row.names = 1, header = T))

##     P1 P2 P3 P4
## Ess 97 36 34 61
## Me  62 24 49  4
## S1   4  6  6 75
## Sr  71 78 99 36
## Vm  76 29 34  2
## Bg  58 49 47 19
## Ef   9 34 91 35
## Dg  98 14 95 14
```

Il contient 4 communautés.

Pour obtenir tous les éléments de l'équation (5.21) :

- $H_\Delta(\mathcal{P})$: diversité totale. Calculer $\mathcal{P}$ en sommant les communautés. Calculer ensuite l'entropie quadratique avec `divc` :

```
(tabF <- apply(effectifs[names(phyf$leaves), ], MARGIN = 1, sum))

##  Sr Ess  Me  S1  Bg  Ef  Vm  Dg
## 284 228 139  91 173 169 141 221

divc(as.data.frame(tabF), phyf$Wdist)

##      diversity
## tabF  5.870667
```

- $H_{\Delta}(\mathcal{P}_i)$: diversité de chaque communauté, fournie par `divc` :

```
divc(effectifs[names(phyf$leaves), ], phyf$Wdist)

##    diversity
## P1  5.816719
## P2  5.897210
## P3  5.744483
## P4  4.790586
```

- La variabilité inter-communautés peut être calculée par `divc` en utilisant la distance fournie par `disc`, en définissant la matrice des poids (égale aux effectifs) :

```
divc(as.data.frame(colSums(effectifs)), disc(effectifs[names(phyf$leaves),
    ], phyf$Wdist))

##                    diversity
## colSums(effectifs) 0.2362187
```

La fonction `disc` calcule $\sqrt{2D_{H_{\Delta}}(\mathcal{P}_i, \mathcal{P}_j)}$:

```
disc(effectifs[names(phyf$leaves), ], phyf$Wdist)

##           P1        P2        P3
## P2 0.4698121
## P3 0.4888729 0.6487916
## P4 1.0018983 0.8776582 1.3157902
```

Une interprétation géométrique de la décomposition est la suivante.[65] Les espèces peuvent être placées dans un espace multidimensionnel construit par une Analyse en Coordonnées Principales[66] de la matrice de dissimilarité Δ (les distances dans l'espace multidimensionnel sont $\sqrt{2d_{s's''}}$; autrement dit : la dissimilarité entre deux espèces est la moitié du carré de la distance entre elles dans la représentation géométrique). Chaque communauté se trouve au centre de gravité des espèces qu'elle contient, pondérées par leur fréquence. La moitié du carré de la distance entre deux communautés dans ce même espace est le coefficient de dissimilarité entre communautés de Rao. La distance entre communautés est donc interprétable directement comme une mesure de diversité β,[67] dans la tradition de l'utilisation de la dissimilarité entre paires de communautés (chapitre 6).

Cette partition suppose que les poids des communautés sont proportionnels à leur nombre d'individus.[68] Hardy et Senterre[69] décomposent (par différence entre γ et α) l'indice de Rao de communautés de poids égaux. Hardy et Jost[70] montrent que les deux pondérations sont valides mais l'absence de cadre général

[65] Ibid.

[66] J. C. Gower (1966). « Some distance properties of latent root and vector methods used in multivariate analysis ». In : *Biometrika* 53.3, p. 325–338.

[67] C. Ricotta et al. (2015). « A classical measure of phylogenetic dissimilarity and its relationship with beta diversity ». In : *Basic and Applied Ecology* 16.1, p. 10–18.

[68] Rao (1982). « Diversity and dissimilarity coefficients : a unified approach », cf. note 5, p. 83 ; S. Villéger et D. Mouillot (2008). « Additive partitioning of diversity including species differences : a comment on Hardy & Senterre (2007) ». In : *Journal of Ecology* 96.5, p. 845–848.

[69] O. J. Hardy et B. Senterre (2007). « Characterizing the phylogenetic structure of communities by an additive partitioning of phylogenetic diversity ». In : *Journal of Ecology* 95.3, p. 493–506.

[70] O. J. Hardy et L. Jost (2008). « Interpreting and estimating measures of community phylogenetic structuring ». In : *Journal of Ecology* 96.5, p. 849–852.

assurant que la diversité γ de Rao est supérieure à la diversité α[71] motive Guiasu et Guiasu[72] à proposer une alternative à l'indice de Rao, l'indice de Gini-Simpson quadratique pondéré, dont la concavité est démontrée (ce qui implique que $\gamma \geq \alpha$). Il sera traité dans le chapitre consacré à la diversité de paires (page 155). La pondération peut en fait être quelconque tant que la matrice dont les éléments sont la racine carré des éléments de la matrice de dissimilarité est euclidienne.[73] Ce résultat sera obtenu pour les arbres ultramétriques dans le cadre plus général de la décomposition de l'entropie phylogénétique (voir section 5.2 page 130).

La partition peut être faite avec la fonction `DivPart` :

```
# Création d'une méta-communauté, poids égaux aux effectifs
MC <- MetaCommunity(effectifs, colSums(effectifs))
summary(DivPart(q = 2, MC, Tree = phyf, Normalize = FALSE))

## HCDT diversity partitioning of order 2 of metaCommunity MC
##
## Phylogenetic or functional diversity was calculated
## according to the tree
## phyf
##
## Diversity is not normalized
##
## Alpha diversity of communities:
##       P1       P2       P3       P4
## 24.19203 24.70000 23.75363 19.16686
## Total alpha diversity of the communities:
##        9
## 23.11553
## Beta diversity of the communities:
##        9
## 1.061198
## Gamma diversity of the metacommunity:
##        9
## 24.53015
```

La décomposition de la diversité (et non seulement de l'entropie) de Rao a été établie par Ricotta et Szeidl.[74]

5.2.6 Synthèse

La diversité mesurée par l'entropie généralisée peut être décomposée dans tous les cas de figure, y compris lorsque les poids des communautés ne sont pas égaux. Les formules d'entropie et de diversité sont résumées dans le Tableau 5.3. L'entropie β est une mesure de divergence entre la distribution de chaque communauté et celle de la méta-communauté, comme l'avaient suggéré Ricotta et Burrascano.[75]

L'entropie de Shannon est un cas particulier dans lequel toutes les controverses disparaissent : l'entropie β est indépendante de l'entropie α, et la pondération de Jost se confond avec celle de Routledge.

L'entropie de Simpson peut être décomposée de deux façons : comme une variance quand les poids des communautés sont donnés par leurs effectifs, ou selon le cas général. Les deux décompositions

[71] F. De Bello et al. (2010). « The partitioning of diversity : showing Theseus a way out of the labyrinth ». In : *Journal of Vegetation Science* 21.5, p. 992–1000.

[72] R. C. Guiasu et S. Guiasu (2011). « The weighted quadratic index of biodiversity for pairs of species : a generalization of Rao's index ». In : *Natural Science* 3.9, p. 795–801.

[73] Champely et Chessel (2002). « Measuring biological diversity using Euclidean metrics », cf. note 83, p. 99.

[74] Ricotta et Szeidl (2009). « Diversity partitioning of Rao's quadratic entropy », cf. note 101, p. 104.

[75] C. Ricotta et S. Burrascano (2009). « Testing for differences in beta diversity with asymmetric dissimilarities ». In : *Ecological Indicators* 9.4, p. 719–724.

produisent les mêmes valeurs d'entropie β, mais la contribution de chaque communauté n'est pas la même.

L'entropie peut être décomposée hiérarchiquement sur plusieurs niveaux.[76]

Les probabilités d'occurrence des espèces ne sont pas connues mais estimées à partir des données. Les diversités α et γ sont sous-estimées et la diversité β surestimée,[77] d'autant plus que l'ordre de diversité est faible (le biais est négligeable au-delà de l'indice de Simpson). Des méthodes de correction existent, mais pas pour l'entropie β à l'exception de Shannon. La méthode générale consiste donc à corriger les entropies α et γ, calculer l'entropie β par différence et transformer les résultats en nombres équivalents.[78]

[76] T. O. Crist et al. (2003). « Partitioning species diversity across landscapes and regions : A hierarchical analysis of alpha, beta, and gamma diversity ». In : *The American Naturalist* 162.6, p. 734–743 ; Marcon et al. (2012). « The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity », cf. note 13, p. 60.

[77] Marcon et al. (2012). « The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity », cf. note 13, p. 60 ; J. Beck et al. (2013). « Undersampling and the measurement of beta diversity ». In : *Methods in Ecology and Evolution* 4.4, p. 370–382.

[78] Marcon et al. (2014a). « Generalization of the Partitioning of Shannon Diversity », cf. note 38, p. 64 ; Marcon et al. (2014b). « The Decomposition of Similarity-Based Diversity and its Bias Correction », cf. note 115, p. 110.

5.3 Normalisation

La mesure de diversité ${}^qD_\beta$ est le nombre équivalent de communautés totalement distinctes qui fourniraient ce niveau de diversité.

5.3.1 Nécessité de normaliser la diversité β

Selon une première approche, cette mesure n'a de sens que comparée à nombre de communautés. La diversité entre un hectare de forêt tropicale et un hectare de forêt tempérée (de même poids, sans espèce commune) est égale à 2 quel que soit q. La diversité entre un nombre suffisant d'échantillons de forêt tropicale relativement homogène (avec de nombreuses espèces communes) peut facilement dépasser 2 : c'est le cas des 50 carrés de BCI pour $q = 0$. Pour ne pas conclure de façon erronée que les carrés de BCI sont plus différents entre eux que BCI et la forêt de Fontainebleau, Jost[79] suggère de diviser la diversité β par le nombre de communautés pour la normaliser. Ce raisonnement a un sens quand la diversité maximale envisageable est effectivement égale au nombre de communautés, par exemple si les communautés sont choisies dans des habitats différents : la diversité β mesure cette différence.

[79] Jost (2007). « Partitioning diversity into independent alpha and beta components », cf. note 42, p. 17.

Un autre cas est envisageable : les communautés peuvent être des échantillons d'une méta-communauté clairement définie. La diversité β mesure alors la variabilité de l'échantillonnage à l'intérieur de la méta-communauté à une échelle donnée. Pour une taille d'échantillon fixée (par exemple un hectare de forêt), la diversité β ne dépend pas du nombre de communautés : l'entropie β est l'espérance (estimée par la moyenne, équation 5.11, page 134) de la divergence de Jensen-Shannon entre la distribution des espèces d'une communauté et celle de la méta-communauté, fixe. Cette variabilité dépend de l'échelle de l'échantillonnage dans le sens où elle diminue si les échantillons sont de taille plus grande : l'entropie α de deux hectares de forêts est la diversité α moyenne

de chacun des deux hectares plus l'entropie β entre eux. L'entropie β entre les échantillons de deux hectares est donc diminuée de cette entropie β « intra » qui est absorbée par la nouvelle entropie α quand la taille des échantillons augmente.

La Figure 5.3 [80] présente les valeurs estimées de diversité neutre β entre carrés de BCI tirés aléatoirement, en fonction du nombre de carrés.

L'estimation de la diversité présente les difficultés classiques dues au biais d'estimation : la diversité γ est sous-estimée si l'échantillonnage est trop faible, la diversité β augmente avec le nombre de communautés comme la diversité γ augmente avec la taille de l'échantillon ; il s'agit d'un problème d'estimation, pas de normalisation. Le biais est sensible pour $q = 1$ malgré la correction fournie par l'estimateur de l'entropie (page 67). Pour $q = 2$, l'estimateur est sans biais, la diversité est très bien estimée même avec un échantillonnage réduit.

La diversité β ne doit donc pas être normalisée systématiquement, mais seulement dans le premier cas.

[80] Code pour réaliser la figure :

```
library(gplots)
MC <- MetaCommunity(as.data.frame(t(BCI)))

# Création d'une méta-communauté à partir de n carrés de BCI
SamplePlots <- function (n, q) {
  Plots <- sample(1:50, n)
  Nsi <- MC$Nsi[, Plots]; MCsample <- MetaCommunity(Nsi)
  dp <- DivPart(q, MCsample, Biased = FALSE)
  # Retourne les entropies alpha et gamma et la diversité beta
  return(c(alpha=as.numeric(dp$TotalAlphaEntropy),
           beta=as.numeric(dp$TotalBetaDiversity),
           gamma=as.numeric(dp$GammaEntropy)))
}
# Calcul de la diversité moyenne à partir des simulations
BetaPlots <- function (q, nPlots, nSimulations) {
  buiw <- b <- a <- g <- vector("numeric", length(nPlots)); i <- 1
  for (Size in nPlots) {
    Sims <- replicate(nSimulations, SamplePlots(Size, q))
    a[i] <- mean(Sims["alpha", ]); g[i] <- mean(Sims["gamma", ])
    b[i] <- expq(g[i], q)/expq(a[i], q)
    buiw[i] <- sd(Sims["beta", ])
    i <- i+1
  }
  dpBCI <- DivPart(q, MC, Biased = FALSE)
  plotCI(x=nPlots, y=b, uiw=buiw, col="black", barcol="blue", lwd=1,
    xlab="Nombre de carrés",
    ylab=expression(paste("Diversité ", beta, " d'ordre ", q)))
  abline(h=dpBCI$TotalBetaDiversity, lty=2)
}

BetaPlots(q=1, nPlots=c(2, 5, 10, 15, 20, 25, 30),
          nSimulations=100)
```

```
BetaPlots(q=2, nPlots=c(2, 5, 10, 15, 20, 25, 30),
          nSimulations=100)
```

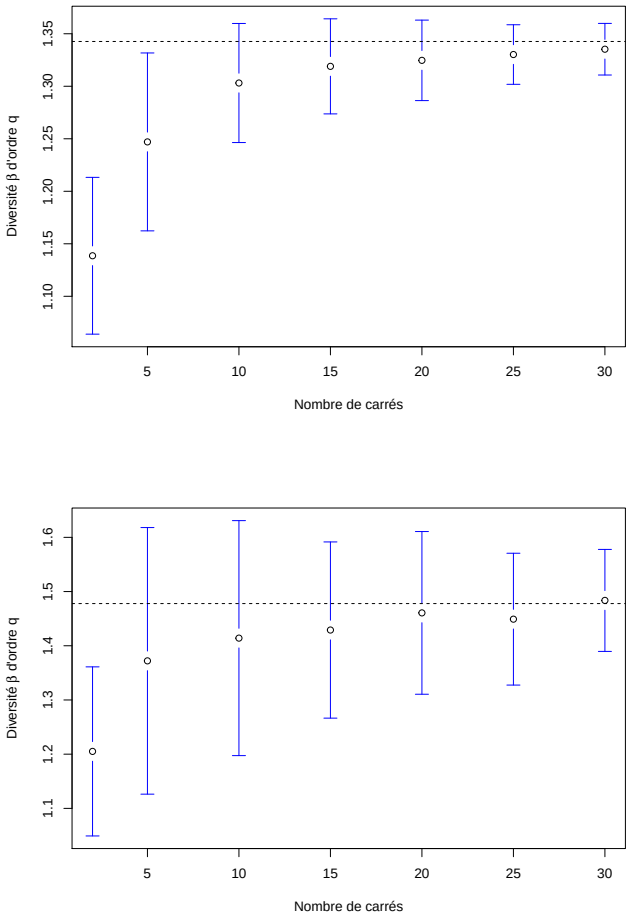


FIGURE 5.3 – Estimation de la diversité β entre carrés de BCI. La diversité est calculée à partir du nombre de carrés en abscisse, de 2 à 30, tirés aléatoirement. La valeur en ordonnée est la diversité moyenne (précisément la diversité calculée à partir de la moyenne des entropies) sur 100 tirages, les barres représentent l'écart-type. En haut, pour $q = 1$, la diversité est sous-estimée parce que la correction du biais d'estimation n'est pas suffisamment efficace. En bas, pour $q = 2$, la diversité β est correctement estimée avec de petits échantillons. Quand l'estimation est correcte, la valeur de la diversité β ne dépend pas du nombre de carrés.

5.3.2 L'indice de chevauchement C_{qN}

Plutôt qu'une simple normalisation, Chao *et al.*[81] définissent un indice de chevauchement, c'est-à-dire la proportion d'espèces partagées en moyenne par une communauté. Si N communautés (le nombre de communautés est noté N ici et non I pour conserver le nom de l'indice établi dans la littérature) de mêmes diversités α et γ que les données, contenaient chacune S espèces équifréquentes, dont A seraient communes à toutes les communautés et les autres représentées dans une seule, A/S serait cet indice de chevauchement.

[81] A. CHAO et al. (2008). « A Two-Stage Probabilistic Approach to Multiple-Community Similarity Indices ». In : *Biometrics* 64.4, p. 1178–1186.

Pour $q \neq 1$:

$$C_{qN} = \left[\left(\frac{1}{{}^{q}D_{\beta}}\right)^{q-1} - \left(\frac{1}{N}\right)^{q-1}\right] \Big/ \left[1 - \left(\frac{1}{N}\right)^{q-1}\right] \qquad (5.22)$$

Et :

$$C_{1N} = \frac{1}{\ln N} \sum_{s=1}^{S} \sum_{i=1}^{N} \frac{p_{s,i}}{N} \ln \left(1 + \frac{\sum_{j \neq i} p_{sj}}{p_{s,i}} \right) = 1 - \frac{H_\beta}{\ln N} \quad (5.23)$$

C_{qN} a été défini à l'origine pour $q \geq 2$. Chao *et al.*[82] l'étendent sans précaution à q quelconque. Sa valeur peut être négative pour $q < 1$, on se limitera donc à $q \geq 1$.

[82] Chao et al. (2012). « Proposing a resolution to debates on diversity partitioning », cf. note 5, p. 123.

5.4 Décomposition de la diversité phylogénétique

L'entropie phylogénétique est une combinaison linéaire de l'entropie HCDT donc sa décomposition est identique.[83] En combinant l'équation (4.22) et le Tableau 5.3, on obtient :

[83] É. Marcon et B. Hérault (2015a). « Decomposing Phylodiversity ». In : *Methods in Ecology and Evolution* 6.3, p. 333–339.

$$ {}^q\bar{H}_\gamma (T) = {}^q\bar{H}_\alpha (T) + {}^q\bar{H}_\beta (T) \quad (5.24)$$

L'entropie γ est celle de la méta-communauté, l'entropie α est la somme pondérée de celles des communautés, la pondération est libre mais doit respecter $p_s = \sum_i w_i p_{s,i}$:

$$ {}^q\bar{H}_\alpha (T) = \sum_i w_i \, {}^q_i\bar{H}_\alpha (T) \quad (5.25)$$

L'entropie β est :

$$ {}^q\bar{H}_\beta (T) = \sum_i w_i \, {}^q_i\bar{H}_\beta (T) \quad (5.26)$$

$$= \sum_i w_i \sum_k \frac{T_k}{T} \, {}^q_{ik}H_\beta \quad (5.27)$$

$$= \sum_i w_i \sum_k \frac{T_k}{T} \sum_u p^q_{k,u,i} \ln_q \frac{p_{k,u,i}}{p_{k,u}} \quad (5.28)$$

où $p_{k,u}$ est la probabilité d'observer une des espèces de la branche u de l'arbre à la période k (Figure 4.7) dans la méta-communauté, et $p_{k,u,i}$ la même probabilité dans la communauté i.

La décomposition de l'entropie phylogénétique (au sens d'Allen *et al.*[84]), c'est-à-dire dans la cas particulier $q = 1$, a été établie par Mouchet et Mouillot.[85]

[84] Allen et al. (2009). « A New Phylogenetic Diversity Measure Generalizing the Shannon Index and Its Application to Phyllostomid Bats », cf. note 95, p. 101.

[85] M. A. Mouchet et D. Mouillot (2011). « Decomposing phylogenetic entropy into α, β and γ components ». In : *Biology Letters* 7.2, p. 205–209.

La fonction `DivPart` de *entropart* permet de décomposer l'entropie. Les résultats contiennent la valeur de la diversité :

```
data(Paracou618)
summary(DivPart(q = 1, Paracou618.MC, Biased = FALSE,
                Tree=Paracou618.Taxonomy))
```

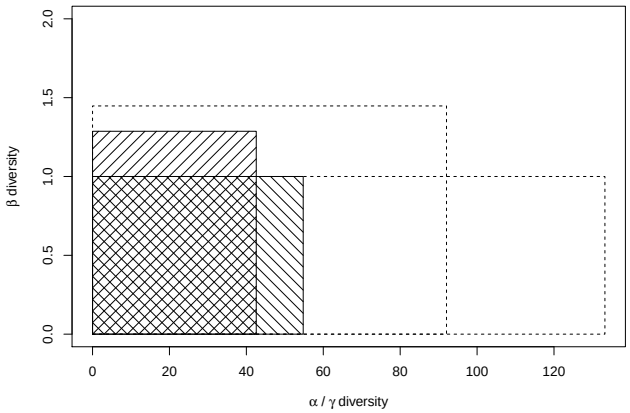


FIGURE 5.4 – Représentation graphique de la diversité des deux parcelles du jeu de données `Paracou618`. Les rectangles transparents représentent la diversité neutre, les rectangles hachurés la diversité phylogénétique (l'arbre est la taxonomie des espèces, $q = 1$). Dans chaque cas, le rectangle le plus large, de hauteur 1, représente la diversité γ. Le rectangle plus haut a pour largeur et longueur les diversités α et β. Les surfaces des deux rectangles sont identiques.

```
## HCDT diversity partitioning of order 1 of metaCommunity
## Paracou618.MC
##  with correction: Best
## Phylogenetic or functional diversity was calculated
## according to the tree
## Paracou618.Taxonomy
##
## Diversity is normalized
##
## Alpha diversity of communities:
##     P006     P018
## 36.66108 51.98951
## Total alpha diversity of the communities:
## [1] 42.57235
## Beta diversity of the communities:
## [1] 1.286827
## Gamma diversity of the metacommunity:
## [1] 54.78324
```

La Figure 5.4 [86] donne une représentation graphique de la décomposition de la diversité. La différence de taille entre les rectangles hachurés et les rectangles transparents est due à la prise en compte de la phylogénie. La diversité γ (rectangles de hauteur 1) est égale au produit des diversités α et β (côtés du rectangle plus haut). Si les communautés étaient complètement différentes, la hauteur du rectangle serait égale à leur nombre (2

[86] Code pour réaliser la figure :

```
data(Paracou618)
q <- 1
Neutre <- DivPart(q, Paracou618.MC, Biased = FALSE)
plot(c(0, Neutre$GammaDiversity), c(0, Paracou618.MC$Ncommunities),
     type = "n",
     xlab = expression(paste(alpha, " / ", gamma, " diversity")),
     ylab = expression(paste(beta, " diversity")))
rect(0, 0, Neutre$GammaDiversity, 1, lty=2)
rect(0, 0, Neutre$TotalAlphaDiversity, Neutre$TotalBetaDiversity, lty=2)
Phylo <- DivPart(q, Paracou618.MC, Biased = FALSE,
                 Tree=Paracou618.Taxonomy)
rect(0, 0, Phylo$GammaDiversity, 1, angle=135, density=10)
rect(0, 0, Phylo$TotalAlphaDiversity, Phylo$TotalBetaDiversity,
     density=10)
```

ici).

5.5 Partitionnement de la diversité de Leinster et Cobbold

Selon les mêmes principes que pour la décomposition de l'entropie HCDT, Marcon *et al.*[87] ont décomposé l'entropie de Ricotta et Szeidl.[88] L'entropie α de la communauté i est :

$$ {}^{q}_{i}H^{\boldsymbol{Z}}_{\alpha} = \frac{1 - \sum_s p_{s,i} (\boldsymbol{Z}\boldsymbol{p})^{q-1}_{s,i}}{q-1} \tag{5.29} $$

L'entropie α est la somme pondérée des entropies des communautés :

$$ {}^{q}H^{\boldsymbol{Z}}_{\alpha} = \sum_i w_i \, {}^{q}_{i}H^{\boldsymbol{Z}}_{\alpha} = \frac{1 - \sum_i w_i \sum_s p_{s,i} (\boldsymbol{Z}\boldsymbol{p})^{q-1}_{s,i}}{q-1} \tag{5.30} $$

L'entropie β est similaire à l'entropie β HCDT : c'est la divergence généralisée de Jensen-Shannon entre les distributions de $\boldsymbol{Z}\boldsymbol{p}$ des communautés. Formellement :

$$ {}^{q}H^{\boldsymbol{Z}}_{\beta} = \sum_i w_i \sum_s p_{s,i} \left(\ln_q \frac{1}{(\boldsymbol{Z}\boldsymbol{p})_s} - \ln_q \frac{1}{(\boldsymbol{Z}\boldsymbol{p})_{s,i}} \right) \tag{5.31} $$

La diversité β est obtenue en calculant l'exponentielle déformée de la décomposition additive de l'entropie :

$$ {}^{q}D^{\boldsymbol{Z}}_{\beta} = e_q^{\frac{{}^{q}H^{\boldsymbol{Z}}}{1+(1-q)\,{}^{q}H^{\boldsymbol{Z}}_{\alpha}}} \tag{5.32} $$

La décomposition de la diversité peut être faite directement, sans passer par l'entropie.

L'inverse de ${}^{q}D^{\boldsymbol{Z}}_{\gamma}$ est la moyenne généralisée d'ordre $q-1$ de $(\boldsymbol{Z}\boldsymbol{p})_s$:

$$ \frac{1}{{}^{q}D^{\boldsymbol{Z}}_{\gamma}} = \left[\sum_i p_s (\boldsymbol{Z}\boldsymbol{p})_s{}^{q-1} \right]^{\frac{1}{q-1}} \tag{5.33} $$

Partant de ${}^{q}D_{\gamma} = {}^{q}D_{\alpha} \, {}^{q}D_{\beta}$ et de la définition de Routledge de la diversité α, on obtient :

$$ {}^{q}D^{\boldsymbol{Z}}_{\beta} = \left[\sum_i w_i \left(\frac{1/\,{}^{q}_{i}D^{\boldsymbol{Z}}_{\alpha}}{1/\,{}^{q}D^{\boldsymbol{Z}}_{\gamma}} \right)^{q-1} \right]^{\frac{1}{q-1}} \tag{5.34} $$

De la même façon que $1/\,{}^{q}D^{\boldsymbol{Z}}_{\gamma}$ est la banalité moyenne des espèces, ${}^{q}D^{\boldsymbol{Z}}_{\beta}$ est la banalité moyenne normalisée des communautés,

[87] Marcon et al. (2014b). « The Decomposition of Similarity-Based Diversity and its Bias Correction », cf. note 115, p. 110.

[88] Ricotta et Szeidl (2006). « Towards a unifying approach to diversity measures : Bridging the gap between the Shannon entropy and Rao's quadratic index », cf. note 109, p. 108.

où la banalité normalisée de la communauté i est définie comme $\left(1/{}_{i}^{q}D_{\alpha}^{Z}\right)/\left(1/{}^{q}D_{\gamma}^{Z}\right)$, c'est-à-dire la banalité moyenne de ses espèces divisée par celle de la méta-communauté.

La fonction `DivPart` (comme les autres fonctions avancées de *entropart*) supporte la décomposition de la diversité de Leinster et Cobbold quand une matrice de similarité lui est passée comme argument à la place d'un arbre phylogénétique pour le calcul de la phylodiversité HCDT :

```
# Matrice de distances fonctionnelles
DistanceMatrix <- as.matrix(Paracou618.dist)
# Transformation en matrice de similarité
Z <- 1 - DistanceMatrix/max(DistanceMatrix)
# Partition de la diversité
summary(DivPart(q = 1, Paracou618.MC, Z = Z, Biased = FALSE))

## Similarity-based diversity partitioning of order 1
## of metaCommunity Paracou618.MC
##  with correction: Best
## Phylogenetic or functional entropy was calculated
## according to the similarity matrix
## Z
##
## Alpha diversity of communities:
##     P006     P018
## 1.483722 1.482119
## Total alpha diversity of the communities:
## [1] 1.483036
## Beta diversity of the communities:
##     Best
## 1.003449
## Gamma diversity of the metacommunity:
##     Best
## 1.488151
```

Le partitionnement de la diversité fonctionnelle de la méta-communauté `Paracou618.MC` selon la matrice de distances des traits fonctionnels montre que la diversité β est extrêmement faible entre les deux parcelles.

5.6 Autres approches

Des définitions différentes de la diversité β ou du partitionnement de la diversité sont présentées ici parce qu'elles sont importantes dans la littérature même si elles n'entrent pas dans le cadre précédent.

5.6.1 L'indice de Simpson spatialement explicite

Chave et Leigh[89] écrivent un modèle neutre de dispersion des espèces forestières, appliqué à l'échelle de l'Amazonie occidentale.[90] Ils définissent dans ce cadre la fonction $F(r)$ égale à la probabilité que deux individus situés à distance r l'un de l'autre soient de la même espèce, c'est-a-dire l'indice de concentration de Simpson (2.46), page 46, appliqué à des paires d'individus distants. $\beta_s(r) = 1 - F(r)$ est appelé « indice de Simpson spatialement explicite » par Shen *et al.*.[91] En pratique il est estimé à

[89] J. Chave et E. G. Leigh (2002). « A Spatially Explicit Neutral Model of β-Diversity in Tropical Forests ». In : *Theoretical Population Biology* 62.2, p. 153–168.

[90] R. Condit et al. (2002). « Beta-diversity in tropical forest trees ». In : *Science* 295.5555, p. 666–669.

[91] G. Shen et al. (2013). « Quantifying spatial phylogenetic structures of fully stem-mapped plant communities ». In : *Methods in Ecology and Evolution* 4.12, p. 1132–1141.

partir d'individus situés dans des parcelles distantes de r, l'ordre de grandeur de r étant nettement plus grand que la taille des parcelles.

De la même façon que l'indice de Simpson est généralisable à l'indice de Rao, $\beta_s(r)$ est généralisable à $\beta_{phy}(r)$, l'indice de Rao spatialement explicite, égal à la divergence moyenne entre deux individus situés à distance r.

Les deux mesures sont des entropies β, au sens où ce sont des diversités de différenciation,[92] voir page 127, qui mesurent à quel point des unités spatiales distantes de r sont différentes. Ce ne sont pas des mesures de diversité proportionnelle parce que ce ne sont pas les différences entre les entropies γ (de la méta-communauté composée de toutes les parcelles) et α (la moyenne des entropies des parcelles). L'avantage de ces mesures est d'être facilement interprétables, estimables sans biais (en tant qu'entropies de Simpson ou de Rao), et indépendantes de la diversité α.

[92] Jurasinski et al. (2009). « Inventory, differentiation, and proportional diversity : a consistent terminology for quantifying species diversity », cf. note 23, p. 128.

Une décomposition due à Nei[93] reprise par Chave *et al.*[94] montre que $\beta_s(r)$ est en réalité un mélange d'entropie α et β. Notons $E_{i,j}$ la probabilité que deux individus tirés respectivement dans les parcelles i et j soient d'espèces différentes :

[93] Nei (1973). « Analysis of Gene Diversity in Subdivided Populations », cf. note 60, p. 137.

[94] Chave et al. (2007). « The importance of phylogenetic structure in biodiversity studies », cf. note 8, p. 83.

$$E_{i,j} = 1 - \sum_s p_{s,i} p_{sj} \tag{5.35}$$

E_{ii} est l'entropie de Simpson calculée dans la parcelle i. L'entropie β de Simpson (équation 5.17, dans le cadre de la décomposition de la variance) pour des parcelles de poids égal vaut :

$$E_\beta = \frac{1}{I^2} \sum_i \sum_{j \neq i} \left(E_{i,j} - \frac{E_{ii} + E_{jj}}{2} \right) \tag{5.36}$$

En se limitant à deux parcelles i et j situées à distance r, le premier terme de la somme $E_{i,j}$ est l'indice de Simpson spatialement explicite, qui doit être corrigé du deuxième terme de la somme, l'entropie α de la paire de parcelle, pour obtenir l'entropie β après la normalisation nécessaire ($E_{i,j} = E_{j,i}$ est compté deux fois dans la somme mais divisé par $I^2 = 4$). L'extension à $\beta_{phy}(r)$ est immédiate, en remplaçant l'entropie de Simpson par celle de Rao.

Il est donc préférable de calculer la diversité β de Simpson ou de Rao plutôt que $\beta_s(r)$ ou $\beta_{phy}(r)$.

5.6.2 Le variogramme de complémentarité

Wagner[95] définit le variogramme de complémentarité à partir des techniques de géostatistique. Les espèces sont inventoriées dans des parcelles de taille identique, en présence-absence. Les distances entre parcelles sont regroupées par classes. L'indicatrice

[95] H. H. Wagner (2003). « Spatial covariance in plant communities : Integrating ordination, geostatistics, and variance testing ». In : *Ecology* 84.4, p. 1045–1057.

$\mathbf{1}_{s,i}$ vaut 1 si l'espèce s est présente dans la parcelle i, 0 sinon. n_r est le nombre de paires de parcelles i et j situées dans la classe de distances dont la valeur centrale est r (leur distance exacte est $d(i,j)$). La semi-variance empirique à distance r de la variable indicatrice de la présence de l'espèce s est :

$$\hat{\gamma}_s(r) = \frac{1}{2n_r} \sum_{i,j;d(i,j)\approx r} (\mathbf{1}_{s,i} - \mathbf{1}_{s,j})^2 \tag{5.37}$$

Chaque terme de la somme vaut 1 si l'espèce est présente sur une seule des deux parcelles : $\hat{\gamma}_s(r)$ estime donc la probabilité que l'espèce ne soit présente que sur une seule parcelle. Le variogramme de la variable indicatrice est obtenue en traçant la valeur de $\hat{\gamma}_s(r)$ en fonction de r. Ce variogramme univarié, classique en géostatistique, est généralisé aux données multivariées : le vecteur de présence des espèces $\boldsymbol{S_i}$ dont chaque terme est $\mathbf{1}_{s,i}$. La semi-variance de ce vecteur de composition spécifique est :

$$\hat{\gamma}(r) = \frac{1}{2n_r} \sum_{i,j;d(i,j)\approx r} \|\boldsymbol{S_i} - \boldsymbol{S_j}\|^2 = \sum_s \hat{\gamma}_s(r) \tag{5.38}$$

$\|\boldsymbol{S_i} - \boldsymbol{S_j}\|$ est la distance euclidienne entre les deux vecteurs. Puisque $\hat{\gamma}(r)$ est la somme des semi-variances spécifiques, c'est un estimateur de l'espérance du nombre d'espèces présentes sur une seule parcelle de chaque paire. Le variogramme de la composition spécifique est appelé « variogramme de complémentarité » pour cette raison.

Bacaro et Ricotta[96] l'utilisent en tant que mesure de diversité β.

[96] G. Bacaro et C. Ricotta (2007). « A spatially explicit measure of beta diversity ». In : *Community Ecology* 8.1, p. 41–46.

$\mathbb{E}(\mathbf{1}_s)$ est la probabilité qu'une parcelle quelconque contienne l'espèce s. Ce n'est pas la probabilité p_s qu'un individu appartienne à l'espèce s : comme toutes les parcelles sont de taille identique, p_s est égal à $\mathbb{E}(\mathbf{1}_s)/\sum_s \mathbb{E}(\mathbf{1}_s)$.

En absence de structuration spatiale, ou à partir de la valeur de r appelée *portée*, au-delà de laquelle la variance se stabilise à son *palier*, $\hat{\gamma}(r)$ est égal à $\sum_s \mathbb{E}(\mathbf{1}_s)[1 - \mathbb{E}(\mathbf{1}_s)]$. $\hat{\gamma}(r)$ n'est pas une forme de l'indice de Simpson parce que la somme des $\mathbb{E}(\mathbf{1}_s)$ ne vaut pas 1.

Le variogramme peut être standardisé en divisant toutes les valeurs par la valeur attendue du palier, $\sum_s \mathbb{E}(\mathbf{1}_s)[1 - \mathbb{E}(\mathbf{1}_s)]$, pour éliminer l'effet de la diversité et se concentrer sur la structuration spatiale.

Le variogramme peut être complété par un covariogramme. La covariance entre les indicatrices de présence des espèces s et t est :

$$\hat{\gamma}_{s,t}(r) = \frac{1}{2n_r} \sum_{i,j;d(i,j)\approx r} (\mathbf{1}_{s,i} - \mathbf{1}_{s,j})(\mathbf{1}_{t,i} - \mathbf{1}_{t,j}) \tag{5.39}$$

Table 5.4 – Tableau d'occurrence des espèces.

			Forêt k	...	Total
			Parcelle j	...	
			Placette i	...	
			Arbre h	...	
Famille u	Genre t	Espèce s	$n_{sh} = 1$ si l'arbre h appartient à l'espèce s, 0 sinon. $n_{s\cdot} = n_{s+}/n$		
...	...	...			
Total			$q_{+h} = p_{+h} = 1$		1

Wagner montre que l'estimateur de la variance de la richesse spécifique par parcelle est :

$$\hat{\mathrm{Var}}(S) = \sum_r \frac{n_r}{\sum_r n_r} \sum_{s,t} \hat{\gamma}_{s,t}(r) \tag{5.40}$$

Si les espèces sont distribuées indépendamment les unes des autres, la covariance est nulle pour toutes les paires d'espèces à toutes les distances. Le test d'égalité entre $\mathrm{Var}(S)$ et $\gamma_{s,t}(r)$ à chaque valeur de r (obtenu par la méthode de Monte-Carlo en redistribuant aléatoirement la présence de chaque espèce sur le même nombre de parcelles que dans les données réelles) permet de rejeter l'indépendance entre les espèces.

5.6.3 Le partitionnement de la diversité selon Pélissier et Couteron

Cadre

Un cadre général de partitionnement de la diversité, applicable aux trois indices classiques, est proposé par Pélissier et Couteron,[97] en conclusion d'une série d'articles.[98]

Les données sont désagrégées jusqu'au niveau de l'individu. Suivant les notations précédentes, les placettes sont considérées comme le regroupement d'arbres, indicés par h.

Le Tableau 5.4 est appelé tableau d'occurrence des espèces. Il ne contient que des 0 et des 1. Chaque arbre (colonne) a une seule valeur 1 dans la cellule correspondant à son espèce.

Un arbre moyen s'il pouvait exister appartiendrait partiellement à chaque espèce, proportionnellement à son effectif. Chaque arbre h apporte une quantité d'information égale à la somme sur toutes les espèces des écarts quadratiques entre la valeur n_{sh}, indicatrice de son appartenance à l'espèce s, et la valeur $n_{s\cdot}$ correspondant à l'arbre moyen : $\sum_s (n_{sh} - n_{s\cdot})^2$. La somme peut être pondérée pour donner une importance plus ou moins grande des espèces selon leur fréquence : $\sum_s w_s (n_{sh} - n_{s\cdot})^2$. La mesure totale, sommée sur tous les arbres, est notée TSS :

[97] R. Pélissier et P. Couteron (2007). « An operational, additive framework for species diversity partitioning and beta-diversity analysis ». In : *Journal of Ecology* 95.2, p. 294–300.

[98] R. Pélissier et al. (2003). « Consistency between ordination techniques and diversity measurements : Two strategies for species occurrence data ». In : *Ecology* 84.1, p. 242–251 ; P. Couteron et R. Pélissier (2004). « Additive apportioning of species diversity : towards more sophisticated models and analyses ». In : *Oikos* 107.1, p. 215–221.

$$TSS = \sum_s w_s \sum_h (n_{sh} - n_{s\cdot})^2 \tag{5.41}$$

Si le poids w_s vaut toujours 1, TSS/n est l'indice de Simpson.

Si $w_s = \ln(1/n_{s\cdot})/(1 - n_{s\cdot})$, on obtient l'indice de Shannon, et $w_s = 1/w_{s\cdot}$ donne le nombre d'espèces moins 1.

Décomposition

L'intérêt de la méthode est de permettre l'utilisation de modèles linéaires et des outils informatiques associés classiques pour décomposer la diversité de différentes manières. Nous nous intéresserons ici à la décomposition de la diversité totale γ en diversités α et β. Dans ce cadre, la diversité β est la partie expliquée par le modèle (les différentes communautés sont les paramètres explicatifs), notée *MSS*, alors que la diversité α est résiduelle, notée *RSS*.

La diversité totale est TSS/n. La diversité α est calculée de la même manière, en prenant pour référence la fréquence des espèces à l'intérieur de chaque placette. Dans la placette i :

$$RSS_i = \sum_s w_s \sum_{h \in \mathcal{I}} \left(n_{sh} - \frac{\sum_{h \in \mathcal{I}} n_{sh}}{n} \right)^2 \tag{5.42}$$

La diversité β est calculée de façon similaire par la différence entre les fréquences des espèces dans chaque placette $(\sum_{h \in \mathcal{I}} n_{sh})/n$ et leur fréquence après regroupement $n_{s\cdot}$.

Le problème est que le poids attribué à chaque espèce, w_s, dépend de la distribution totale des espèces. *RSS* n'est donc pas une diversité α puisque sa valeur dépend de données extérieures à la placette.

Le poids $w_s = 1$ pour toutes les espèces règle le problème. Le modèle est alors celui de la décomposition de l'indice de Simpson présenté plus haut.

5.7 Tests de significativité

5.7.1 Test de non-nullité

L'objectif est de tester si deux communautés ne sont pas simplement deux échantillons d'une même communauté, dont les différences ne sont que des fluctuations dues au hasard. Sous l'hypothèse nulle, les observations $\hat{q}_{s,i}$ sont des réalisations des mêmes probabilités p_s. Il est possible[99] mais controversé[100] de tester la significativité de la différence entre communautés. Une meilleure approche consiste à calculer l'intervalle de confiance de la diversité β due à l'incertitude sur les estimateurs de probabilités.

Le test est réalisé de la façon suivante :[101]

[99] Crist et al. (2003). « Partitioning species diversity across landscapes and regions : A hierarchical analysis of alpha, beta, and gamma diversity », cf. note 76, p. 141.

[100] D. Jones et N. Matloff (1986). « Statistical Hypothesis Testing in Biology : A Contradiction in Terms ». In : *Journal of Economic Entomology* 79.5, p. 1156–1160.

[101] Marcon et al. (2014a). « Generalization of the Partitioning of Shannon Diversity », cf. note 38, p. 64.

- Les effectifs de chaque communauté i sont tirés dans une loi multinomiale $\mathcal{M}(n_{+i}, n_{s,i}/n_{+i})$ et $\tilde{H}_\beta$ est estimé ;
- La simulation est répétée un grand nombre de fois, par exemple 10000, et les valeurs extrêmes sont éliminées. Au seuil de risque $\alpha = 5\%$, les 251^{e} et 9750^{e} valeurs simulées définissent les bornes de l'intervalle de confiance de l'hypothèse nulle.

L'hypothèse nulle est rejetée si la valeur observée de $\tilde{H}_\beta$ n'est pas dans cet intervalle, en général au-delà de la borne supérieure. Il peut arriver que les deux communautés soient plus semblables que sous l'hypothèse nulle, c'est-à-dire que les fréquences varient moins que dans le tirage d'une loi multinomiale, si deux communautés ont été créées artificiellement avec le même nombre d'individus de chaque espèce par exemple.

Lorsque les données sont issues de communautés réelles, le sens même de ce type de test est remis en question[102] : les communautés réelles ne pouvant pas être exactement identiques, il suffit d'augmenter la taille de l'échantillonnage pour prouver leur différence.

[102] Jones et Matloff (1986). « Statistical Hypothesis Testing in Biology : A Contradiction in Terms », cf. note 100, p. 151.

Intervalle de confiance

L'intervalle de confiance de l'estimateur de H_β[103] peut être calculé en simulant les communautés par des tirages dans des lois multinomiales suivant leurs fréquences : $\mathcal{M}\left(n_{+i}, n_s i/n_{+i}\right)$.

Si l'intervalle de confiance ne contient pas 0, l'égalité des distributions est rejetée.

Calcul sous R : la fonction `DivEst` produit un graphique avec l'intervalle de confiance (Figure 5.5 [104]).

[103] Marcon et al. (2014a). « Generalization of the Partitioning of Shannon Diversity », cf. note 38, p. 64.

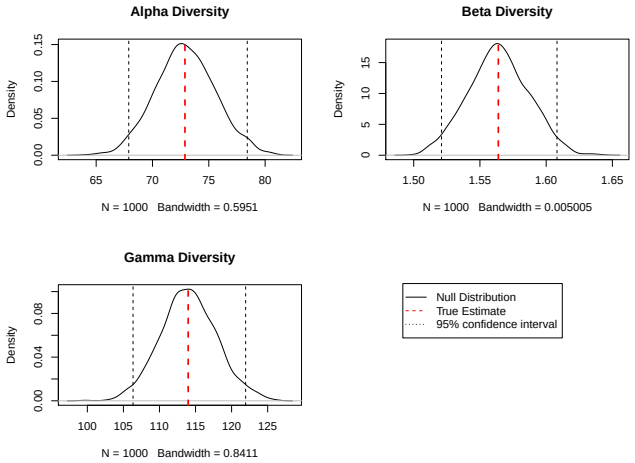


Figure 5.5 – Estimation de la diversité de la méta-communauté `Paracou618`.

[104] Code R :

```
de <- DivEst(q = 1, Paracou618.MC, Simulations = 1000)
plot(de)
```

Par défaut, `DivEst` corrige les biais. Les résultats sont présentés sous forme de diversité et non d'entropie, la valeur attendue en cas d'égalité des communautés est 1.

Correction des biais

Les simulations nécessaires aux tests créent un biais d'estimation : les espèces les plus rares dans les communautés sont souvent éliminées par les tirages. La correction du biais des tirages recentre leur distribution autour des valeurs originales des communautés non débiaisées. Il n'existe pas de correction analytique pour corriger successivement le biais dû aux simulations (dû à la perte des espèces rares des communautés réelles) puis celui dû à l'échantillonnage des communautés elles-mêmes (dû à la non-observation des espèces rares de la communauté). Marcon *et al.*,[105] comme Chao et Jost,[106] effectuent la deuxième correction par un recentrage de la distribution de H_β simulée autour de sa valeur observée débiaisée, ce qui permet d'obtenir l'intervalle de confiance de l'estimateur de H_β.

La distribution de l'estimateur sous l'hypothèse nulle ne peut pas être débiaisée complètement : elle est systématiquement surestimée.

[105] Marcon et al. (2012). « The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity », cf. note 13, p. 60.

[106] Chao et Jost (2015). « Estimating diversity and entropy profiles via discovery rates of new species », cf. note 66, p. 68.

5.7.2 Analyse de la variabilité

On se place maintenant dans un cadre un peu différent : les communautés sont elles-mêmes obtenues par regroupement de placettes d'échantillonnage. Pour tester la significativité de la diversité β entre communautés, l'approche pourrait être celle de l'analyse de variance : l'entropie totale de la méta-communauté est la somme des entropies des sites, de l'entropie β entre sites des communautés et de l'entropie β entre communautés. L'entropie β inter-communautés pourrait être comparée à l'entropie β intra-communautés en calculant une statistique F. Cette approche souffre de la dépendance de l'entropie β à l'entropie α : si une communauté est plus diverse que l'autre, son entropie β sera forcément plus faible.

Le test correct porte sur la diversité plutôt que sur l'entropie.[107] La statistique de test est le ratio de la diversité β inter-communautés sur la la diversité β intra-communautés. Sa valeur observée est comparée à la distribution de sa valeur simulée en redistribuant les placettes aléatoirement dans les communautés.

[107] C. Richard-Hansen et al. (2015). « Landscape patterns influence communities of medium- to large-bodied vertebrate in undisturbed terra firme forests of French Guiana ». In : *Journal of Tropical Ecology* 31.5, p. 423–436.

Chapitre 6

Diversité de paires

La distance entre les paires d'espèces est à la base de la définition de la diversité fonctionnelle ou phylogénétique, mais les nombres effectifs utilisés pour mesurer la diversité sont des nombres d'espèces. Le nombre effectif de paires d'espèces équiprobables de distance égale à une valeur de référence est une extension assez naturelle des nombres de Hill introduite récemment dans la littérature. La diversité β est en revanche classiquement abordée sous l'angle de la distance entre les paires de communautés.

6.1 Variations sur l'entropie quadratique

Izsák et Pavoine[1] proposent deux variantes de l'indice de Rao dans lesquelles la distance entre espèces dépend de leurs probabilités :

$$L_1 = \sum_s \sum_t p_s p_t (p_s - p_t)^2 \tag{6.1}$$

$$L_2 = \sum_s \sum_t p_s p_t (\ln p_s - \ln p_t)^2 \tag{6.2}$$

Ces indices sont proposés pour leurs propriétés mathématiques sans support écologique bien établi.

Δ est la matrice de dissimilarité dont les éléments sont $d_{s,t}$, la dissimilarité entre l'espèce s et l'espèce t. Guiasu et Guiasu[2] proposent l'indice GS_D pour ses propriétés mathématiques et fournissent sa décomposition :

$$GS_D = \sum_s \sum_t d_{s,t} p_s p_t \left(1 - p_s p_t\right) \tag{6.3}$$

Chiu et Chao[3] proposent de pondérer l'entropie quadratique par la distance entre les paires d'espèces et obtiennent un nombre de Hill permettant de mesurer la diversité fonctionnelle à partir d'une matrice de dissimilarité :

[1] J. Izsák et S. Pavoine (2011). « New concentration measures as kinds of the quadratic entropy ». In : *Ecological Indicators* 11.2, p. 540–544.

[2] Guiasu et Guiasu (2011). « The weighted quadratic index of biodiversity for pairs of species : a generalization of Rao's index », cf. note 72, p. 140 ; R. C. Guiasu et S. Guiasu (2012). « The Weighted Gini-Simpson Index : Revitalizing an Old Index of Biodiversity ». In : *International Journal of Ecology* 2012.478728, p. 1–10.

[3] Chiu et Chao (2014). « Distance-based functional diversity measures and their decomposition : a framework based on hill numbers. », cf. note 111, p. 109.

$$ {}^{q}D\left(Q\right) = \left[\sum_{s}\sum_{t}\frac{d_{s,t}}{Q}\left(p_s p_t\right)^{q}\right]^{\frac{1}{2(1-q)}} \tag{6.4} $$

$$ {}^{1}D\left(Q\right) = e^{\frac{1}{2}\left[\sum_{s}\sum_{t} p_s p_t \frac{d_{s,t}}{Q}\ln p_s p_t\right]} \tag{6.5} $$

La notation ${}^{q}D(Q)$ fait référence à l'entropie quadratique $Q = {}^{2}\bar{H}(T)$. Une approche complémentaire est développée ci-dessous. Chiu et Chao notent que $GS_D = Q - {}^{q}D(Q)$.

6.2 Diversité fonctionnelle

La définition de Chiu et Chao revient à calculer l'entropie des paires d'individus à partir de la fonction d'information $I(p_s p_t) = \ln_q[{}^{(d_{s,t}/Q)^{1/1-q}}/_{p_s p_t}]$ pour $q \neq 1$ et $I(p_s p_t) = \ln[{}^{1}/_{(p_s p_t)^{d_{s,t}/Q}}]$ pour $q = 1$. Le nombre effectif de paires est $[{}^{q}D(Q)]^2$. Le nombre effectif d'espèces est donc ${}^{q}D(Q)$.

Le nombre effectif de paires est le nombres de paires équifréquentes dont la distance entre les deux espèces est Q. Le problème de cette approche est que les paires constituées de la même espèce doivent aussi avoir une distance $d_{s,s}$ égale à Q. Chiu et Chao considèrent ce point comme la possibilité de prendre en compte la variabilité intraspécifique, mais elle doit être identique à la variabilité interspécifique, ce qui n'est pas très convaincant. En absence de variabilité intraspécifique, Chiu et Chao redéfinissent la distance de référence : $d_{s,s} = 0$ et $d_{s,t} = ({}^{D}/_{D-1})Q$. Il n'est donc pas possible de comparer le nombre effectif d'espèces de deux communautés différentes puisque la définition même du nombre effectif dépend de la diversité, ce qui invalide cette définition de la diversité fonctionnelle.

6.3 Diversité β

6.3.1 Données de présence-absence

Indices de similarité

Les indices les plus simples et probablement les plus connus pour calculer la diversité β sont l'indice de Jaccard[4] et de Sørensen.[5] Ils sont définis pour deux communautés et comptent la proportion d'espèces qu'elles partagent. Plus elle est faible, moins la composition spécifique des communautés est semblable, plus la diversité β est considérée comme grande.

Précisément, notons $s_{1,1}$ les espèces observées dans les deux communautés, $s_{1,0}$ celles observées dans la première communauté mais pas la deuxième (et de même $s_{0,1}$). L'indice de Jaccard est le rapport entre le nombre d'espèces communes et le nombre d'espèces total :

[4] P. Jaccard (1901). « Étude comparative de la distribution florale dans une portion des Alpes et du Jura ». In : *Bulletin de la Société Vaudoise des Sciences Naturelles* 37, p. 547–579.

[5] L. R. Dice (1945). « Measures of the amount of ecologic association between species ». In : *Ecology* 26.3, p. 297–302 ; T. Sørensen (1948). *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*. T. 5. Biologiske Skrifter 4. København : I kommission hos E. Munksgaard, p. 1–34.

$$S_3 = \frac{s_{1,1}}{s_{1,1} + s_{1,0} + s_{0,1}} \tag{6.6}$$

L'indice de Sørensen vaut deux fois le nombre d'espèces communes divisé par la somme du nombre d'espèces des deux communautés :

$$S_7 = \frac{2s_{1,1}}{2s_{1,1} + s_{1,0} + s_{0,1}} \tag{6.7}$$

Les notations S_n correspondent à celles de l'article de référence de Gower et Legendre[6] qui fait la revue de ces indices de similarité (13 autres sont traités) et en étudie les propriétés. Certains indices prennent en compte les espèces absentes des deux communautés mais qui auraient pu être observées $s_{0,0}$.

[6] J. C. Gower et P. Legendre (1986). « Metric and Euclidean properties of dissimilarity coefficients ». In : *Journal of classification* 48, p. 5–48.

Deux autres indices sont mentionnés ici, celui de Sokal et Sneath[7] et celui d'Ochiai[8] :

$$S_5 = \frac{s_{1,1}}{s_{1,1} + 2\left(s_{1,0} + s_{0,1}\right)} \tag{6.8}$$

$$S_{12} = \frac{s_{1,1}}{\sqrt{\left(s_{1,1} + s_{1,0}\right)\left(s_{1,1} + s_{0,1}\right)}} \tag{6.9}$$

[7] R. R. Sokal et P. H. A. Sneath (1963). *Principles of numerical taxonomy*. San Francisco : W. H. Freeman & Co.

[8] A. Ochiai (1957). « Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions ». In : *Bulletin of the Japanese Society of Scientific Fisheries* 22.9, p. 526–530.

Une propriété importante est la possibilité de comparer plusieurs communautés en calculant la matrice de leurs dissimilarités deux à deux et en en faisant une représentation graphique. Pour cela, la similarité est transformée en dissimilarité $D_n = 1 - S_n$ (les similarités sont construites pour être comprises entre 0 et 1). D_n n'est pas toujours une distance euclidienne, mais $\sqrt{D_n}$ l'est pour la majorité des indices. On utilisera donc généralement la matrice de distances euclidiennes Δ dont chaque élément $d_{i,j}$ est la racine de la dissimilarité $\sqrt{D_n}$ calculée entre les communautés i et j.

Exemple : l'indice de Jaccard peut être calculé pour tous les carrés de BCI comparés deux à deux. La fonction `dist.binary` du package *ade4* calcule $\sqrt{D_n}$ pour 10 indices de similarité étudiés par Gower et Legendre (voir l'aide de la fonction). Le résultat est une demi-matrice de distances de classe `dist`.

```
DistJaccard <- dist.binary(BCI, method = 1)
```

La distribution des valeurs de l'indice de Jaccard $S_3 = 1 - D_3^2$ est en figure 6.1 [9]. Le mode est proche de 0,5 : la moitié des espèces sont communes entre deux carrés.

Pour afficher l'indice de Jaccard entre les carrés 1 et 2, il faut transformer la demi-matrice des distances en matrice de similarité :

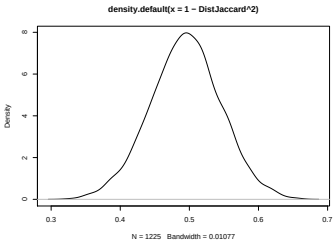


Figure 6.1 – Distribution des indices de Jaccard calculés pour les 50 carrés de BCI deux à deux.

[9] Code R pour la figure :

```
plot(density(1 - DistJaccard^2))
```

```
Jaccard <- 1 - (as.matrix(DistJaccard))^2
Jaccard[1, 2]

## [1] 0.5663717
```

Biais d'estimation

Tous les indices de similarité vus plus haut sont sensibles au biais d'estimation. Chao *et al.*[10] fournissent un estimateur de la probabilité qu'un individu de la première communauté appartienne à une espèce commune aux deux communautés :

[10] A. Chao et al. (2004). « A new statistical approach for assessing similarity of species composition with incidence and abundance data ». In : *Ecology Letters* 8.2, p. 148–159.

$$\hat{U} = \sum_{s}^{s_{1,1}} \frac{n_{s,1}}{n} + \frac{n_{+2}-1}{n_{+2}} \frac{s_1^{(2)}}{2s_2^{(2)}} \sum_{s}^{s_{1,1}} \frac{n_s}{n} \mathbf{1}\left(n_{s,2}=1\right) \tag{6.10}$$

Les sommes s'entendent pour toutes les espèces communes, indicées de 1 à $s_{1,1}$. Le premier terme de la somme est donc la probabilité qu'un individu de la première communauté appartienne à une espèce commune aux deux communautés. n_{+2} est le nombre d'individus de la deuxième communauté. $s_1^{(2)}$ et $s_2^{(2)}$ sont le nombre d'espèces communes qui sont des singletons et doubletons dans la deuxième communauté. L'indicatrice $\mathbf{1}(n_{s,2}=1)$ vaut 1 si l'espèce s est un singleton dans la deuxième communauté : la dernière somme est donc la probabilité qu'un individu de la première communauté appartienne à une espèce commune qui soit un singleton dans la deuxième communauté. S'il n'y a pas de doubletons dans la deuxième communauté, leur nombre est remplacé par 1. Enfin, il est possible que l'estimateur soit supérieur à 1, il est alors fixé à 1.

L'estimateur de la probabilité qu'un individu de la deuxième communauté appartienne à une espèce commune est noté $\hat{V}$ et est calculé de façon symétrique à $\hat{U}$. Le nombre d'espèces communes $s_{1,1}$ est estimé par $\hat{U}\hat{V}$, $s_{1,0}$ par $(1-\hat{U})\hat{V}$ et $s_{1,0}$ par $\hat{U}(1-\hat{V})$. L'estimateur de l'indice de Jaccard est donc :

$$\hat{S}_3 = \frac{\hat{U}\hat{V}}{\hat{U}+\hat{V}-\hat{U}\hat{V}} \tag{6.11}$$

L'estimateur de l'indice de Sørensen est :

$$\hat{S}_7 = \frac{2\hat{U}\hat{V}}{\hat{U}+\hat{V}} \tag{6.12}$$

La matrice de dissimilarité corrigée est assez différente de l'originale. L'indice de Jaccard entre les deux premiers carrés devient :

```
# Calcul de U et V, N2Col est une matrice ou un dataframe de
# deux colonnes contenant les effectifs de chaque parcelle
ChaoUV <- function(N2Col) {
```

```
    NCommunity <- colSums(N2Col)
    Common <- (N2Col[, 1] > 0) & (N2Col[, 2] > 0)
    Single <- N2Col == 1
    NSingleCommon <- colSums(Common & Single)
    Double <- N2Col == 2
    NDoubleCommon <- colSums(Common & Double)
    NsCommon <- N2Col[Common, ]
    PsCommon <- t(t(NsCommon)/colSums(N2Col))
    U <- sum(PsCommon[, 1]) + (NCommunity[2] - 1)/NCommunity[2] *
        NSingleCommon[2]/2/max(NDoubleCommon[2], 1) * sum(PsCommon[,
        1][(NsCommon == 1)[, 2]])
    V <- sum(PsCommon[, 2]) + (NCommunity[1] - 1)/NCommunity[1] *
        NSingleCommon[1]/2/max(NDoubleCommon[1], 1) * sum(PsCommon[,
        2][(NsCommon == 1)[, 1]])
    return(list(U = U, V = V))
}
# Indice de Jaccard corrigé
bcJaccard <- function(N2Col) {
    ChaoEst <- ChaoUV(N2Col)
    return(min(ChaoEst$U * ChaoEst$V/(ChaoEst$U + ChaoEst$V -
        ChaoEst$U * ChaoEst$V), 1))
}
# Calcul de l'indice entre les carrés i et j de BCI, i<j.
BCIjaccard <- function(i, j) {
    if (i >= j)
        return(NA) else return(bcJaccard(t(BCI[c(i, j), ])))
}
# Calcul de la matrice de dissimilarité entre les 50 carrés
ChaoJaccard <- sapply(1:50, function(i) {
    sapply(1:50, function(j) BCIjaccard(i, j))
})
# Indice de Jaccard, carrés 1 et 2, corrigé
BCIjaccard(1, 2)

## [1] 0.9391222
```

La distribution des valeurs de l'indice corrigé est en figure 6.2 [11].

La majorité des paires de carrés a une similarité supérieure à 0,8 après correction.

Représentation graphique

La matrice de distances peut être représentée par une analyse en coordonnées principales (PCoA)[12] :

```
DistChaoJaccard <- as.dist(sqrt(1 - ChaoJaccard))
PcoChaoJaccard <- dudi.pco(DistChaoJaccard, scannf = FALSE)
```

La représentation de l'analyse est en figure 6.3 [13]. Les 50 carrés de BCI sont représentés sur le plan des deux premiers facteurs. Les valeurs propres sont représentées dans le cartouche en haut à gauche de la figure. Les distances entre les points sont les valeurs de $\sqrt{D_3}$, projetées sur le plan. La matrice de distances n'est pas

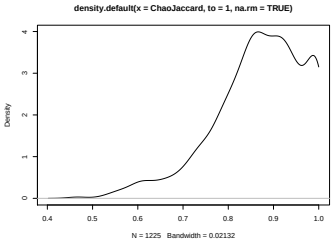


FIGURE 6.2 – Distribution des indices de Jaccard corrigés du biais d'etimation pour les 50 carrés de BCI deux à deux.

[11] Code R pour la figure :

```
plot(density(ChaoJaccard, na.rm = TRUE, to = 1))
```

[12] GOWER (1966). « Some distance properties of latent root and vector methods used in multivariate analysis », cf. note 66, p. 139.

[13] Code R pour la figure :

```
scatter(PcoChaoJaccard, posieig = "bottomright")
```

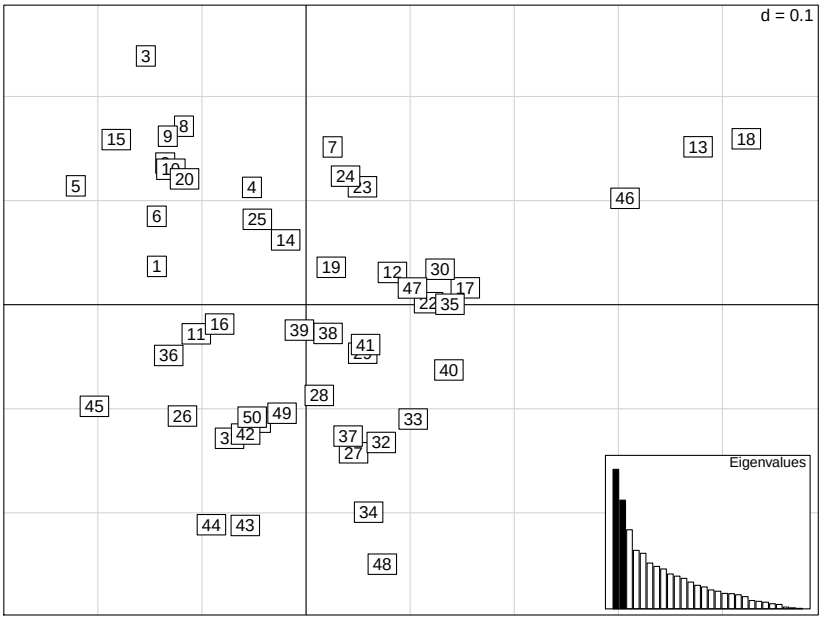


FIGURE 6.3 – Représentation des 50 carrés de BCI dans une analyse en coordonnées principale fondée sur les distances de Jaccard.

euclidienne à cause des approximations de l'estimation mais les conséquences sont faibles sur les premiers axes.

Emboîtement et substitution

L'indice de Sørensen et les indices similaires prennent en compte deux types de dissimilarité : l'emboîtement des composition spécifiques (*nestedness*) et la substitution d'espèces (*turnover*). On parle d'emboîtement quand une communauté diffère d'une autre parce que sa composition est un sous-ensemble des espèces de l'autre. La substitution correspond au remplacement de certaines espèces d'une communauté par de nouvelles espèces. Baselga[14] montre que la diversité β de longicornes en Europe du nord est similaire à celle observée en Europe du sud, mais pour des raisons totalement différentes. En Europe du nord, les communautés perdent des espèces avec la latitude : leur dissimilarité est due principalement à l'emboîtement. En Europe du sud, les communautés sont composées d'espèces différentes : la diversité β est due à la substitution des espèces. Baselga décompose l'indice de dissimilarité de Sørensen, $1 - S_7$, en un terme mesurant la substitution et un mesurant l'emboîtement.

[14] A. BASELGA (2010b). « Partitioning the turnover and nestedness components of beta diversity ». In : *Global Ecology and Biogeography* 19.1, p. 134–143.

Le premier composant est l'indice de dissimilarité de Simpson :[15]

$$D_{sim} = \frac{\min(s_{1,0}, s_{0,1})}{s_{1,1} + \min(s_{1,0}, s_{0,1})} \tag{6.13}$$

[15] G. G. SIMPSON (1943). « Mammals and the Nature of Continents ». In : *American Journal of Science* 241, p. 1–31.

Il vaut 0 quand les communautés sont emboîtées. Son complément D_{nes} prend en compte la partie de la dissimilarité due à l'emboîtement :

$$D_{nes} = \frac{s_{1,0} + s_{0,1}}{2s_{1,1} + s_{1,0} + s_{0,1}} - \frac{\min(s_{1,0}, s_{0,1})}{s_{1,1} + \min(s_{1,0}, s_{0,1})} \tag{6.14}$$

Ces mesures de dissimilarité sont généralisées au-delà de deux communautés : les formules sont disponibles dans l'article de Baselga.

Les mesures de diversité β fondées sur l'entropie HCDT (chapitre 5) ne permettent pas de distinguer directement emboîtement et substitution. La variabilité des contributions ${}^{q}_{i}H_{\beta}$ des communautés à l'entropie β augmente avec l'emboîtement, mais cette voie n'a pas été explorée.

6.3.2 Données d'abondance

La distance la plus évidente entre communautés quand les effectifs des espèces sont connus est simplement la distance euclidienne (Figure 6.4), éventuellement normalisée par le nombre d'espèces (Dissimilarité D_1 de Gower et Legendre[16]). Son problème principal est qu'elle est très sensible aux ordres de grandeurs relatifs des variables.

[16] Gower et Legendre (1986). « Metric and Euclidean properties of dissimilarity coefficients », cf. note 6, p. 157.

La dissimilarité de Steinhaus,[17] plus connue sous le nom de Bray-Curtis[18] (voir l'analyse historique de Lengendre et Legendre[19]) est très utilisée (par défaut dans plusieurs analyses du package *vegan* par exemple) :

$$D_8 = \frac{\sum_s |n_{s,1} - n_{s,2}|}{\sum_s (n_{s,1} + n_{s,2})} \tag{6.15}$$

[17] J. Motyka (1947). *O celach i metodach badan geobotanicznych. Sur les buts et les méthodes des recherches géobotaniques*. Lublin, Poland : Nakladem universytftu Marii Curie-Sklodowskiej.

[18] E. P. Odum (1950). « Bird Populations of the Highlands (North Carolina) Plateau in Relation to Plant Succession and Avian Invasion ». In : *Ecology* 31.4, p. 587–605 ; J. R. Bray et J. T. Curtis (1957). « An Ordination of the Upland Forest Communities of Southern Wisconsin ». In : *Ecological Monographs* 27.4, p. 325–349.

[19] Legendre et Legendre (2012). *Numerical Ecology*, cf. note 27, p. 86, note de bas de page, eq. 7.58.

$n_{s,1}$ et $n_{s,2}$ sont le nombre d'individus de l'espèce s dans la première et la deuxième communauté. La racine carrée de la dissimilarité de Bray-Curtis est euclidienne.

Gower[20] propose une extension de la dissimilarité de Bray-Curtis :

$$D_G = \frac{\sum_s w_s |n_{s,1} - n_{s,2}|}{R_s \sum_s w_s} \tag{6.16}$$

[20] Gower (1971). « A General Coefficient of Similarity and Some of Its Properties », cf. note 21, p. 85.

R_s est l'étendue ou l'écart-type de l'abondance de l'espèce s, qui permet de ne pas surpondérer l'effet des espèces très abondantes. w_s est le poids attribué à l'espèce s, qui peut être utilisé pour éliminer les doubles absences ($w_s = 0$ si l'espèce est absente des deux communautés, 1 sinon).

Anderson *et al.*[21] proposent une mesure de dissimilarité (de Gower modifiée) donnant le même poids à une différence d'un ordre de grandeur entre les abondances qu'à la différence entre présence et absence, qui permet donc de se passer de la normalisation par R_s. Les effectifs sont transformés de la façon suivante : $x_s = \log n_s + 1$ si $n_s \neq 0$, $x_s = 1$ sinon (le logarithme est décimal). La dissimilarité est :

[21] M. J. Anderson et al. (2006). « Multivariate dispersion as a measure of beta diversity. » In : *Ecology letters* 9.6, p. 683–93.

$$D_{MG} = \frac{\sum_s w_s |x_{s,1} - x_{s,2}|}{\sum_s w_s} \tag{6.17}$$

Elle n'est pas euclidienne.

6.3.3 Données d'abondance et similarité des espèces

Dissimilarité minimale moyenne

Ricotta et Bacaro[22] définissent une mesure de dissimilarité paramétrique entre deux communautés. Une matrice de dissimilarité est définie entre toutes les paires d'espèces. $\min d_{s,2}$ désigne la dissimilarité entre l'espèce s présente dans la première communauté et l'espèce la plus proche d'elle dans la deuxième communauté. $\min d_{t,1}$ est la même notation pour l'espèce t de la deuxième communauté et l'espèces la plus proche dans la communauté 1. Ces valeurs sont nulles pour toutes les espèces communes.

$\pi_{s(n)}$ désigne la contribution de l'espèce s à l'indice de Hurlbert, déjà vue dans le cadre de l'originalité taxonomique (page 92). La contribution des espèces rares augmente avec le paramètre n (Figure 4.5, page 93). Pour chaque communauté, la moyenne des dissimilarités minimales entre chaque espèce et l'autre communauté est calculée, pondérée par la contribution de l'espèce à l'indice de Hurlbert.

La dissimilarité pondérée (W pour *weighted*) entre les deux communautés (C pour *community*) en est la moyenne entre les deux communautés :

$$d_{CW(n)} = \frac{1}{2}\left(\sum_s \pi_{s(n)} \min d_{s,2} + \sum_t \pi_{t(n)} \min d_{t,1}\right) \tag{6.18}$$

Sa version non pondérée (U pour *unweighted*) par les abondances est :[23]

$$d_{CU} = \frac{\sum_s \min d_{s,2} + \sum_t \min d_{t,1}}{2s_{1,1} + s_{1,0} + s_{0,1}} \tag{6.19}$$

Similarité moyenne entre individus

Pavoine et Ricotta[24] définissent une famille de similarités entre communautés construite à partir de la similarité moyenne entre les paires d'individus de chaque communauté et les paires intercommunautaires. Une matrice de similarité $\boldsymbol{Z}$ est définie entre les paires d'espèces : $z_{s,t}$ est la similarité entre l'espèce s et l'espèce t, comprise entre 0 et 1 de la même façon que pour le calcul de la diversité ${}^{q}D^{\boldsymbol{Z}}$ (page 105). La matrice $\boldsymbol{Z}$ doit de plus être semi-définie positive, ce qui implique que $\sqrt{1-\boldsymbol{Z}}$ est euclidienne : de cette façon, les espèces peuvent être représentées graphiquement par une PCoA. Les fréquences des espèces dans les deux communautés sont $\mathcal{P}$ et $\mathcal{Q}$. La similarité moyenne dans chaque communauté est $\sum_{s,t} p_s p_t z_{s,t}$ et $\sum_{s,t} q_s q_t z_{s,t}$. La similarité moyenne des paires d'individus de communautés différentes est $\sum_{s,t} p_s q_t z_{s,t}$.

La dissimilarité entre communautés de Rao (page 138) peut être normalisée (en la divisant par sa valeur maximale pour qu'elle

[22] C Ricotta et G Bacaro (2010). « On plot-to-plot dissimilarity measures based on species functional traits ». In : *Community Ecology* 11.1, p. 113–119.

[23] C. Ricotta et S. Burrascano (2008). « Beta diversity for functional ecology ». In : *Preslia* 80.1, p. 61–72.

[24] Pavoine et Ricotta (2014). « Functional and phylogenetic similarity among communities », cf. note 120, p. 114.

soit comprise entre 0 et 1) puis transformée en similarité (en prenant son complément à 1) pour obtenir :

$$S_{Sorensen} = \frac{\sum_{s,t} p_s q_t z_{s,t}}{\frac{1}{2}\sum_{s,t} p_s p_t z_{s,t} + \frac{1}{2}\sum_{s,t} q_s q_t z_{s,t}} \tag{6.20}$$

Les probabilités peuvent être remplacées par n'importe quel vecteur de valeurs positives, par exemple les abondances ou des valeurs 1 et 0 pour des données de présence-absence. Dans ce dernier cas, et si $\boldsymbol{Z}$ est la matrice identité $\boldsymbol{I}$ (la similarité de deux espèces différentes est nulle et $z_{s,s} = 1$), $S_{Sorensen}$ se réduit à l'indice de Sørensen vu plus haut.

La généralisation de l'indice de Jaccard aux données d'abondances est :

$$S_{Jaccard} = \frac{\sum_{s,t} p_s q_t z_{s,t}}{\sum_{s,t} p_s p_t z_{s,t} + \sum_{s,t} q_s q_t z_{s,t} - \sum_{s,t} p_s q_t z_{s,t}} \tag{6.21}$$

Il se réduit à l'indice de Jaccard original pour les données de présence-absence et $\boldsymbol{Z} = \boldsymbol{I}$.

L'indice de Sokal et Sneath se généralise par :

$$S_{Sokal} = \frac{\sum_{s,t} p_s q_t z_{s,t}}{2\sum_{s,t} p_s p_t z_{s,t} + 2\sum_{s,t} q_s q_t z_{s,t} - 3\sum_{s,t} p_s q_t z_{s,t}} \tag{6.22}$$

Celui d'Ochiai par :

$$S_{Ochiai} = \frac{\sum_{s,t} p_s q_t z_{s,t}}{\sqrt{\sum_{s,t} p_s p_t z_{s,t}}\sqrt{\sum_{s,t} q_s q_t z_{s,t}}} \tag{6.23}$$

Ces quatre similarités permettent une représentation des communautés par PCoA de la même façon que l'indice de Jaccard traité plus haut.

6.3.4 Distance entre paires de communautés et diversité β

Ter Braak[25] a le premier rapproché l'entropie de Simpson et l'Analyse en Composantes Principales.[26] En représentant les communautés dans l'espace des espèces, la norme du vecteur de la communauté i est $(\sum_s p_{s,i})^2$, c'est-à-dire 1 moins l'entropie de Simpson. La représentation graphique de l'ACP (non centrée, non réduite) montre les communautés les moins diverses (les plus éloignées de l'origine du repère). Ter Braak *définit* la diversité β comme la distance entre deux communautés (et la moyenne des distances entre paires de communautés si leur nombre est supérieur à 2). Cette « définition » est valide puisque cette distance

[25] Braak (1983). « Principal components biplots and alpha and beta diversity », cf. note 7, p. 123.

[26] K. Pearson (1901). « On lines and planes of closest fit to systems of points in space ». In : *Philosophical Magazine* 2, p. 559–572.

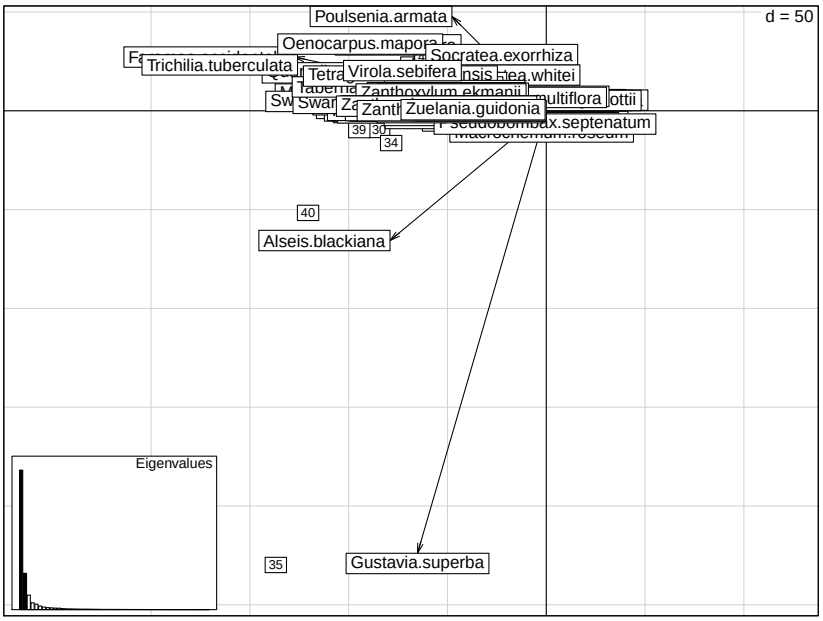


FIGURE 6.4 – Représentation des deux premiers axes factoriels de l'ACP des 50 carrés de BCI. Les données sont les nombres d'individus par espèce, ni centrées ni réduites pour l'ACP. Les carrés sont représentés avec les espèces. Les carrés 35 et 40 ont une diversité faible et sont associés aux espèces *Gustavia superba* et *Alseis blackiana*.

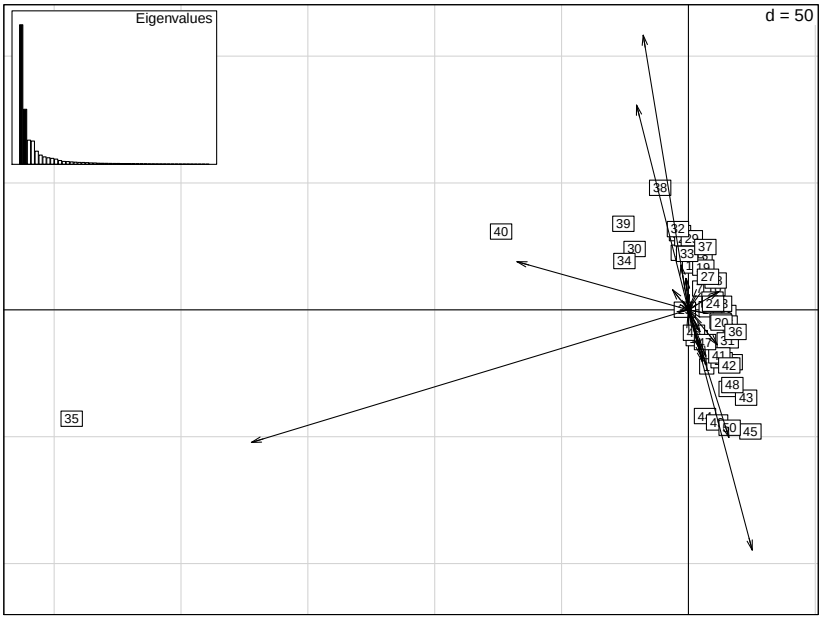


FIGURE 6.5 – ACP centrée des 50 carrés de BCI pour visualiser la diversité β.

est la composante inter de la décomposition de l'entropie β de Simpson (page 136).

L'ACP non centrée des 50 carrés de BCI est en figure 6.4 [27]. L'ACP centrée (mais non réduite, figure 6.5) [28] permet de mieux représenter la diversité β : la distance de chaque carré au centre du repère est sa contribution.

Legendre et Anderson[29] ont montré plus largement que si les communautés sont représentées dans l'espace vectoriel des espèces, la moyenne des carrés des distances euclidiennes entre communautés égale la somme des carrés des écarts à la moyenne des distributions des espèces :

$$\frac{1}{I}\sum_{i=1}^{I-1}\sum_{j=i+1}^{I} d_{i,j}^2 = \sum_{s=1}^{S}\sum_{i=1}^{I}(y_{s,i}-y_s)^2 \tag{6.24}$$

$y_{s,i}$ peut être le nombre d'individus $n_{s,i}$ ou la probabilité $p_{s,i}$. Les communautés ont ici le même poids, y_s est la moyenne des valeurs de l'espèce s dans chaque communauté. La somme des carrés des écarts est une mesure de diversité β. Si $y_{s,i}$ est la probabilité $p_{s,i}$, il s'agit de l'entropie β de Simpson.

Ce résultat peut être étendu à toute matrice de distances euclidiennes entre communautés[30] : l'argument est qu'il existe une matrice de $y_{s,i}$, transformation des $p_{s,i}$, telle que la distance entre les communautés est cette matrice de distances. Pour un certain nombre de distances, la transformation de $p_{s,i}$ en $y_{s,i}$ est connue.[31]

La moyenne des dissimilarités D_n ou des compléments à 1 des similarités de Pavoine et Ricotta (qui sont les carrés des distances euclidiennes de la PCoA) est donc une mesure de diversité β.

6.4 Diversité de Kosman

Kosman[32] nomme *diversité génétique* la diversité des individus d'une population (non regroupés par espèces) pour lesquels une mesure de dissimilarité a été choisie. Les mesures de dissimilarité acceptables[33] sont par exemple les mêmes que celles utilisées pour la diversité β des communautés appliquées aux présences/absences de bandes de profils AFLP.

[27] Code R pour la figure :

```
pcaBCI <- dudi.pca(BCI, scannf = FALSE, nf = 2, center = FALSE,
    scale = FALSE)
scatter(pcaBCI, posieig = "bottomleft")
```

[28] Code R :

```
pcaBCI <- dudi.pca(BCI, scannf = FALSE, nf = 2, center = TRUE,
    scale = FALSE)
scatter(pcaBCI, clab.col = 0)
```

[29] Legendre et Anderson (1999). « Distance-based redundancy analysis : testing multispecies responses in multifactorial ecological experiments », cf. note 11, p. 126.

[30] Legendre et al. (2005). « Analyzing Beta Diversity : Partitioning the Spatial Variation of Community Composition Data », cf. note 11, p. 126.

[31] Legendre et Gallagher (2001). « Ecologically meaningful transformations for ordination of species data », cf. note 11, p. 126.

[32] E. Kosman (2013). « Measuring diversity : from individuals to populations ». In : *European Journal of Plant Pathology* 138.3, p. 467–486.

[33] E. Kosman et K. J. Leonard (2005). « Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. » In : *Molecular ecology* 14.2, p. 415–24.

La mesure de dissimilarité moyenne entre les individus, notée ADW_ρ pour *Average Dissimilarity Within*, où ρ représente la dissimilarité choisie, est simplement l'entropie quadratique de Rao. Si la dissimilarité n'est pas euclidienne, la concavité de ADW_ρ n'est pas garantie, ce qui motive Kosman à proposer une autre mesure,[34] KW_ρ, qui est la dissimilarité moyenne des n paires d'individus les plus dissimilaires (au lieu de toutes les $n(n-1)/2$ paires). Les n paires d'individus sont construites de façon que chaque individu apparaisse une fois en tant que premier et une fois en tant que second membre de la paire, selon un algorithme non trivial.[35] Tous les individus contribuent donc à la mesure de la diversité.

[34] E. Kosman (1996). « Difference and diversity of plant pathogen populations : a new approach for measuring ». In : *Phytopathology* 86.11, p. 1152–1155 ; E. Kosman et K. J. Leonard (2007). « Conceptual analysis of methods applied to assessment of diversity within and distance between populations with asexual or mixed mode of reproduction. » In : *The New phytologist* 174.3, p. 683–96.

[35] R. Bellman et al. (1970). *Algorithms, Graphs and Computers*. New York : Academic Press, Inc.

Pour un assemblage de I populations, la diversité de Kosman KW_ρ de l'assemblage des populations est toujours plus grande que la moyenne des diversités intra-populations KW_ρ pondérées par le poids des populations (c'est-à-dire la diversité α selon la définition de Routledge) parce que KW_ρ est concave. La différence est la diversité entre populations, interprétable comme une diversité β selon la définition additive de la décomposition.

Pour obtenir une diversité β indépendante des diversités α et γ, Kosman définit KA_d comme la moyenne des distances entre les I paires de populations assignées selon le même algorithme que précédemment. La distance entre populations d est de préférence la distance de Kosman entre populations, KB_ρ : la moyenne des n paires d'individus appartenant chacun à une population les plus dissimilaires (si la taille des populations n'est pas identique, un échantillon de la plus grande population de la taille de la plus petite est tiré aléatoirement, la valeur moyenne de la distance obtenue sur plusieurs tirages est calculée). La diversité γ est la somme de la moyenne des diversités intra-populations KW_ρ pondérées par le poids des populations et de la diversité inter-populations KA_d. Ce n'est pas la diversité KW_ρ de l'assemblage des populations, ce qui limite fortement l'intérêt de cette approche.

6.5 Tests de significativité

6.5.1 Test de non-nullité

Chase *et al.*[36] testent la valeur de diversité β observée contre sa valeur attendue sous l'hypothèse nulle d'une distribution aléatoire des individus dans les communautés.

[36] J. M. Chase et al. (2011). « Using null models to disentangle variation in community dissimilarity from variation in α-diversity ». In : *Ecosphere* 2.2, art. 24.

Leur motivation principale est d'éliminer la dépendance de la diversité β à la diversité α due à la mesure de diversité qu'ils utilisent : l'indice de Jaccard, appliqué aux données de présence-absence. Des communautés théoriques sont tirées selon le modèle nul (selon une distribution multinomiale correspondant aux probabilités des espèces dans la méta-communauté et la taille de chaque communauté dans le cas le plus courant). La distribution

de la diversité β sous l'hypothèse nulle est obtenue en répétant les simulations. La différence entre la diversité β réelle et la moyenne simulée est normalisée par la moyenne simulée pour obtenir une valeur comprise entre -1 et +1 (l'indice de Jaccard est compris entre 0 et 1), comparée à sa distribution sous l'hypothèse nulle pour évaluer sa significativité.

Tucker *et al.*[37] étendent cette méthode aux données d'abondance en utilisant la dissimilarité de Bray-Curtis et testent, à partir de communautés simulées, sa capacité à distinguer les mécanismes d'assemblage des communautés : niche contre modèle neutre, et modèles déterministes contre stochastiques (c'est-à-dire dont les paramètres varient aléatoirement). Ces deux oppositions peuvent être traitées indépendamment, pour éviter la confusion entre niche et déterminisme et entre neutralité et stochasticité. Ils concluent qu'une valeur de diversité β significativement différente de celle du modèle nul permet de détecter le mécanisme de niche (alors que le modèle neutre donne une valeur identique à celle de l'hypothèse nulle, 0 étant donné la normalisation). Chase[38] interprétait les valeurs significatives comme preuve de déterminisme (par opposition à la stochasticité). Tucker *et al.* montrent que ce n'est pas le cas : le niveau de stochasticité influe peu sur les résultats quand la diversité est mesurée à partir de données d'abondance. En revanche, la méthode n'est pas fiable quand elle est utilisée avec des données de présence-absence : la stochasticité ne permet alors plus de distinguer les mécanismes.

[37] C. M. Tucker et al. (2015). « Differentiating between niche and neutral assembly in metacommunities using null models of β-diversity ». In : *Oikos* in press.

[38] J. M. Chase et J. A. Myers (2011). « Disentangling the importance of ecological niches from stochastic processes across scales ». In : *Philosophical Transactions of the Royal Society B : Biological Sciences* 366.1576, p. 2351–2363.

6.5.2 Analyse de la variabilité

Anderson[39] fournit un test d'égalité de la diversité β entre groupes de communautés. Les communautés sont placées dans un espace euclidien, dans lequel les distances entre elles sont une mesure de dissimilarité présentée dans ce chapitre. Si les dissimilarités ne sont pas euclidiennes, une PCoA permet de les placer dans un espace euclidien dont certains axes sont imaginaires, sans que cette difficulté invalide la méthode.

[39] Anderson et al. (2006). « Multivariate dispersion as a measure of beta diversity. », cf. note 21, p. 161.

Dans chaque groupe, la distance entre chaque communauté et le centre de gravité du groupe est calculée. Une Anova est réalisée à partir de ces distances. La statistique F de l'Anova suit une loi de Fisher si la distribution des distances aux centres de gravité est normale. Si ce n'est pas le cas, le test peut être réalisé par permutation des distances entre les groupes. La statistique de test de l'Anova est le rapport de la dispersion des centres de gravité des groupes à celle des dispersion des communautés dans chaque groupe. Elle est comparée à sa distribution obtenue par permutation.

Une alternative consiste à utiliser la distance aux médianes spatiales des groupes. La médiane spatiale est le point qui minimise la somme des distances entre les points et lui (alors que le centre

de gravité minimise la somme des carrés de ces distances).

La fonction `betadisper` du package *vegan* implémente ce test.

Ricotta et Burrascano[40] utilisent comme point central de chaque groupe celui représentant l'assemblage de ses communautés. Ils définissent ainsi la diversité β comme la dissimilarité moyenne entre chaque communauté et la méta-communauté.

Bacaro *et al.*[41] préfèrent tester la distribution des dissimilarités entre paires des communautés plutôt que la dissimilarité entre chaque communauté et un point de référence. Une statistique F est calculée à partir de ces dissimilarités et comparée à sa distribution obtenue par permutation. La méthode de permutation originale est inappropriée et donc corrigée :[42] les dissimilarités doivent être redistribuées aléatoirement entre les groupes.

[40] Ricotta et Burrascano (2009). « Testing for differences in beta diversity with asymmetric dissimilarities », cf. note 75, p. 140.

[41] G. Bacaro et al. (2012). « Testing for Differences in Beta Diversity From Plot-To-Plot Dissimilarities ». In : *Ecological Research* 27.2, p. 285–292.

[42] G. Bacaro et al. (2013). « Beta diversity reconsidered ». In : *Ecological Research* 28.3, p. 537–540.

Chapitre 7

Diversité jointe, diversité structurelle

La notion de diversité structurelle a été définie par les forestiers.[1] En plus des espèces, d'autres critères (par exemple les classes de diamètres) sont pris en compte pour définir la structure forestière. Une revue des approches utilisées pour décrire cette structure, et plus particulièrement sa complexité, est fournie par Bacaro *et al.* Certaines se sont attachées à définir une mesure de diversité structurelle : notamment, Das et Nautiyal[2] définissent un indice qui prend en compte la variabilité des diamètres.

Bacaro *et al.* suggèrent d'utiliser l'entropie de Shannon pour mesurer la diversité structurelle. Cette idée est explorée ici pour définir une mesure de diversité bidimensionnelle (croisant par exemple les espèces et les classes de diamètres) et étendue à la diversité HCDT.

Gregorius[3] argumente en faveur de la prise en compte de ce qu'il définit comme la « diversité jointe » dans le cadre classique où les classes sont des communautés pour des raisons fondamentales de définition de la diversité. Son approche consiste à traiter de façon symétrique l'appartenance à une classe ou à une espèce : contrairement aux chapitres précédents où l'échantillonnage des individus avait lieu séparément dans chaque communauté, l'échantillonnage a lieu ici globalement dans la méta-communauté et l'appartenance d'un individu à une communauté apporte autant d'information que son appartenance à une espèce. Cette approche convient bien à des questions de choix de localisation traités en économie (voir par exemple la définition de la localisation globale,[4] qui traite de façon symétrique le secteur d'activité et la région d'implantation des firmes). L'entropie des lignes est alors la diversité de localisation des firmes ; son complément à la valeur maximale est une mesure de concentration spatiale (l'indice de Theil pour $q = 1$). En écologie, cette entropie permet par exemple de caractériser le niveau de spécialisation d'une espèce si les communautés sont de même taille (plus précisément de même

[1] G. Bacaro et al. (2014). « Using Shannon's recursivity to summarize forest structural diversity ». In : *Forests Trees and Livelihoods* 23.3, p. 211–216.

[2] J. K. Das et J. Nautiyal (2004). « Forest variability index : a vector quantifying forest stand diversity and forest compactness ». In : *Forest Policy and Economics* 6.3-4, p. 271–288.

[3] Gregorius (2010). « Linking Diversity and Differentiation », cf. note 44, p. 64 ; Gregorius (2014). « Partitioning of diversity : the "within communities" component », cf. note 27, p. 61.

[4] E. Cutrini (2010). « Specialization and Concentration from a Twofold Geographical Perspective : Evidence from Europe ». In : *Regional Studies* 44.3, p. 315–336.

TABLE 7.1 – Notations des effectifs, tableau espèces-classes.

	Classe i	...	Total : communauté
Espèce s	$n_{s,i}$: nombre d'individus de l'espèce s dans la classe i. $\hat{p}^*_{s,i} = n_{s,i}/n$ est l'estimateur de la probabilité $p^*_{s,i}$ qu'un individu soit de la classe i et de l'espèce s.		$n_{s+} = \sum_i n_{s,i}$ $p_s = \sum_i w_i p_{s,i}$
...			
Total	n_{+i} : nombre d'individus de la classe. $n_{+i}/n = w_i$: poids de la classe		n : nombre total d'individus échantillonnés

probabilité de choix *a priori*) et situées dans des environnements différents.

La décomposition de la diversité jointe apporte en plus de la décomposition de la diversité γ une information sur l'ubiquité des espèces. La décomposition de Chiu *et al.*[5] se situe en réalité dans ce cadre.

[5] CHIU et al. (2014b). « Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers », cf. note 36, p. 130.

Pour la suite, les données sont organisées comme dans le tableau 7.1 : les individus appartiennent à l'espèce s et à la classe i avec la probabilité $p^*_{s,i}$ telle que $\sum_s \sum_i p^*_{s,i} = 1$ (au lieu de $\sum_i p_{s,i} = 1$ dans les chapitres précédents).

7.1 Récursivité de l'entropie

7.1.1 Entropie de Shannon

L'entropie de Shannon (et ses multiples) sont les seules fonctions H telles que :

$$H\left(\{p^*_{s,i}\}\right) - H\left(\{p_i\}\right) = \sum_i p_i H\left(\{\frac{p^*_{s,i}}{p_i}\}\right) \tag{7.1}$$

$$H\left(\{p^*_{s,i}\}\right) - H\left(\{p_s\}\right) = \sum_s p_s H\left(\{\frac{p^*_{s,i}}{p_s}\}\right) \tag{7.2}$$

Cette propriété est appelée récursivité.[6] Le premier terme de chaque équation est la différence entre l'entropie de la distribution complète $(p^*_{s,i})$, c'est-à-dire la diversité jointe de Gregorius qui sera appelée plutôt entropie jointe ici, et celle de la distribution marginale (l'entropie de la distribution des classes ou celle des espèces). Une démonstration rigoureuse est fournie par Baez *et al.*[7] d'après Faddeev,[8] mais il s'agit d'un résultat assez répandu dans la littérature, sous des formes diverses.[9] $H(\{p^*_{s,i}\}) - H(\{p_i\}) - H(\{p_s\})$ est l'information mutuelle, vue page 131.

Le terme de droite de l'équation 7.1 est l'entropie α de l'ensemble des classes si on les pondère par p_i car $p^*_{s,i}/p_i = p_{s,i}$. Par ailleurs, $H(\{p_s\})$ est l'entropie γ. En combinant les équations 7.1

[6] BACARO et al. (2014). « Using Shannon's recursivity to summarize forest structural diversity », cf. note 1, p. 169.

[7] J. C. BAEZ et al. (2011). « A characterization of entropy in terms of information loss ». In : *Entropy* 13.11, p. 1945–1957.

[8] D. K. FADDEEV (1956). « On the concept of entropy of a finite probabilistic scheme ». In : *Uspekhi Mat. Nauk* 1.67, p. 227–231.

[9] J. ACZÉL et Z. DARÓCZY (1975). *On measures of information and their characterizations*. New York : Academic Press, p. 234 ; RÉNYI (1961). « On Measures of Entropy and Information », cf. note 17, p. 60 ; BOURGUIGNON (1979). « Decomposable Income Inequality Measures », cf. note 47, p. 133.

et 7.2, on voit que l'entropie β est l'opposée de l'information mutuelle. L'information mutuelle est nulle quand les distributions des lignes et des colonnes sont indépendantes ($p^*_{s,i} = p_s p_i$), autrement dit quand les distributions marginales $\{p_s\}$ et $\{p_i\}$ suffisent à décrire la méta-communauté : l'information contenue par chaque communauté se limite à son poids. L'écart à l'indépendance implique que les distributions de probabilités sont différentes dans les communautés, ce qui définit la diversité β.

Si les colonnes sont les classes de diamètre par exemple, les informations fournies par la décomposition de la diversité sont nombreuses :

- Les diversités des colonnes sont les classiques diversités spécifiques, par classe de diamètre (α), globale (γ), et interclasses (β) ;
- La diversité des lignes renseigne sur l'homogénéité des diamètres de chaque espèce et globalement ;
- La diversité jointe est le nombre effectif de catégories (espèces $\times$ classes de diamètre) de la communauté.

7.1.2 Entropie HCDT

Baez *et al.* généralisent la récursivité à l'entropie HCDT et montrent que :

$$ {}^qH\left(\{p^*_{s,i}\}\right) - {}^qH\left(\{p_i\}\right) = \sum_i p_i^q \, {}^qH\left(\{\frac{p^*_{s,i}}{p_i}\}\right) \tag{7.3} $$

$$ {}^qH\left(\{p^*_{s,i}\}\right) - {}^qH\left(\{p_s\}\right) = \sum_s p_s^q \, {}^qH\left(\{\frac{p^*_{s,i}}{p_s}\}\right) \tag{7.4} $$

Par le même raisonnement que pour l'entropie de Shannon, on trouve donc que l'entropie entre les classes est l'opposée de l'information mutuelle calculée dans le cadre de l'entropie HCDT, à condition de définir les poids des classes égaux à p_i et l'entropie intraclasse comme la somme des entropies des classes pondérées par leur poids à la puissance q.

7.1.3 Intérêt de l'information mutuelle

Si les individus sont classés dans une matrice selon deux critères, et non dans un simple vecteur d'espèces, l'entropie HCDT de la distribution appelée *entropie jointe* est calculée par ${}^qH(\{p^*_{s,i}\})$.

L'entropie jointe peut être décomposée en la somme des entropies marginales et de l'information mutuelle. L'entropie intra-lignes (respectivement intra-colonnes) est définie comme la moyenne pondérée par le poids des lignes (resp. des colonnes) à la puissance q de l'entropie de chaque ligne (resp. colonne). Les lignes et les colonnes jouent un rôle interchangeable.

$$ {}^qH_{intra} = \sum_i p_i^q \, {}^qH(\{p_{s,i}\}) \tag{7.5} $$

$$ {}^qH_{inter} = -\left[{}^qH\left(\{p^*_{s,i}\}\right) - {}^qH\left(\{p_s\}\right) - {}^qH\left(\{p_i\}\right)\right] \tag{7.6} $$

$$ {}^qH_{marginale} = {}^qH\left(\{p_s\}\right) \tag{7.7} $$

Le rôle symétrique des lignes et des colonnes signifie que la taille des communautés n'est pas un choix arbitraire d'échantillonnage : la distribution des poids des communautés est équivalente à celle des probabilités des espèces. De la même façon que l'ordre de la diversité, q, donne plus ou moins d'importance aux espèces rares dans le calcul de la diversité intra et marginale, il donne plus ou moins d'importance aux communautés de petite taille dans le calcul des diversités intra et inter.

L'entropie ne peut être transformée en diversité que pour $q = 1$ (diversité de Shannon). Pour $q \neq 1$, l'entropie intra-classes n'est pas une entropie α (elle n'est pas égale à l'entropie de chacune des classes si toutes les classes sont identiques), le nombre effectif d'espèces correspondant n'est pas défini. L'information mutuelle de l'entropie HCDT n'a donc qu'un intérêt limité pour mesurer la diversité.

7.2 Décomposition de la diversité jointe

Gregorius[10] propose une décomposition de la diversité jointe qui complète celle de la diversité γ vue au chapitre précédent. Il se place dans le cadre d'un échantillonnage global de la méta-communauté, dans lequel le poids de chaque communauté est égal à la probabilité qu'un individu lui appartienne, et non un choix arbitraire.

La diversité β mesure la divergence entre la distribution moyenne des espèces dans les communautés et leur distribution dans la méta-communauté. Gregorius la qualifie de « diversité de répartition » (*apportionment*).[11] L'exemple suivant mesure la diversité β dans une méta-communauté très simple (Figure 7.1 [12]) :

[10] H.-R. Gregorius (2009). « Distribution of variation over populations. » In : *Theory in Biosciences* 128.3, p. 179–89 ; Gregorius (2010). « Linking Diversity and Differentiation », cf. note 44, p. 64.

[11] Gregorius (2014). « Partitioning of diversity : the "within communities" component », cf. note 27, p. 61.

```
q <- 2
df1 <- data.frame(Co1 = c(10, 20, 30, 10), Co2 = c(50, 20, 20,
    5), row.names = c("S1", "S2", "S3", "S4"))
MC1 <- MetaCommunity(df1)
```

```
summary(dp1 <- DivPart(q, MC1))

## HCDT diversity partitioning of order 2 of metaCommunity MC1
##
## Alpha diversity of communities:
```

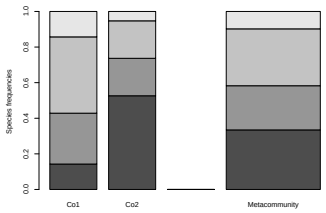


Figure 7.1 – Méta-communauté simple.

[12] Le code R nécessaire pour réaliser la figure est :

```
plot(MC1)
```

```
##      Co1      Co2
## 3.266667 2.714286
## Total alpha diversity of the communities:
## [1] 2.964968
## Beta diversity of the communities:
## [1] 1.182674
## Gamma diversity of the metacommunity:
##     None
## 3.506591
```

Répliquer les communautés à l'identique n'a aucun effet sur la diversité : la diversité α est la moyenne des diversités des communautés et la diversité γ est celle de leur assemblage dont les proportions ne changent pas. L'exemple suivant concerne une méta-communauté contenant les mêmes communautés, mais en double (Figure 7.2) :

```
df2 <- cbind(df1, df1)
MC2 <- MetaCommunity(df2)
```

```
summary(dp2 <- DivPart(q, MC2))

## HCDT diversity partitioning of order 2 of metaCommunity MC2
##
## Alpha diversity of communities:
##      Co1      Co2    Co1.1    Co2.1
## 3.266667 2.714286 3.266667 2.714286
## Total alpha diversity of the communities:
## [1] 2.964968
## Beta diversity of the communities:
## [1] 1.182674
## Gamma diversity of the metacommunity:
##     None
## 3.506591
```

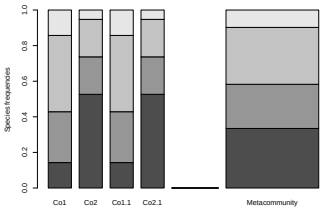


FIGURE 7.2 – Méta-communauté doublant les communautés de la précédente.

La diversité jointe ${}^qD(\{p^*_{s,i}\})$ est en revanche doublée au cours de l'opération : elle mesure le nombre effectif de catégories (combinaison de l'espèce et de la communauté) du tableau de données :

```
# Méta-communauté simple
Diversity(as.ProbaVector(df1), q)

##     None
## 5.642487

# Méta-communauté double
Diversity(as.ProbaVector(df2), q)

##     None
## 11.28497
```

La distribution de référence pour définir ce nombre effectif correspond à l'équitabilité maximale : S espèces équiprobables sont réparties dans I communautés équiprobables. Le nombre effectif est alors SI. Si nombre effectif d'espèces est égal à la diversité γ, le rapport entre la diversité jointe et la diversité γ, noté $D_{I|S}$, est le nombre effectif de communautés produisant la diversité jointe observée, compte-tenu de la diversité γ. Il mesure l'ubiquité des espèces, c'est-à-dire le nombre effectif de communautés dans lesquelles elles se trouvent, et complète la diversité γ en fournissant un niveau d'information supplémentaire. Gregorius le qualifie de « diversité de différenciation » (*differentiation*) parce que $D_{I|S}$

compare la distribution réelle à une distribution de référence, et pas seulement les distributions des communautés entre elles. En résumé :

- Le nombre effectif d'espèces ${}^q_iD$ de la communauté i est le nombre d'espèces équifréquentes nécessaires pour obtenir la diversité observée dans la communauté. La diversité α est le nombre d'espèces équifréquentes nécessaires dans chaque communauté pour obtenir la diversité observée ;
- La diversité β est le nombre de communautés totalement distinctes (chaque espèce n'est présente que dans une seule communauté), de même poids et de même diversité (cette diversité étant égale à la diversité α), égal au rapport entre la diversité γ et la diversité α ;
- La diversité γ est le nombre effectif d'espèces de l'assemblage des communautés ;
- L'ubiquité des espèces est le nombre effectif de communautés dans lesquelles les espèces sont présentes.

La décomposition complète de la diversité jointe s'écrit :

$$ {}^qD\left(\{p^*_{s,i}\}\right) = {}^qD_\alpha\, {}^qD_\beta\, {}^qD_{I|S} \tag{7.8}$$

La méta-communauté suivante illustre cette décomposition : elle contient $D_{I|S} = 3$ répliques identiques de $D_\beta = 2$ communautés équiprobables totalement distinctes, comprenant chacune $D_\beta = 2$ espèces équiprobables (Figure 7.3). Sa diversité jointe est 12 (le nombre de cellules non nulles dans la matrice) :

```
q <- 2  # Peu importe q
df1 <- data.frame(Co1 = c(10, 10, 0, 0), Co2 = c(0, 0, 10, 10),
    row.names = c("S1", "S2", "S3", "S4"))
df3 <- cbind(df1, df1, df1)
MC <- MetaCommunity(df3)
```

```
summary(dp <- DivPart(q, MC))

## HCDT diversity partitioning of order 2 of metaCommunity MC
##
## Alpha diversity of communities:
##   Co1   Co2 Co1.1 Co2.1 Co1.2 Co2.2
##     2     2     2     2     2     2
## Total alpha diversity of the communities:
## [1] 2
## Beta diversity of the communities:
## [1] 2
## Gamma diversity of the metacommunity:
## None
##    4

(Djointe <- Diversity(as.ProbaVector(df3), q))

## None
##   12

(Dis <- Djointe/dp$GammaDiversity)

## None
##    3
```

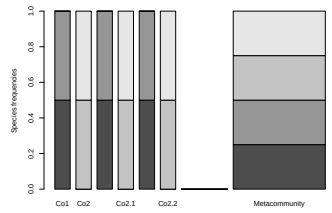


FIGURE 7.3 – Méta-communauté illustrant la décomposition complète de la diversité jointe.

La définition de la diversité α de Chiu *et al.*[13] (page 130) est

[13] CHIU et al. (2014b). « Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers », cf. note 36, p. 130.

la diversité jointe, normalisée par le nombre de communautés :

$$ {}^qD_\alpha = \frac{1}{I}\left(\sum_s \sum_i (w_i p_{s,i})^q\right)^{1/(1-q)} = \frac{1}{I}\,{}^qD(\{p^*_{s,i}\}) \tag{7.9} $$

En introduisant cette définition dans la décomposition de la diversité jointe, équation 7.8, il vient :

$$ {}^qD_\beta = \frac{{}^qD_\gamma}{{}^qD_\alpha} = \frac{I}{{}^qD_{I|S}} \tag{7.10} $$

La définition de la diversité α de Chiu *et al.* implique que la diversité β est le rapport entre le nombre de communautés et l'ubiquité des espèces, c'est-à-dire une mesure de différenciation et non de répartition.

$D_{S|I}$ est le nombre effectif d'espèces des communautés au sens de la diversité de différenciation. Il est calculé symétriquement à $D_{I|S}$, en divisant la diversité jointe par la diversité des poids des communautés :

$$ {}^qD_{S|I} = \frac{{}^qD\left(\{p^*_{s,i}\}\right)}{{}^qD\left(\{p_i\}\right)} \tag{7.11} $$

Gregorius montre que ${}^qD_{S|I}$ égale la diversité α selon Jost (page 129) :

$$ {}^qD_{S|I} = \left(\sum_s \sum_i \frac{p_i^q}{\sum_i p_i^q} p_{s,i}^q\right)^{1/(1-q)} \tag{7.12} $$

La diversité α au sens de Jost donc est elle aussi une diversité de différenciation.

Chapitre 8

Conclusion

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'occurrence 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.

Chao, Chiu et Jost[1] proposent un cadre fondé exclusivement sur les nombres de Hill sans référence à l'entropie. Les principales différences avec celui présenté ici concernent la définition de la diversité α et donc celle de la diversité β et l'absence d'estimateurs à biais réduit.

L'entropie phylogénétique généralise les indices de diversité classique (Figure 8.1), intègre si nécessaire la distance entre espèces, peut être décomposée 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 diversité d'ordre 2 présente de nombreuses particularités et avantages qui font qu'elle est très employée :

Estimation Tous les estimateurs de ${}^qH$ sont biaisés, d'autant plus que q est petit. Seul l'estimateur de l'indice de Simpson, ${}^2\tilde{H} = ({}^n/_{n-1})(1 - \sum_s \hat{p}_s^2)$ est non biaisé. Comme l'entropie de Rao est la somme sur les périodes de l'arbre des entropies de Simpson, ${}^2\tilde{H}(T) = ({}^n/_{n-1})^2\bar{H}(T)$ est donc sans biais.

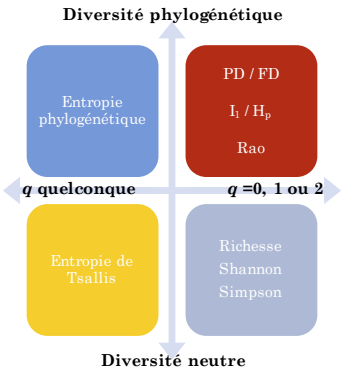


Figure 8.1 – Généralisation des mesures classiques de diversité. La figure présente les mesures d'entropie, qui peuvent toutes être transformées en mesures de diversité au sens strict. Toutes ces entropies sont des cas particuliers de l'entropie de Ricotta et Szeidl, qui peut être transformée en diversité de Leinster et Cobbold.

[1] A. Chao et al. (2014b). « Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity and Related Similarity/Differentiation Measures Through Hill Numbers ». In : *Annual Review of Ecology, Evolution, and Systematics* 45, p. 297–324.

Partition L'entropie de Simpson peut être décomposée comme une variance. La contribution de chaque communauté à la variance inter n'est pas la même que dans la décomposition générale de l'entropie HCDT mais la variance inter est bien égale à l'entropie β. La décomposition de la diversité de Pélissier et Couteron est identique. Dans le cas particulier de l'entropie de Rao, l'entropie phylogénétique, sans normalisation par la hauteur de l'arbre, est égale à la somme pondérée des dissimilarités (divergences de Jensen-Shannon) entre les paires de communautés. Cette propriété permet de représenter dans un même espace les espèces et les communautés par une double analyse en composantes principales.

Égalité des diversités Pour $q = 2$ seulement, toutes les mesures de diversité sont identiques :

- Le nombre équivalents d'espèces de Hurlbert est égal à ${}^2D$, le nombre de Hill ;
- La diversité de Leinster et Cobbold est égale à la diversité phylogénétique.

La diversité d'ordre 1 a en revanche des propriétés avantageuses pour la décomposition : toutes les définitions de diversité α sont dans ce cas identiques, et l'information mutuelle se confond avec la diversité β.

Le package *entropart* permet de calculer avec R la diversité d'une méta-communauté à partir de données d'abondances et d'un arbre (phylogénétique) ou d'une matrice de distances (fonctionnelles).

D'autres mesures de diversité ont été présentées ici, d'autres encore peuvent être trouvées dans la littérature. Elles ont en général un moins bon support mathématique, surtout les mesures de diversité β, elles ne peuvent pas être décomposées, corrigées ou interprétées. La diversité de Hurlbert possède toutes ses propriétés et une correction très efficace du biais d'estimation, mais elle n'est pas généralisée à la diversité phylogénétique.

Les principales limites des mesures de diversité sont d'ordre conceptuel et pratique. L'axiome de symétrie interdit de donner un rôle particulier à une espèce : il n'est donc pas possible conceptuellement de prendre en compte les espèces focales[2] dans la quantification de la diversité (espèces emblématiques, espèces clé de voûte,[3] espèces parapluie,[4] espèces ingénieures[5]...). Leur présence peut être décisive pour la conservation, mais il s'agit clairement d'une question distincte de celle de la diversité.

Sur le plan pratique, la mesure de la diversité se réduit souvent à un groupe biologique (dans ce document, les arbres). Estimer la biodiversité totale est un objectif hors d'atteinte. Si la diversité à prendre en compte est au-delà des données disponibles, il est nécessaire d'utiliser des *indicateurs*[6] qui peuvent être la diversité

[2] M. A. Zacharias et J. C. Roff (2001). « Use of focal species in marine conservation and management : a review and critique ». In : *Aquatic Conservation : Marine and Freshwater Ecosystems* 11.1, p. 59–76.

[3] R. T. Paine (1969). « A note on trophic complexity and community stability ». In : *American naturalist* 103.929, p. 91–93.

[4] J.-M. Roberge et P. Angelstam (2004). « Usefulness of the Umbrella Species Concept ». In : *Conservation Biology* 18.1, p. 76–85.

[5] C. G. Jones et al. (1994). « Organisms as Ecosystem Engineers ». In : *Oikos,* 69.3, p. 373–386.

[6] A. Balmford et al. (2003). « Measuring the changing state of nature ». In : *Trends in Ecology & Evolution* 18.7, p. 326–330.

d'un ou quelques groupes représentatifs, voire la présence d'espèces indicatrices.[7] Des indicateurs indirects sont la diversité des habitats ou même l'étendue de certains habitats reconnus comme les plus favorables à un haut niveau de diversité. Une réflexion en cours[8] consiste à établir une liste de variables pertinentes (*Essential biodiversity variables*) qui puissent constituer des indicateurs standardisés à l'échelle globale pour permettre la comparaison des données. Les indicateurs dévient souvent de la notion stricte de diversité et peuvent mesurer tout autre chose, comme l'indice BII (*Biodiversity Intactness Index*[9]) qui est la moyenne pondérée par la surface des habitats et le nombre d'espèces concernées de la variation des effectifs des populations induite par l'anthropisation. Les effectifs ne sont pas pris en compte dans la mesure de la diversité (sauf dans le calcul des biais d'estimation), seules les proportions importent : l'indice BII mesure donc la dégradation des habitats plus que la variation de la biodiversité. L'argument, développé par Balmford *et al.*,[10] est que la diminution des effectifs des espèces suivies est corrélé à la disparition des espèces plus vulnérables ; l'effectif d'une population est aussi un indicateur de sa diversité génétique. La moyenne géométrique des abondances relatives[11] compare les abondances des espèces au cours du temps à leur valeur initiale. L'indice LPI (*Living Planet Index*)[12] s'appuie sur ce même principe : il mesure l'évolution dans le temps des effectifs d'un grand nombre de populations de vertébrés.

La recherche sur la mesure de la diversité est très active, ce document évolue donc régulièrement pour prendre en compte les avancées récentes.

[7] F. E. Clements (1916). *Plant succession ; an analysis of the development of vegetation*. Washington, D.C. : Carnegie Institution of Washington.

[8] H. M. Pereira et al. (2013). « Essential biodiversity variables ». In : *Science* 339, p. 277–278.

[9] R. J. Scholes et R. Biggs (2005). « A Biodiversity Intactness Index ». In : *Nature* 434.7029, p. 45–49.

[10] Balmford et al. (2003). Cf. note 6.

[11] S. T. Buckland et al. (2011). « The geometric mean of relative abundance indices : a biodiversity measure with a difference ». In : *Ecosphere* 2.9, art100.

[12] J. Loh et al. (2005). « The Living Planet Index : using species population time series to track trends in biodiversity. » In : *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 360.1454, p. 289–295.

Bibliographie

Aczél, J. et Z. Daróczy (1975). *On measures of information and their characterizations*. New York : Academic Press, p. 234 (cf. p. 170).

Adelman, M. A. (1969). « Comment on the "H" Concentration Measure as a Numbers-Equivalent ». In : *The Review of Economics and Statistics* 51.1, p. 99–101 (cf. p. 61).

Agapow, P. M., O. R. P. Binindal-Emonds, K. A. Crandall, J. L. Gittleman, G. M. Mace, J. C. Marshall et A. Purvis (2004). « The impact of species concept on biodiversity studies ». In : *The Quarterly Review of Biology* 79.2, p. 161–179 (cf. p. 45).

Alatalo, R. V. (1981). « Problems in the Measurement of Evenness in Ecology ». In : *Oikos* 37.2, p. 199–204 (cf. p. 77, 80).

Allen, B., M. Kon et Y. Bar-Yam (2009). « A New Phylogenetic Diversity Measure Generalizing the Shannon Index and Its Application to Phyllostomid Bats ». In : *American Naturalist* 174.2, p. 236–243 (cf. p. 101, 144).

Alonso, D. et A. J. McKane (2004). « Sampling Hubbell's neutral theory of biodiversity ». In : *Ecology Letters* 7.10, p. 901–910 (cf. p. 14).

Anderson, M. J., K. E. Ellingsen et B. H. McArdle (2006). « Multivariate dispersion as a measure of beta diversity. » In : *Ecology letters* 9.6, p. 683–93 (cf. p. 161, 167).

Anderson, M. J., T. O. Crist, J. M. Chase, M. Vellend, B. D. Inouye, A. L. Freestone, N. J. Sanders, H. V. Cornell, L. S. Comita, K. F. Davies, S. P. Harrison, N. J. B. Kraft, J. C. Stegen et N. G. Swenson (2011). « Navigating the multiple meanings of β diversity : a roadmap for the practicing ecologist ». In : *Ecology Letters* 14.1, p. 19–28 (cf. p. 123–126).

Arrhenius, O. (1921). « Species and Area ». In : *Journal of Ecology* 9.1, p. 95–99 (cf. p. 10, 11).

Ashbridge, J. et I. B. J. Goudie (2000). « Coverage-adjusted estimators for mark-recapture in heterogeneous populations ». In : *Communications in Statistics - Simulation and Computation* 29.4, p. 1215–1237 (cf. p. 49).

Bacaro, G. et C. Ricotta (2007). « A spatially explicit measure of beta diversity ». In : *Community Ecology* 8.1, p. 41–46 (cf. p. 149).

Bacaro, G., M. Gioria et C. Ricotta (2012). « Testing for Differences in Beta Diversity From Plot-To-Plot Dissimilarities ». In : *Ecological Research* 27.2, p. 285–292 (cf. p. 168).

— (2013). « Beta diversity reconsidered ». In : *Ecological Research* 28.3, p. 537–540 (cf. p. 168).

Bacaro, G., C. Ricotta, M. Marignani, D. Torri et A. Chiarucci (2014). « Using Shannon's recursivity to summarize forest structural diversity ». In : *Forests Trees and Livelihoods* 23.3, p. 211–216 (cf. p. 169, 170).

Baez, J. C., T Fritz et T Leinster (2011). « A characterization of entropy in terms of information loss ». In : *Entropy* 13.11, p. 1945–1957 (cf. p. 170).

Balmford, A., M. J. B. Green et M. G. Murray (1996a). « Using higher-taxon richness as a surrogate for species richness : I. Regional tests ». In : *Proceedings of the Royal Society B* 263, p. 1267–1274 (cf. p. 43).

Balmford, A., A. H. M. Jayasuriya et M. J. B. Green (1996b). « Using higher-taxon richness as a surrogate for species richness : II. Local applications ». In : *Proceedings of the Royal Society B* 263, p. 1571–1575 (cf. p. 42).

Balmford, A., R. E. Green et M. Jenkins (2003). « Measuring the changing state of nature ». In : *Trends in Ecology & Evolution* 18.7, p. 326–330 (cf. p. 178, 179).

Baraloto, C., C. E. T. Paine, S. Patiño, D. Bonal, B. Hérault et J. Chave (2010). « Functional trait variation and sampling strategies in species rich plant communities ». In : *Functional Ecology* 24, p. 208–216 (cf. p. 85).

Barberousse, A. et S. Samadi (2014). « La taxonomie et les collections d'histoire naturelle à l'heure de la sixième extinction ». In : *La biodiversité en question. Enjeux philosophiques, éthiques et scientifiques*. Sous la dir. d'E. Casetta et J. Delord. Paris : Editions Matériologiques. Chap. 6, p. 155–182 (cf. p. 45).

Baselga, A. (2010a). « Multiplicative partition of true diversity yields independent alpha and beta components ; additive partition does not ». In : *Ecology* 91.7, p. 1974–1981 (cf. p. 123, 128).

— (2010b). « Partitioning the turnover and nestedness components of beta diversity ». In : *Global Ecology and Biogeography* 19.1, p. 134–143 (cf. p. 160).

Basharin, G. P. (1959). « On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables ». In : *Theory of Probability and its Applications* 4.3, p. 333–336 (cf. p. 48).

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 et M. Leponce (2012). « Arthropod Diversity in a Tropical Forest ». In : *Science* 338.6113, p. 1481–1484 (cf. p. 28).

Beck, C. (2009). « Generalised Information and Entropy Measures in Physics ». In : *Contemporary Physics* 50.4, p. 495–510 (cf. p. 75).

Beck, J. et W. Schwanghart (2010). « Comparing measures of species diversity from incomplete inventories : an update ». In : *Methods in Ecology and Evolution* 1.1, p. 38–44 (cf. p. 34, 49).

Beck, J., J. D. Holloway et W. Schwanghart (2013). « Undersampling and the measurement of beta diver-

sity ». In : *Methods in Ecology and Evolution* 4.4, p. 370–382 (cf. p. 141).

Bellman, R., K. L. Cooke et J. A. Lockett (1970). *Algorithms, Graphs and Computers*. New York : Academic Press, Inc. (cf. p. 166).

Berger, W. H. et F. L. Parker (1970). « Diversity of planktonic foraminifera in deep-sea sediments ». In : *Science* 168.3937, p. 1345–1347 (cf. p. 65).

Blandin, P. (2014). « La diversité du vivant avant (et après) la biodiversité : repères historiques et épistémologiques ». In : *La biodiversité en question. Enjeux philosophiques, éthiques et scientifiques*. Sous la dir. d'E. Casetta et J. Delord. Paris : Editions Matériologiques. Chap. 1, p. 31–68 (cf. p. 1).

Bonachela, J. A., H. Hinrichsen et M. A. Muñoz (2008). « Entropy estimates of small data sets ». In : *Journal of Physics A : Mathematical and Theoretical* 41.202001, p. 1–9 (cf. p. 50).

Bordenave, B. G. et J. J. De Granville (1998). « Les mesures de la biodiversité : un outil de conservation en forêt guyanaise ». In : *JATBA, Revue d'Ethnobiologie* 40.1-2, p. 433–446 (cf. p. 15).

Bordenave, B. G., J. J. De Granville et K Steyn (2011). « Quantitative botanical diversity descriptors to set conservation priorities in Bakhuis Mountains rainforest, Suriname ». In : *Botanical Journal of the Linnean Society* 167.1, p. 94–130 (cf. p. 15).

Borland, L., A. R. Plastino et C. Tsallis (1998). « Information gain within nonextensive thermostatistics ». In : *Journal of Mathematical Physics* 39.12, p. 6490–6501 (cf. p. 134).

Bourguignon, F. (1979). « Decomposable Income Inequality Measures ». In : *Econometrica* 47.4, p. 901–920 (cf. p. 133, 170).

Braak, C. J. F. ter (1983). « Principal components biplots and alpha and beta diversity ». In : *Ecology* 64.3, p. 454–462 (cf. p. 123, 163).

Braun-Blanquet, J. (1928). *Pflanzensoziologie : Grundzüge der Vegetationskunde*. Berlin : Springer (cf. p. 9).

Bray, J. R. et J. T. Curtis (1957). « An Ordination of the Upland Forest Communities of Southern Wisconsin ». In : *Ecological Monographs* 27.4, p. 325–349 (cf. p. 161).

Brillouin, L. (1962). *Science and information theory*. 2nd ed. Oxford : Academic Press. Chap. 351 (cf. p. 47).

Broms, K. M., M. B. Hooten et R. M. Fitzpatrick (2014). « Accounting for Imperfect Detection in Hill Numbers for Biodiversity Studies ». In : *Methods in Ecology and Evolution* 6.1, p. 99–108 (cf. p. 74).

Brose, U., N. D. Martinez et R. J. Williams (2003). « Estimating species richness : Sensitivity to sample coverage and insensitivity to spatial patterns ». In : *Ecology* 84.9, p. 2364–2377 (cf. p. 34, 35, 70).

Brülhart, M. et R. Traeger (2005). « An Account of Geographic Concentration Patterns in Europe ». In : *Regional Science and Urban Economics* 35.6, p. 597–624 (cf. p. 75).

Buckland, S. T., A. C. Studeny, A. E. Magurran, J. B. Illian et S. E. Newson (2011). « The geometric mean of relative abundance indices : a biodiversity measure with a difference ». In : *Ecosphere* 2.9, art100 (cf. p. 179).

Bulla, L. (1994). « An index of evenness and its associated diversity measure ». In : *Oikos* 70.1, p. 167–172 (cf. p. 77).

Bulmer, M. G. (1974). « On Fitting the Poisson Lognormal Distribution to Species-Abundance Data ». In : *Biometrics* 30.1, p. 101–110 (cf. p. 48).

Burnham, A. K. P. et W. S. Overton (1978). « Estimation of the Size of a Closed Population when Capture Probabilities vary Among Animals ». In : *Biometrika* 65.3, p. 625–633 (cf. p. 31).

Burnham, K. P. et W. S. Overton (1979). « Robust Estimation of Population Size When Capture Probabilities Vary Among Animals ». In : *Ecology* 60.5, p. 927–936 (cf. p. 28, 31, 70).

Buuren, S. van et K. Groothuis-Oudshoorn (2011). « mice : Multivariate Imputation by Chained Equations in R ». In : *Journal of Statistical Software* 45.3, p. 1–67 (cf. p. 85).

Buuren, S. van, J. P. L. Brand, C. G. M. Groothuis-Oudshoorn et D. B. Rubin (2006). « Fully conditional specification in multivariate imputation ». In : *Journal of Statistical Computation and Simulation* 76.12, p. 1049–1064 (cf. p. 85).

Cadotte, M. W., T. J. Davies, J Regetz, S. W. Kembel, E Cleland et T. H. Oakley (2010). « Phylogenetic diversity metrics for ecological communities : integrating species richness, abundance and evolutionary history ». In : *Ecology Letters* 13.1, p. 96–105 (cf. p. 94).

Cailliez, F. (1983). « The analytical solution of the additive constant problem ». In : *Psychometrika* 48, p. 305–310 (cf. p. 84).

Caldarelli, G., A. Capocci, P. De Los Rios et M. A. Muñoz (2002). « Scale-free networks from varying vertex intrinsic fitness. » In : *Physical Review Letters* 89.25, p. 258702 (cf. p. 43).

Campos, D. et J. F. Isaza (2009). « A geometrical index for measuring species diversity ». In : *Ecological Indicators* 9.4, p. 651–658 (cf. p. 115).

Cao, L. et M. Grabchak (2014). *EntropyEstimation : Estimation of Entropy and Related Quantities* (cf. p. 13).

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 et S Naeem (2012). « Biodiversity loss and its impact on humanity ». In : *Nature* 486.7401, p. 59–67 (cf. p. 2).

Cartozo, C. C., D. Garlaschelli, C. Ricotta, M. Barthélemy et G. Caldarelli (2008). « Quantifying the taxonomic diversity in real species communities ». In : *Journal of Physics A : Mathematical and Theoretical* 41.22 (cf. p. 43).

Casetta, E. (2014). « Évaluer et conserver la biodiversité face au problème des espèces ». In : *La biodiversité en question. Enjeux philosophiques, éthiques et scientifiques*. Sous la dir. d'E. Casetta et J. Delord. Paris : Editions Matériologiques. Chap. 5, p. 139–154 (cf. p. 44).

Cavender-Bares, J., K. H. Kozak, P. V. A. Fine et S. W. Kembel (2009). « The merging of community ecology and phylogenetic biology ». In : *Ecology Letters* 12.7, p. 693–715 (cf. p. 86).

Ceriani, L. et P. Verme (2012). « The origins of the Gini index : extracts from Variabilità e Mutabilità (1912) by Corrado Gini ». In : *Journal of Economic Inequality* 10.3, p. 421–443 (cf. p. 76).

Champely, S. et D. Chessel (2002). « Measuring biological diversity using Euclidean metrics ». In : *Environmental and Ecological Statistics* 9.2, p. 167–177 (cf. p. 99, 140).

Chao, A. (1984). « Nonparametric estimation of the number of classes in a population ». In : *Scandinavian Journal of Statistics* 11, p. 265–270 (cf. p. 28, 29, 31).

— (1987). « Estimating the population size for capture-recapture data with unequal catchability ». In : *Biometrics* 43.4, p. 783–791 (cf. p. 29, 30).

— (2004). « Species richness estimation. » In : *Encyclopedia of Statistical Sciences*. Sous la dir. de N Balakrish-

nan, C. B. Read et B Vidakovic. 2nd ed. New York : Wiley (cf. p. 29).

Chao, A. et L. Jost (2012). « Coverage-based rarefaction and extrapolation : standardizing samples by completeness rather than size ». In : *Ecology* 93.12, p. 2533–2547 (cf. p. 26, 28, 50, 69).

— (2015). « Estimating diversity and entropy profiles via discovery rates of new species ». In : *Methods in Ecology and Evolution* 6.8, p. 873–882 (cf. p. 68, 71, 153).

Chao, A. et S.-M. Lee (1992). « Estimating the Number of Classes Via Sample Coverage ». In : *Journal of the American Statistical Association* 87.417, p. 210–217 (cf. p. 30).

Chao, A. et C.-W. Lin (2012). « Nonparametric Lower Bounds for Species Richness and Shared Species Richness under Sampling without Replacement ». English. In : *Biometrics* 68.3, p. 912–921 (cf. p. 28).

Chao, A. et T.-J. Shen (2003). « Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample ». In : *Environmental and Ecological Statistics* 10.4, p. 429–443 (cf. p. 49, 69, 110).

— (2010). *Program SPADE : Species Prediction And Diversity Estimation. Program and user's guide.* Hsin-Chu, Taiwan (cf. p. 31).

Chao, A., S.-M. Lee et T.-C. Chen (1988). « A generalized Good's nonparametric coverage estimator ». In : *Chinese Journal of Mathematics* 16, p. 189–199 (cf. p. 27).

Chao, A., R. L. Chazdon, R. K. Colwell et T.-J. Shen (2004). « A new statistical approach for assessing similarity of species composition with incidence and abundance data ». In : *Ecology Letters* 8.2, p. 148–159 (cf. p. 158).

Chao, A., L. Jost, S. C. Chiang, Y. H. Jiang et R. L. Chazdon (2008). « A Two-Stage Probabilistic Approach to Multiple-Community Similarity Indices ». In : *Biometrics* 64.4, p. 1178–1186 (cf. p. 143).

Chao, A., C.-H. Chiu et L. Jost (2010). « Phylogenetic diversity measures based on Hill numbers ». In : *Philosophical Transactions of the Royal Society B* 365.1558, p. 3599–3609 (cf. p. 27, 64, 102, 104, 105).

Chao, A., C.-H. Chiu et T. C. Hsieh (2012). « Proposing a resolution to debates on diversity partitioning ». In : *Ecology* 93.9, p. 2037–2051 (cf. p. 123, 128, 129, 136, 144).

Chao, A., Y.-T. Wang et L. Jost (2013). « Entropy and the species accumulation curve : a novel entropy estimator via discovery rates of new species ». In : *Methods in Ecology and Evolution* 4.11, p. 1091–1100 (cf. p. 13, 50, 68, 70, 72, 131).

Chao, A., N. J. Gotelli, T. C. Hsieh, E. L. Sander, K. H. Ma, R. K. Colwell et A. M. Ellison (2014a). « Rarefaction and extrapolation with Hill numbers : A framework for sampling and estimation in species diversity studies ». In : *Ecological Monographs* 84.1, p. 45–67 (cf. p. 54, 74).

Chao, A., C.-H. Chiu et L. Jost (2014b). « Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity and Related Similarity/Differentiation Measures Through Hill Numbers ». In : *Annual Review of Ecology, Evolution, and Systematics* 45, p. 297–324 (cf. p. 177).

Chao, A., C.-H. Chiu, T. C. Hsieh, T. Davis, D. A. Nipperess et D. P. Faith (2015a). « Rarefaction and extrapolation of phylogenetic diversity ». In : *Methods in Ecology and Evolution* 6.4, p. 380–388 (cf. p. 96).

Chao, A., T. C. Hsieh, R. L. Chazdon, R. K. Colwell et N. J. Gotelli (2015b). « Unveiling the Species-Rank Abundance Distribution by Generalizing Good-Turing Sample Coverage Theory ». In : *Ecology* 96.5, p. 1189–1201 (cf. p. 23, 70, 72).

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 et S. Diaz (2000). « Consequences of changing biodiversity ». In : *Nature* 405.6783, p. 234–242 (cf. p. 2).

Chase, J. M. et J. A. Myers (2011). « Disentangling the importance of ecological niches from stochastic processes across scales ». In : *Philosophical Transactions of the Royal Society B : Biological Sciences* 366.1576, p. 2351–2363 (cf. p. 167).

Chase, J. M., N. J. B. Kraft, K. G. Smith, M. Vellend et B. D. Inouye (2011). « Using null models to disentangle variation in community dissimilarity from variation in α-diversity ». In : *Ecosphere* 2.2, art. 24 (cf. p. 166).

Chave, J. et E. G. Leigh (2002). « A Spatially Explicit Neutral Model of β-Diversity in Tropical Forests ». In : *Theoretical Population Biology* 62.2, p. 153–168 (cf. p. 147).

Chave, J., G. Chust et C. Thébaud (2007). « The importance of phylogenetic structure in biodiversity studies ». In : *Scaling biodiversity*. Sous la dir. de D. Storch, P. Marquet et J. Brown. Santa Fe : Institute Editions, p. 150–167 (cf. p. 83, 85, 148).

Chiu, C.-H. et A. Chao (2014). « Distance-based functional diversity measures and their decomposition : a framework based on hill numbers. » In : *PloS one* 9.7, e100014 (cf. p. 109, 114, 155).

Chiu, C.-H., Y.-T. Wang, B. A. Walther et A. Chao (2014a). « An Improved Nonparametric Lower Bound of Species Richness via a Modified Good-Turing Frequency Formula ». In : *Biometrics* 70.3, p. 671–682 (cf. p. 26, 30, 35).

Chiu, C.-H., L. Jost et A. Chao (2014b). « Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers ». In : *Ecological Monographs* 84.1, p. 21–44 (cf. p. 130, 170, 174).

Chytrý, M. et Z. Otýpková (2003). « Plot sizes used for phytosociological sampling of European vegetation ». In : *Journal of Vegetation Science* 14.4, p. 563–570 (cf. p. 9).

Clarke, K. R. et R. M. Warwick (2001). « A further biodiversity index applicable to species lists : variation in taxonomic distinctness ». In : *Marine Ecology-Progress Series* 216, p. 265–278 (cf. p. 9, 84, 91).

Clements, F. E. (1916). *Plant succession ; an analysis of the development of vegetation.* Washington, D.C. : Carnegie Institution of Washington (cf. p. 179).

Clench, H. K. (1979). « How to make regional lists of butterflies : some thoughts ». In : *Journal of the Lepidopterists' Society* 33.4, p. 216–231 (cf. p. 39).

Coleman, B. D. (1981). « On random placement and species-area relations ». In : *Mathematical Biosciences* 54.3-4, p. 191–215 (cf. p. 12, 20).

Coleman, B. D., M. A. Mares, M. R. Willig et Y.-H. Hsieh (1982). « Randomness, area, and species richness ». In : *Ecology* 63.4, p. 1121–1133 (cf. p. 20, 21).

Colwell, R. K. et J. A. Coddington (1994). « Estimating Terrestrial Biodiversity through Extrapolation ». In : *Philosophical Transactions : Biological Sciences* 345.1311, p. 101–118 (cf. p. 32, 39).

Colwell, R. K., A. Chao, N. J. Gotelli, S.-Y. Lin, C. X. Mao, R. L. Chazdon et J. T. Longino (2012). « Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages ». In : *Journal of Plant Ecology* 5.1, p. 3–21 (cf. p. 13).

Conceição, P. et P. Ferreira (2000). *The Young Person's Guide to the Theil Index : Suggesting Intuitive Inter-*

pretations and Exploring Analytical Applications. Rapp. tech. Austin, Texas, 54 p. (Cf. p. 48, 59).

Condit, R., N. Pitman, E. G. J. Leigh, J. Chave, J. Terborgh, R. B. Foster, P. Núñez, S. Aguilar, R. Valencia, G. Villa, H. C. Muller-Landau, E. Losos et S. P. Hubbell (2002). « Beta-diversity in tropical forest trees ». In : *Science* 295.5555, p. 666–669 (cf. p. 147).

Connell, J. H. (1978). « Diversity in Tropical Rain Forests and Coral Reefs ». In : *Science* 199.4335, p. 1302–1310 (cf. p. 44).

Connor, E. F. et E. D. McCoy (1979). « The Statistics and Biology of the Species-Area Relationship ». In : *The American Naturalist* 113.6, p. 791–833 (cf. p. 11).

Cormack, R. M. (1989). « Log-Linear Models for Capture-Recapture ». In : *Biometrics* 45.2, p. 395–413 (cf. p. 31).

Cornelissen, J. H. C., S. Lavorel, E. Garnier, S. Diaz, N. Buchmann, D. E. Gurvich, P. B. Reich, H. ter Steege, H. D. Morgan, M. G. A. van der Heijden, J. G. Pausas et H. Poorter (2003). « A handbook of protocols for standardised and easy measurement of plant functional traits worldwide ». In : *Australian Journal of Botany* 51.4, p. 335–380 (cf. p. 85).

Cousins, S. H. (1991). « Species diversity measurement : Choosing the right index ». In : *Trends in Ecology and Evolution* 6.6, p. 190–192 (cf. p. 83).

Couteron, P. et R. Pélissier (2004). « Additive apportioning of species diversity : towards more sophisticated models and analyses ». In : *Oikos* 107.1, p. 215–221 (cf. p. 150).

Cowell, F. (2000). « Measurement of Inequality ». In : *Handbook of Income Distribution*. Sous la dir. d'A. B. Atkison et F Bourguignon. T. 1. 6. North Holland. Chap. 2, p. 87–166 (cf. p. 75).

Cracraft, J. (1983). « Species Concepts and Speciation Analysis ». English. In : *Current Ornithology Volume 1*. Sous la dir. de R. F. Johnston. T. 1. Current Ornithology. Springer US, p. 159–187 (cf. p. 45).

Cressie, N. et T. R. C. Read (1984). « Multinomial Goodness-of-Fit Tests ». In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 46.3, p. 440–464 (cf. p. 75, 134).

Cressie, N. A. (1993). *Statistics for spatial data*. New York : John Wiley & Sons, p. 1–900 (cf. p. 22).

Crist, T. O., J. A. Veech, J. C. Gering et K. S. Summerville (2003). « Partitioning species diversity across landscapes and regions : A hierarchical analysis of alpha, beta, and gamma diversity ». In : *The American Naturalist* 162.6, p. 734–743 (cf. p. 141, 151).

Crozier, R. H. (1997). « Preserving the Information Content of Species : Genetic Diversity , Phylogeny , and Conservation Worth Interaction with Reasons Justifying the Preservation ». In : *Annual Review of Ecology and Systematics* 28, p. 243–268 (cf. p. 85).

Cutrini, E. (2010). « Specialization and Concentration from a Twofold Geographical Perspective : Evidence from Europe ». In : *Regional Studies* 44.3, p. 315–336 (cf. p. 169).

Dalton, H. (1920). « The measurement of the inequality of incomes ». In : *The Economic Journal* 30.119, p. 348–361 (cf. p. 60, 76).

Daróczy, Z. (1970). « Generalized information functions ». In : *Information and Control* 16.1, p. 36–51 (cf. p. 62).

Das, J. K. et J. Nautiyal (2004). « Forest variability index : a vector quantifying forest stand diversity and forest compactness ». In : *Forest Policy and Economics* 6.3-4, p. 271–288 (cf. p. 169).

Dauby, G. et O. J. Hardy (2012). « Sampled-based estimation of diversity sensu stricto by transforming Hurlbert diversities into effective number of species ». In : *Ecography* 35.7, p. 661–672 (cf. p. 27, 53, 54, 64).

Davis, H. T. (1941). *The theory of econometrics*. Bloomington, Indiana : The Principia Press (cf. p. 57).

De Bello, F., S. Lavergne, C. N. Meynard, J. Lepš et W. Thuiller (2010). « The partitioning of diversity : showing Theseus a way out of the labyrinth ». In : *Journal of Vegetation Science* 21.5, p. 992–1000 (cf. p. 140).

DeLong, D. C. J. (1996). « Defining Biodiversity ». In : *Wildlife Society Bulletin* 24.4, p. 738–749 (cf. p. 1).

Delord, J. (2014). « La biodiversité : imposture scientifique ou ruse épistémologique ? » In : *La biodiversité en question. Enjeux philosophiques, éthiques et scientifiques*. Sous la dir. d'E. Casetta et J. Delord. Paris : Editions Matériologiques. Chap. 3, p. 83–118 (cf. p. 1).

Dengler, J. (2009). « Which function describes the species-area relationship best ? A review and empirical evaluation ». In : *Journal of Biogeography* 36.4, p. 728–744 (cf. p. 6, 24).

Devictor, V., D. Mouillot, C. Meynard, F. Jiguet, W. Thuiller et N. Mouquet (2010). « 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, p. 1030–40 (cf. p. 9, 87).

Dice, L. R. (1945). « Measures of the amount of ecologic association between species ». In : *Ecology* 26.3, p. 297–302 (cf. p. 156).

Diggle, P. J. (1983). *Statistical analysis of spatial point patterns*. London : Academic Press, p. 1–148 (cf. p. 22).

Dobzhansky, T. (1937). *Genetics and the Origin of Species*. New York : Columbia University Press (cf. p. 44).

Drakare, S., J. J. Lennon et H. Hillebrand (2006). « The imprint of the geographical, evolutionary and ecological context on species–area relationships ». In : *Ecology Letters* 9.2, p. 215–227 (cf. p. 11).

Ellerman, D. (2013). « An Introduction To Logical Entropy and Its Relation To Shannon Entropy ». In : *International Journal of Semantic Computing* 7.02, p. 121–145 (cf. p. 46).

Ellison, A. M. (2010). « Partitioning diversity ». In : *Ecology* 91.7, p. 1962–1963 (cf. p. 123).

Engen, S. (1974). « On species models frequency ». In : *Biometrika* 61.2, p. 263–270 (cf. p. 24).

— (1977). « Exponential and Logarithmic Species-Area Curves ». In : *The American Naturalist* 111.979, p. 591–594 (cf. p. 15, 24, 78).

Eren, M. I., A. Chao, W.-H. Hwang et R. K. Colwell (2012). « Estimating the Richness of a Population When the Maximum Number of Classes Is Fixed : A Nonparametric Solution to an Archaeological Problem ». English. In : *Plos One* 7.5 (cf. p. 30).

Escalas, A., T. Bouvier, M. A. Mouchet, F. Leprieur, C. Bouvier, M. Troussellier et D. Mouillot (2013). « A unifying quantitative framework for exploring the multiple facets of microbial biodiversity across diverse scales ». In : *Environmental Microbiology* 15.10, p. 2642–2657 (cf. p. 98).

Esty, W. W. (1983). « A Normal Limit Law for a Nonparametric Estimator of the Coverage of a Random Sample ». In : *The Annals of Statistics* 11.3, p. 905–912 (cf. p. 27).

Faddeev, D. K. (1956). « On the concept of entropy of a finite probabilistic scheme ». In : *Uspekhi Mat. Nauk* 1.67, p. 227–231 (cf. p. 170).

Faith, D. P. (1992). « Conservation evaluation and phylogenetic diversity ». In : *Biological Conservation* 61.1, p. 1–10 (cf. p. 96).

Faith, D. P. (2008). « Threatened species and the potential loss of phylogenetic diversity : conservation scenarios based on estimated extinction probabilities and phylogenetic risk analysis. » In : *Conservation biology : the journal of the Society for Conservation Biology* 22.6, p. 1461–70 (cf. p. 92).

Felsenstein, J. (1985). « Phylogenies and the Comparative Method ». In : *The American Naturalist* 25.1, p. 1–15 (cf. p. 86, 94).

— (2004). *Inferring Phylogenies*. 2nd Ed. Sinauer Associates (cf. p. 86).

Fisher, R. A., A. S. Corbet et C. B. Williams (1943). « The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population ». In : *Journal of Animal Ecology* 12, p. 42–58 (cf. p. 8, 13).

Gadagkar, R. (1989). « An undesirable property of Hill ' s diversity index N2 ». In : *Oecologia* 80, p. 140–141 (cf. p. 65).

Gárcia Martín, H. et N. Goldenfeld (2006). « On the origin and robustness of power-law species-area relationships in ecology ». In : *Proceedings of the National Academy of Sciences* 103.27, p. 10310–10315 (cf. p. 23).

Garnier, E., J. Cortez, G. Billès, M.-L. Navas, C. Roumet, M. Debussche, G. Laurent, A. Blanchard, D. Aubry, A. Bellmann, C. Neill et J.-P. Toussaint (2004). « Plant functional markers capture ecosystem properties during secondary succession ». In : *Ecology* 85.9, p. 2630–2637 (cf. p. 101).

Gaston, K. J. (2000). « Global patterns in biodiversity ». In : *Nature* 405.6783, p. 220–227 (cf. p. 5).

Gini, C. (1912). *Variabilità e mutabilità*. Bologna : C. Cuppini (cf. p. 46, 76).

Gleason, H. A. (1922). « On the Relation Between Species and Area ». In : *Ecology* 3.2, p. 158–162 (cf. p. 11).

Good, I. J. (1953). « The Population Frequency of Species and the Estimation of Population Parameters ». In : *Biometrika* 40.3/4, p. 237–264 (cf. p. 26, 46).

Gosselin, F. (2001). « Lorenz partial order : the best known logical framework to define evenness indices ». In : *Community Ecology* 2.2, p. 197–207 (cf. p. 76).

Gotelli, N. J. et R. K. Colwell (2001). « Quantifying biodiversity : procedures and pitfalls in the measurement and comparison of species richness ». In : *Ecology Letters* 4.4, p. 379–391 (cf. p. 10–12, 74).

Gower, J. C. (1966). « Some distance properties of latent root and vector methods used in multivariate analysis ». In : *Biometrika* 53.3, p. 325–338 (cf. p. 139, 159).

— (1971). « A General Coefficient of Similarity and Some of Its Properties ». In : *Biometrics* 27.4, p. 857–871 (cf. p. 85, 161).

Gower, J. C. et P. Legendre (1986). « Metric and Euclidean properties of dissimilarity coefficients ». In : *Journal of classification* 48, p. 5–48 (cf. p. 157, 161).

Grassberger, P. (1988). « Finite sample corrections to entropy and dimension estimates ». In : *Physics Letters A* 128.6-7, p. 369–373 (cf. p. 51, 69, 72).

— (2003). « Entropy Estimates from Insufficient Samplings ». In : *ArXiv Physics e-prints* 0307138.version 2 (cf. p. 51).

Gregorius, H.-R. (1991). « On the concept of effective number. » In : *Theoretical population biology* 40.2, p. 269–83 (cf. p. 61).

— (2009). « Distribution of variation over populations. » In : *Theory in Biosciences* 128.3, p. 179–89 (cf. p. 172).

— (2010). « Linking Diversity and Differentiation ». In : *Diversity* 2.3, p. 370–394 (cf. p. 64, 129, 169, 172).

— (2014). « Partitioning of diversity : the "within communities" component ». In : *Web Ecology* 14, p. 51–60 (cf. p. 61, 137, 169, 172).

Grilli, J., S. Azaele, J. R. Banavar et A. Maritan (2012). « Spatial aggregation and the species–area relationship across scales ». In : *Journal of Theoretical Biology* 313.0, p. 87–97 (cf. p. 25).

Guiasu, R. C. et S. Guiasu (2011). « The weighted quadratic index of biodiversity for pairs of species : a generalization of Rao's index ». In : *Natural Science* 3.9, p. 795–801 (cf. p. 140, 155).

— (2012). « The Weighted Gini-Simpson Index : Revitalizing an Old Index of Biodiversity ». In : *International Journal of Ecology* 2012.478728, p. 1–10 (cf. p. 155).

Haegeman, B., J. Hamelin, J. Moriarty, P. Neal, J. Dushoff et J. S. Weitz (2013). « Robust estimation of microbial diversity in theory and in practice ». In : *The ISME journal* 7.6, p. 1092–101 (cf. p. 70).

Hankin, R. K. S. (2007). « Introducing untb, an R Package For Simulating Ecological Drift Under the Unified Neutral Theory of Biodiversity ». In : *Journal of Statistical Software* 22.12, p. 1–15 (cf. p. 14).

Hardy, G. H., J. E. Littlewood et G. Pólya (1952). *Inequalities*. Cambridge University Press (cf. p. 106).

Hardy, O. J. et L. Jost (2008). « Interpreting and estimating measures of community phylogenetic structuring ». In : *Journal of Ecology* 96.5, p. 849–852 (cf. p. 139).

Hardy, O. J. et B. Senterre (2007). « Characterizing the phylogenetic structure of communities by an additive partitioning of phylogenetic diversity ». In : *Journal of Ecology* 95.3, p. 493–506 (cf. p. 139).

Harte, J., A. Kinzig et J. Green (1999). « Self-similarity in the distribution and abundance of species ». In : *Science* 284.5412, p. 334–336 (cf. p. 18).

Hausser, J. et K. Strimmer (2009). « Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks ». In : *Journal of Machine Learning Research* 10, p. 1469–1484 (cf. p. 52).

Havrda, J. et F. Charvát (1967). « Quantification method of classification processes. Concept of structural a-entropy ». In : *Kybernetika* 3.1, p. 30–35 (cf. p. 62).

Heller, R. et H. R. Siegismund (2009). « Relationship between three measures of genetic differentiation G(ST), D-EST and G'(ST) : how wrong have we been ? » In : *Molecular Ecology* 18.10, p. 2080–2083 (cf. p. 137).

Helmus, M. R., T. J. Bland, C. K. Williams et A. R. Ives (2007). « Phylogenetic measures of biodiversity ». In : *The American naturalist* 169.3, E68–83 (cf. p. 94).

Heltshe, J. F. et N. E. Forrester (1983). « Estimating species richness using the jackknife procedure ». In : *Biometrics* 39.1, p. 1–11 (cf. p. 31).

Hey, J. (2001). « The mind of the species problem ». In : *Trends in Ecology & Evolution* 16.7, p. 326–329 (cf. p. 2, 45).

Hill, M. O. (1973). « Diversity and Evenness : A Unifying Notation and Its Consequences ». In : *Ecology* 54.2, p. 427–432 (cf. p. 60, 65, 66, 74, 77).

Hoffmann, S. et A. Hoffmann (2008). « Is there a "true" diversity ? » In : *Ecological Economics* 65.2, p. 213–215 (cf. p. 64).

Holdridge, L., W. Grenke, W. Hatheway, T. Liang et J. Tosi (1971). *Forest environments in tropical life zones*. Oxford : Pergamon Press (cf. p. 41).

Horvitz, D. G. et D. J. Thompson (1952). « A generalization of sampling without replacement from a finite universe ». In : *Journal of the American Statistical Association* 47.260, p. 663–685 (cf. p. 49).

Hubbell, S. P. (2001). *The unified neutral theory of biodiversity and biogeography*. Princeton University Press (cf. p. 10, 14, 15, 24).

Hubbell, S. P., R. Condit et R. B. Foster (2005). *Barro Colorado Forest Census Plot Data* (cf. p. 6).

Hurlbert, S. H. (1971). « The Nonconcept of Species Diversity : A Critique and Alternative Parameters ». In : *Ecology* 52.4, p. 577–586 (cf. p. 53, 61, 92).

Hutcheson, K. (1970). « A Test for Comparing Diversities Based on the Shannon Formula ». In : *Journal of Theoretical Biology* 29, p. 151–154 (cf. p. 48).

Hutchinson, G. E. (1957). « Concluding remarks ». In : *Cold Spring Harbor Symposia on Quantitative Biology* 2.2, p. 415–427 (cf. p. 89).

Hwang, W.-H., C.-W. Lin et T.-J. Shen (2014). « Good-Turing frequency estimation in a finite population ». In : *Biometrical journal* 57.2, p. 321–339 (cf. p. 28).

Iknayan, K. J., M. W. Tingley, B. J. Furnas et S. R. Beissinger (2014). « Detecting diversity : emerging methods to estimate species diversity ». In : *Trends in ecology & evolution* 29.2, p. 97–106 (cf. p. 74).

Isaac, N. J., S. T. Turvey, B. Collen, C. Waterman et J. E. Baillie (2007). « Mammals on the EDGE : Conservation Priorities Based on Threat and Phylogeny ». In : *PLoS ONE* 2.3, e296 (cf. p. 92).

Izsák, J. et L. Papp (2000). « A link between ecological diversity indices and measures of biodiversity ». In : *Ecological Modelling* 130, p. 151–156 (cf. p. 98).

Izsák, J. et S. Pavoine (2011). « New concentration measures as kinds of the quadratic entropy ». In : *Ecological Indicators* 11.2, p. 540–544 (cf. p. 155).

— (2012). « Links between the species abundance distribution and the shape of the corresponding rank abundance curve ». In : *Ecological Indicators* 14.1, p. 1–6 (cf. p. 8, 14).

Jaccard, P. (1901). « Étude comparative de la distribution florale dans une portion des Alpes et du Jura ». In : *Bulletin de la Société Vaudoise des Sciences Naturelles* 37, p. 547–579 (cf. p. 156).

James, W. et C. Stein (1961). « Estimation with Quadratic Loss ». In : *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Sous la dir. de J. Neyman. T. Volume 1 : Berkeley, California : University of California Press, p. 361–379 (cf. p. 52).

Jensen, J. L. W. V. (1906). « Sur les fonctions convexes et les inégalités entre les valeurs moyennes ». French. In : *Acta Mathematica* 30.1, p. 175–193 (cf. p. 51).

Jones, C. G., J. H. Lawton et M. Shachak (1994). « Organisms as Ecosystem Engineers ». In : *Oikos,* 69.3, p. 373–386 (cf. p. 178).

Jones, D. et N. Matloff (1986). « Statistical Hypothesis Testing in Biology : A Contradiction in Terms ». In : *Journal of Economic Entomology* 79.5, p. 1156–1160 (cf. p. 151, 152).

Jost, L. (2006). « Entropy and diversity ». In : *Oikos* 113.2, p. 363–375 (cf. p. 62, 64, 127).

— (2007). « Partitioning diversity into independent alpha and beta components ». In : *Ecology* 88.10, p. 2427–2439 (cf. p. 17, 62, 64, 127–129, 133, 134, 137, 141).

— (2008). « GST and its relatives do not measure differentiation ». In : *Molecular Ecology* 17.18, p. 4015–4026 (cf. p. 137).

— (2009a). « D vs. GST : Response to Heller and Siegismund (2009) and Ryman and Leimar (2009) ». In : *Molecular Ecology* 18.10, p. 2088–2091 (cf. p. 137).

— (2009b). « Mismeasuring biological diversity : Response to Hoffmann and Hoffmann (2008) ». In : *Ecological Economics* 68, p. 925–928 (cf. p. 64).

— (2010a). « Independence of alpha and beta diversities ». In : *Ecology* 91.7, p. 1969–1994 (cf. p. 123).

— (2010b). « The Relation between Evenness and Diversity ». In : *Diversity* 2.2, p. 207–232 (cf. p. 4, 76, 80, 88).

Jurasinski, G., V. Retzer et C. Beierkuhnlein (2009). « Inventory, differentiation, and proportional diversity : a consistent terminology for quantifying species diversity ». English. In : *Oecologia* 159.1, p. 15–26 (cf. p. 128, 148).

Kader, G. D. et M. Perry (2007). « Variability for Categorical Variables ». In : *Journal of Statistics Education* 15.2 (cf. p. 91).

Kempton, R. A. et L. R. Taylor (1976). « Models and statistics for species diversity ». In : *Nature* 262.5571, p. 818–820 (cf. p. 17).

Kempton, R. A. et R. W. M. Wedderburn (1978). « A comparison of three measures of species diversity ». In : *Biometrics* 34, p. 25–37 (cf. p. 17).

Keylock, C. J. (2005). « Simpson diversity and the Shannon-Wiener index as special cases of a generalized entropy ». In : *Oikos* 109.1, p. 203–207 (cf. p. 62).

Kindt, R., P. Van Damme et A. J. Simons (2006). « Tree diversity in western Kenya : Using profiles to characterise richness and evenness ». In : *Biodiversity and Conservation* 15.4, p. 1253–1270 (cf. p. 66).

Kosman, E. (1996). « Difference and diversity of plant pathogen populations : a new approach for measuring ». In : *Phytopathology* 86.11, p. 1152–1155 (cf. p. 166).

— (2013). « Measuring diversity : from individuals to populations ». In : *European Journal of Plant Pathology* 138.3, p. 467–486 (cf. p. 165).

Kosman, E. et K. J. Leonard (2005). « Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. » In : *Molecular ecology* 14.2, p. 415–24 (cf. p. 165).

— (2007). « Conceptual analysis of methods applied to assessment of diversity within and distance between populations with asexual or mixed mode of reproduction. » In : *The New phytologist* 174.3, p. 683–96 (cf. p. 166).

Krishnamani, R., A. Kumar et J. Harte (2004). « Estimating species richness at large spatial scales using data from small discrete plots ». In : *Ecography* 27.5, p. 637–642 (cf. p. 33).

Kullback, S. et R. A. Leibler (1951). « On Information and Sufficiency ». In : *The Annals of Mathematical Statistics* 22.1, p. 79–86 (cf. p. 59, 75).

Lamberti, P. W. et A. P. Majtey (2003). « Non-logarithmic Jensen–Shannon divergence ». In : *Physica A : Statistical Mechanics and its Applications* 329.1–2, p. 81–90 (cf. p. 134).

Lande, R. (1996). « Statistics and partitioning of species diversity, and similarity among multiple communities ». In : *Oikos* 76, p. 5–13 (cf. p. 46, 127, 133).

Lande, R., P. J. DeVries et T. R. Walla (2000). « When species accumulation curves intersect : implications for ranking diversity using small samples ». In : *Oikos* 89.3, p. 601–605 (cf. p. 66).

Lavorel, S., K. Grigulis, S. McIntyre, N. S. G. Williams, D. Garden, J. Dorrough, S. Berman, F. Quétier, A. Thébault et A. Bonis (2008). « Assessing functional diversity in the field - Methodology matters ! » In : *Functional Ecology* 22, p. 134–147 (cf. p. 101).

Legendre, P. et M. J. Anderson (1999). « Distance-based redundancy analysis : testing multispecies responses in multifactorial ecological experiments ». In : *Ecological Monographs* 69.1, p. 1–24 (cf. p. 126, 165).

Legendre, P. et E. Gallagher (2001). « Ecologically meaningful transformations for ordination of species data ». In : *Oecologia* 129.2, p. 271–280 (cf. p. 126, 165).

Legendre, P. et L. Legendre (2012). *Numerical Ecology*. 3rd Ed. Elsevier (cf. p. 86, 161).

Legendre, P., D. Borcard et P. R. Peres-Neto (2005). « Analyzing Beta Diversity : Partitioning the Spatial Variation of Community Composition Data ». In : *Ecological Monographs* 75.4, p. 435–450 (cf. p. 126, 165).

Leinster, T. (2013). « The Magnitude of Metric Spaces ». In : *Documenta Mathematica* 18, p. 857–905 (cf. p. 110).

Leinster, T. et C. Cobbold (2012). « Measuring diversity : the importance of species similarity ». In : *Ecology* 93.3, p. 477–489 (cf. p. 54, 66, 105).

Lewontin, R. (1972). « The apportionment of human diversity ». In : *Evolutionary biology* 6, p. 381–398 (cf. p. 130).

Lin, J. (1991). « Divergence Measures Based on the Shannon Entropy ». In : *IEEE Transactions on Information Theory* 37.1, p. 145–151 (cf. p. 130).

Lineweaver, H. et D. Burk (1934). « The Determination of Enzyme Dissociation Constants ». In : *Journal of the American Chemical Society* 56.3, p. 658–666 (cf. p. 40).

Lingoes, J. C. (1971). « Some boundary conditions for a monotone analysis of symmetric matrices ». In : *Psychometrika* 36.2, p. 195–203 (cf. p. 84).

Liu, C., R. J. Whittaker, K. Ma et J. R. Malcolm (2006). « Unifying and distinguishing diversity ordering methods for comparing communities ». In : *Population Ecology* 49.2, p. 89–100 (cf. p. 66).

Liu, D., D. Wang, Y. Wang, J. Wu, V. P. Singh, X. Zeng, L. Wang, Y. Chen, X. Chen, L. Zhang et S. Gu (2015). « Entropy of hydrological systems under small samples : Uncertainty and variability ». In : *Journal of Hydrology* in press (cf. p. 53).

Lloyd, M. et R. J. Ghelardi (1964). « A Table for Calculating the 'Equitability' Component of Species Diversity ». In : *Journal of Animal Ecology* 33.2, p. 217–225 (cf. p. 76, 78).

Loh, J., R. E. Green, T. Ricketts, J. Lamoreux, M. Jenkins, V. Kapos et J. Randers (2005). « The Living Planet Index : using species population time series to track trends in biodiversity. » In : *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 360.1454, p. 289–295 (cf. p. 179).

Lonsdale, W. M. (1999). « Global Patterns of Plant Invasions and the Concept of Invasibility ». In : *Ecology* 80.5, p. 1522–1536 (cf. p. 10).

Loreau, M. (2005). « Discours de clôture ». In : *Actes de la Conférence internationale Biodiversité Science et Gouvernance*. Sous la dir. de R. Barbault et J.-P. Le Duc. Paris, France. : IRD Editions, p. 254–256 (cf. p. 1).

Lorenz, M. O. (1905). « Methods of Measuring the concentration of Wealth ». In : *Quarterly Publications of the American Statistical Association* 9, p. 209–219 (cf. p. 76).

Ludovisi, A. et M. I. Taticchi (2006). « Investigating beta diversity by Kullback-Leibler information measures ». In : *Ecological Modelling* 192.1-2, p. 299–313 (cf. p. 131).

Maasoumi, E. (1993). « A compendium to information theory in economics and econometrics ». In : *Econometric Reviews* 12.2, p. 137–181 (cf. p. 57, 75).

MacArthur, R. H. (1955). « Fluctuations of Animal Populations and a Measure of Community Stability ». In : *Ecology* 36.3, p. 533–536 (cf. p. 60).

— (1957). « On the Relative Abundance of Bird Species ». In : *Proceedings of the National Academy of Sciences of the United States of America* 43.3, p. 293–295 (cf. p. 8).

— (1965). « Patterns of species diversity ». In : *Biological Reviews* 40.4, p. 510–533 (cf. p. 61).

Magurran, A. E. (1988). *Ecological diversity and its measurement*. Princeton, NJ : Princeton University Press (cf. p. 8).

Mao, C. X. (2007). « Estimating Species Accumulation Curves and Diversity Indices ». In : *Statistica Sinica* 17, p. 761–774 (cf. p. 54, 68).

Mao, C. X. et R. K. Colwell (2005). « Estimation of Species Richness : Mixture Models, the Role of Rare Species, and Inferential Challenges ». In : *Ecology* 86.5, p. 1143–1153 (cf. p. 29).

Marcon, É. (2015). « Practical Estimation of Diversity from Abundance Data ». In : *HAL* 01212435.version 2 (cf. p. 35, 70–72).

Marcon, É. et B. Hérault (2015a). « Decomposing Phylodiversity ». In : *Methods in Ecology and Evolution* 6.3, p. 333–339 (cf. p. 144).

— (2015b). « entropart, an R Package to Measure and Partition Diversity ». In : *Journal of Statistical Software* 67.8, p. 1–26 (cf. p. 2).

Marcon, É., B. Hérault, C. Baraloto et G. Lang (2012). « The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity ». In : *Oikos* 121.4, p. 516–522 (cf. p. 60, 71, 128, 130, 141, 153).

Marcon, É., I. Scotti, B. Hérault, V. Rossi et G. Lang (2014a). « Generalization of the Partitioning of Shannon Diversity ». In : *Plos One* 9.3, e90289 (cf. p. 64, 69, 71, 98, 103, 127, 128, 130, 133, 134, 136, 141, 151, 152).

Marcon, É., Z. Zhang et B. Hérault (2014b). « The Decomposition of Similarity-Based Diversity and its Bias Correction ». In : *HAL* 00989454.version 3 (cf. p. 110, 141, 146).

Mason, N. W. H., K. MacGillivray, J. B. Steel et J. B. Wilson (2003). « An index of functional diversity ». In : *Journal of Vegetation Science* 14.4, p. 571–578 (cf. p. 91).

Mason, N. W. H., D. Mouillot, W. G. Lee et J. B. Wilson (2005). « Functional richness, functional evenness and functional divergence : the primary components of functional diversity ». In : *Oikos* 111.1, p. 112–118 (cf. p. 88).

Matérn, B. (1960). « Spatial variation ». In : *Meddelanden från Statens Skogsforskningsinstitut* 49.5, p. 1–144 (cf. p. 23).

May, R. M. (1975). « Patterns of species abundance and diversity ». In : *Ecology and Evolution of Communities*. Sous la dir. de M. L. Cody et J. M. Diamond. Harvard University Press, p. 81–120 (cf. p. 10).

— (1990). « Taxonomy as Destiny ». In : *Nature* 347, p. 129–130 (cf. p. 83).

— (2011). « Why Worry about How Many Species and Their Loss ? » In : *PLoS Biol* 9.8, e1001130 (cf. p. 43).

May, R. M. et M. P. H. Stumpf (2000). « Species-Area Relations in Tropical Forests ». In : *Science* 290.5499, p. 2084–2086 (cf. p. 10).

Mayden, R. L. (1997). « A hierarchy of species concepts : the denouement in the saga of the species problem ». In : *Species. The units of biodiversity*. Sous la dir. de M. F. Claridge, H. A. Dawah et M. R. Wilson. London : Chapman et Hall. Chap. 19, p. 381–424 (cf. p. 44).

Mayr, E. (1942). *Systematics and the origin of species from the viewpoint of a zoologist*. New York : Columbia University Press (cf. p. 44).

— (1970). *Populations, species, and evolution*. Cambridge, Massachussets : Belknap Press of Harvard University Press (cf. p. 45).

McGill, B. J. (2003). « A test of the unified neutral theory of biodiversity ». In : *Nature* 422.6934, p. 881–885 (cf. p. 24).

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 et E. P. White (2007). « Species abundance distributions : moving beyond single prediction theories to integration within an ecological framework ». In : *Ecology Letters* 10.10, p. 995–1015 (cf. p. 8).

Mcintosh, R. P. (1967). « An Index of Diversity and the Relation of Certain Concepts to Diversity ». In : *Ecology* 48.3, p. 392–404 (cf. p. 3).

Meine, C., M. Soulé et R. F. Noss (2006). « "A mission-driven discipline" : The growth of conservation biology ». In : *Conservation Biology* 20.3, p. 631–651 (cf. p. 1).

Mendes, R. S., L. R. Evangelista, S. M. Thomaz, A. A. Agostinho et L. C. Gomes (2008). « A unified index to measure ecological diversity and species rarity ». In : *Ecography* 31.4, p. 450–456 (cf. p. 62, 79).

Mérigot, B. et J.-C. Gaertner (2011). « Incorporation of phylogeny in biological diversity measurement : Drawbacks of extensively used indices, and advantages of quadratic entropy ». In : *BioEssays* 33.11, p. 819–822 (cf. p. 91).

Michaelis, L. et M. L. Menten (1913). « Die Kinetik der Invertinwirkung ». In : *Biochemische Zeitschrift* 49, p. 333–369 (cf. p. 39).

Miller, G. A. (1955). « Note on the bias of information estimates ». In : *Information Theory in Psychology : Problems and Methods*. Sous la dir. de H. Quastler. Free Press, p. 95–100 (cf. p. 49).

Mooers, A. Ø. et S. B. Heard (1997). « Inferring Evolutionary Process from Phylogenetic Tree Shape ». In : *The Quarterly Review of Biology* 72.1, p. 31–54 (cf. p. 87).

Mora, C., D. P. Tittensor, S. Adl, A. G. B. Simpson et B. Worm (2011). « How Many Species Are There on Earth and in the Ocean ? » In : *PLoS Biol* 9.8, e1001127 (cf. p. 43, 44).

Moreno, C. E. et P. Rodríguez (2010). « A consistent terminology for quantifying species diversity ? » In : *Oecologia* 163.2, p. 279–82 (cf. p. 5).

Motomura, I. (1932). « On the statistical treatment of communities ». Japanese. In : *Zoological Magazine* 44, p. 379–383 (cf. p. 8).

Motyka, J. (1947). *O celach i metodach badan geobotanicznych. Sur les buts et les méthodes des recherches géobotaniques*. Lublin, Poland : Nakladem universytftu Marii Curie-Sklodowskiej (cf. p. 161).

Mouchet, M., F. Guilhaumon, S. Villéger, N. W. H. Mason, J.-A. Tomasini et D. Mouillot (2008). « Towards a consensus for calculating dendrogram-based functional diversity indices ». In : *Oikos* 117, p. 794–800 (cf. p. 86, 90).

Mouchet, M. A. et D. Mouillot (2011). « Decomposing phylogenetic entropy into α, β and γ components ». In : *Biology Letters* 7.2, p. 205–209 (cf. p. 144).

Mouillot, D. et A. Leprêtre (1999). « A comparison of species diversity estimators ». In : *Researches on Population Ecology* 41.2, p. 203–215 (cf. p. 9).

Mouillot, D., W. H. N. Mason, O. Dumay et J. B. Wilson (2005a). « Functional regularity : a neglected aspect of functional diversity ». In : *Oecologia* 142.3, p. 353–359 (cf. p. 90).

Mouillot, D., W. Stubbs, M. Faure, O. Dumay, J. A. Tomasini, J. B. Wilson et T. D. Chi (2005b). « Niche overlap estimates based on quantitative functional traits : a new family of non-parametric indices ». In : *Oecologia* 145.3, p. 345–353 (cf. p. 4).

Nei, M. (1972). « Genetic Distance between Populations ». In : *The American Naturalist* 106.949, p. 283–292 (cf. p. 109).

— (1973). « Analysis of Gene Diversity in Subdivided Populations ». In : *Proceedings of the National Academy of Sciences of the United States of America* 70.12, p. 3321–3323 (cf. p. 137, 148).

Nekola, J. C. et P. S. White (1999). « The Distance Decay of Similarity in Biogeography and Ecology ». In : *Journal of Biogeography* 26, p. 867–878 (cf. p. 125).

Neyman, J. et E. L. Scott (1958). « Statistical Approach to Problems of Cosmology ». In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 20.1, p. 1–43 (cf. p. 22).

Norris, J. L. et K. H. Pollock (1998). « Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species ». In : *Environmental and Ecological Statistics* 5.4, p. 391–402 (cf. p. 37).

Ochiai, A. (1957). « Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions ». In : *Bulletin of the Japanese Society of Scientific Fisheries* 22.9, p. 526–530 (cf. p. 157).

Odum, E. P. (1950). « Bird Populations of the Highlands (North Carolina) Plateau in Relation to Plant Succession and Avian Invasion ». In : *Ecology* 31.4, p. 587–605 (cf. p. 161).

O'Hara, R. B. (2005). « Species richness estimators : How many species can dance on the head of a pin ? » In : *Journal of Animal Ecology* 74, p. 375–386 (cf. p. 28).

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 et H. Wagner (2012). *vegan : Community Ecology Package* (cf. p. 12).

Paine, R. T. (1969). « A note on trophic complexity and community stability ». In : *American naturalist* 103.929, p. 91–93 (cf. p. 178).

Pallmann, P., F. Schaarschmidt, L. A. Hothorn, C. Fischer, H. Nacke, K. U. Priesnitz et N. J. Schork (2012). « Assessing group differences in biodiversity by simultaneously testing a user-defined selection of diversity indices ». In : *Molecular Ecology Resources* 12.6, p. 1068–1078 (cf. p. 66).

Patil, G. P. et C. Taillie (1982). « Diversity as a concept and its measurement ». In : *Journal of the American Statistical Association* 77.379, p. 548–561 (cf. p. 57, 58, 60, 62, 66).

Pavoine, S. et M. B. Bonsall (2009). « Biological diversity : Distinct distributions can lead to the maximization of Rao's quadratic entropy ». In : *Theoretical Population Biology* 75.2-3, p. 153–163 (cf. p. 101, 102, 105).

— (2011). « Measuring biodiversity to explain community assembly : a unified approach ». In : *Biological Reviews* 86.4, p. 792–812 (cf. p. 4, 83, 87, 90).

Pavoine, S. et J. Izsák (2014). « New biodiversity measure that includes consistent interspecific and intraspecific components ». In : *Methods in Ecology and Evolution* 5.2, p. 165–172 (cf. p. 115).

Pavoine, S. et C. Ricotta (2014). « Functional and phylogenetic similarity among communities ». In : *Methods in Ecology and Evolution* 5.7, p. 666–675 (cf. p. 114, 162).

Pavoine, S., A.-B. Dufour et D. Chessel (2004). « From dissimilarities among species to dissimilarities among communities : a double principal coordinate analysis ». In : *Journal of Theoretical Biology* 228.4, p. 523–537 (cf. p. 126, 138, 139).

Pavoine, S., S. Ollier et A.-B. Dufour (2005a). « Is the originality of a species measurable ? » In : *Ecology Letters* 8, p. 579–586 (cf. p. 86, 98–100, 113).

Pavoine, S., S. Ollier et D. Pontier (2005b). « Measuring, diversity from dissimilarities with Rao's quadratic entropy : Are any dissimilarities suitable ? » In : *Theoretical Population Biology* 67.4, p. 231–239 (cf. p. 98).

Pavoine, S., M. S. Love et M. B. Bonsall (2009a). « Hierarchical partitioning of evolutionary and ecological pat-

terns in the organization of phylogenetically-structured species assemblages : Application to rockfish (genus : Sebastes) in the Southern California Bight ». In : *Ecology Letters* 12.9, p. 898–908 (cf. p. 101).

Pavoine, S., J. Vallet, A.-B. Dufour, S. Gachet et H. Daniel (2009b). « On the challenge of treating various types of variables : application for improving the measurement of functional diversity ». In : *Oikos* 118.3, p. 391–402 (cf. p. 85).

Pavoine, S., É. Marcon et C. Ricotta (2015). « "Equivalent numbers" for species, phylogenetic, or functional diversity in a nested hierarchy of multiple scales ». In : *Methods in Ecology and Evolution* in rev (cf. p. 129).

Pearson, K. (1900). « On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling ». In : *Philosophical Magazine Series 5* 50.302, p. 157–175 (cf. p. 75).

— (1901). « On lines and planes of closest fit to systems of points in space ». In : *Philosophical Magazine* 2, p. 559–572 (cf. p. 163).

Peet, R. K. (1974). « The measurement of species diversity ». In : *Annual review of ecology and systematics* 5, p. 285–307 (cf. p. 9, 77).

Pélissier, R. et P. Couteron (2007). « An operational, additive framework for species diversity partitioning and beta-diversity analysis ». In : *Journal of Ecology* 95.2, p. 294–300 (cf. p. 150).

Pélissier, R., P. Couteron, S. Dray et D. Sabatier (2003). « Consistency between ordination techniques and diversity measurements : Two strategies for species occurrence data ». In : *Ecology* 84.1, p. 242–251 (cf. p. 150).

Pereira, H. M., S. Ferrier, M. Walters, G. N. Geller, R. H. G. Jongman, R. J. Scholes, M. W. Bruford, N. Brummitt, S. H. M. Butchart, A. C. Cardoso, N. C. Coops, E. Dulloo, D. P. Faith, J. Freyhof, R. D. Gregory, C. Heip, R. Höft, G. Hurtt, W. Jetz, D. S. Karp, M. A. McGeoch, D. Obura, Y. Onoda, N. Pettorelli, B. Reyers, R. Sayre, J. P. W. Scharlemann, S. N. Stuart, E. Turak, M. Walpole et M. Wegmann (2013). « Essential biodiversity variables ». In : *Science* 339, p. 277–278 (cf. p. 179).

Pernès, J., éd. (1984). *Gestion des ressources génétiques des plantes. Tome 2 : Manuel*. Paris : Agence de Coopération culturelle et technique (cf. p. 45).

Petchey, O. L. et K. J. Gaston (2002). « Functional diversity (FD), species richness and community composition ». In : *Ecology Letters* 5, p. 402–411 (cf. p. 90, 96).

Picard, N., M Karembe et P. Birnbaum (2004). « Species-area curve and spatial pattern ». In : *Ecoscience* 11.1, p. 45–54 (cf. p. 22, 23).

Pielou, E. C. (1966a). « Species-diversity and pattern-diversity in the study of ecological succession ». In : *Journal of Theoretical Biology* 10.2, p. 370–383 (cf. p. 50).

— (1966b). « The measurement of diversity in different types of biological collections ». In : *Journal of Theoretical Biology* 13.C, p. 131–144 (cf. p. 48, 78).

— (1969). *An Introduction to Mathematical Ecology*. New York : Wiley (cf. p. 13).

— (1975). *Ecological Diversity*. New York : Wiley (cf. p. 57, 78, 83).

Plotkin, J. B., M. D. Potts, D. W. Yu, S. Bunyavejchewin, R. Condit, R. B. Foster, S. P. Hubbell, J. LaFrankie, N Manokaran, H.-S. Lee, R. Sukumar, M. A. Nowak et P. S. Ashton (2000a). « Predicting species diversity in tropical forests ». In : *Proceedings of the National Academy of Sciences of the United States of America* 97.20, p. 10850–10854 (cf. p. 18–20, 24).

Plotkin, J. B., M. D. Potts, N. Leslie, N Manokaran, J. V. LaFrankie et P. S. Ashton (2000b). « Species-area curves, spatial aggregation, and habitat specialization in tropical forests ». In : *Journal of Theoretical Biology* 207.1, p. 81–99 (cf. p. 22).

Podani, J. (1999). « Extending Gower's General Coefficient of Similarity to Ordinal Characters ». In : *Taxon* 48.2, p. 331–340 (cf. p. 85).

Podani, J. et D. Schmera (2006). « On dendrogram-based measures of functional diversity ». In : *Oikos* 115.1, p. 179–185 (cf. p. 86, 113).

— (2007). « How should a dendrogram-based measure of functional diversity function ? A rejoinder to Petchey and Gaston ». In : *Oikos* 116.8, p. 1427–1430 (cf. p. 86, 113).

Presley, S. J., S. M. Scheiner et M. R. Willig (2014). « Evaluation of an integrated framework for biodiversity with a new metric for functional dispersion. » In : *PloS one* 9.8, e105818 (cf. p. 119).

Preston, F. W. (1948). « The Commonness, And Rarity, of Species ». In : *Ecology* 29.3, p. 254–283 (cf. p. 8, 37).

— (1960). « Time and Space and the Variation of Species ». In : *Ecology* 41.4, p. 611–627 (cf. p. 24).

— (1962). « The Canonical Distribution of Commonness and Rarity : Part I ». In : *Ecology* 43.2, p. 185–215 (cf. p. 10, 20).

Purvis, A. et A. Hector (2000). « Getting the measure of biodiversity. » In : *Nature* 405.6783, p. 212–9 (cf. p. 2).

R Core Team (2015). *R : A Language and Environment for Statistical Computing*. Vienna, Austria : R Foundation for Statistical Computing (cf. p. 2).

Raaijmakers, J. G. W. (1987). « Statistical Analysis of the Michaelis-Menten Equation ». In : *Biometrics* 43.4, p. 793–803 (cf. p. 39, 41).

Rao, C. R. (1982). « Diversity and dissimilarity coefficients : a unified approach ». In : *Theoretical Population Biology* 21, p. 24–43 (cf. p. 83, 98, 139).

Rao, C. R. et T. K. Nayak (1985). « Cross entropy, dissimilarity measures, and characterizations of quadratic entropy ». In : *Information Theory, IEEE Transactions on* 31.5, p. 589–593 (cf. p. 130).

Rényi, A. (1961). « On Measures of Entropy and Information ». In : *4th Berkeley Symposium on Mathematical Statistics and Probability*. Sous la dir. de J. Neyman. T. 1. Berkeley, USA : University of California Press, p. 547–561 (cf. p. 60, 74, 170).

Richard-Hansen, C., G. Jaouen, T. Denis, O. Brunaux, É. Marcon et S. Guitet (2015). « Landscape patterns influence communities of medium- to large-bodied vertebrate in undisturbed terra firme forests of French Guiana ». In : *Journal of Tropical Ecology* 31.5, p. 423–436 (cf. p. 153).

Richards, R. A. (2010). *The Species Problem. A Philosophical Analysis*. Cambridge : Cambridge University Press (cf. p. 44).

Ricotta, C et G Bacaro (2010). « On plot-to-plot dissimilarity measures based on species functional traits ». In : *Community Ecology* 11.1, p. 113–119 (cf. p. 162).

Ricotta, C. (2003). « On parametric evenness measures ». In : *Journal of Theoretical Biology* 222.2, p. 189–197 (cf. p. 78).

— (2004a). « A parametric diversity measure combining the relative abundances and taxonomic distinctiveness of species ». In : *Diversity and Distributions* 10, p. 143–146 (cf. p. 92).

Ricotta, C. (2004b). « A recipe for unconventional evenness measures ». In : *Acta Biotheoretica* 52.2, p. 95–104 (cf. p. 76–78).

— (2005a). « A note on functional diversity measures ». In : *Basic and Applied Ecology* 6.5, p. 479–486 (cf. p. 87).

— (2005b). « Additive partitioning of Rao's quadratic diversity : a hierarchical approach ». In : *Ecological Modelling* 183.4, p. 365–371 (cf. p. 102).

— (2005c). « On hierarchical diversity decomposition ». In : *Journal of Vegetation Science* 16.2, p. 223–226 (cf. p. 128).

— (2005d). « On parametric diversity indices in ecology : A historical note ». In : *Community Ecology* 6.2, p. 241–244 (cf. p. 62).

— (2005e). « Through the jungle of biological diversity ». In : *Acta Biotheoretica* 53.1, p. 29–38 (cf. p. 2).

— (2006). « Strong requirements for weak diversities ». In : *Diversity and Distributions* 12.2, p. 218–219 (cf. p. 93).

— (2007). « A semantic taxonomy for diversity measures ». In : *Acta Biotheoretica* 55.1, p. 23–33 (cf. p. 4, 83, 87).

— (2009). « Computing parametric beta diversity with unequal plot weights : A solution based on resampling methods ». In : *Theoretical Ecology* 2.1, p. 13–17 (cf. p. 130).

Ricotta, C. et G. C. Avena (2003a). « An information-theoretical measure of beta-diversity ». In : *Plant Biosystems* 137.1, p. 57–61 (cf. p. 131).

— (2003b). « An information-theoretical measure of taxonomic diversity ». In : *Acta biotheoretica* 25.51, p. 35–41 (cf. p. 9).

Ricotta, C. et S. Burrascano (2008). « Beta diversity for functional ecology ». In : *Preslia* 80.1, p. 61–72 (cf. p. 162).

— (2009). « Testing for differences in beta diversity with asymmetric dissimilarities ». In : *Ecological Indicators* 9.4, p. 719–724 (cf. p. 140, 168).

Ricotta, C. et M. Marignani (2007). « Computing β-diversity with Rao's quadratic entropy : A change of perspective ». In : *Diversity and Distributions* 13.2, p. 237–241 (cf. p. 128).

Ricotta, C. et M. Moretti (2011). « CWM and Rao's quadratic diversity : A unified framework for functional ecology ». In : *Oecologia* 167, p. 181–188 (cf. p. 101).

Ricotta, C. et L. Szeidl (2006). « 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, p. 237–243 (cf. p. 108, 146).

— (2009). « Diversity partitioning of Rao's quadratic entropy ». In : *Theoretical Population Biology* 76.4, p. 299–302 (cf. p. 104, 140).

Ricotta, C., E. De Zuliani, A. Pacini et G. C. Avena (2001). « On the mutual relatedness of evenness measures ». In : *Community Ecology* 2.1, p. 51–56 (cf. p. 78).

Ricotta, C., G. Bacaro, M. Marignani, S. Godefroid et S. Mazzoleni (2012). « Computing diversity from dated phylogenies and taxonomic hierarchies : Does it make a difference to the conclusions ? » In : *Oecologia* 170.2, p. 501–506 (cf. p. 85).

Ricotta, C., G. Bacaro, M. Caccianiga, B. E. Cerabolini et M. Moretti (2015). « A classical measure of phylogenetic dissimilarity and its relationship with beta diversity ». In : *Basic and Applied Ecology* 16.1, p. 10–18 (cf. p. 139).

Roberge, J.-M. et P. Angelstam (2004). « Usefulness of the Umbrella Species Concept ». In : *Conservation Biology* 18.1, p. 76–85 (cf. p. 178).

Routledge, R. D. (1979). « Diversity indices : Which ones are admissible ? » In : *Journal of Theoretical Biology* 76.4, p. 503–515 (cf. p. 129).

— (1983). « Evenness indices : Are any admissible ? » In : *Oikos* 40.1, p. 149–151 (cf. p. 76).

Runnegar, B. (1987). « Rates and Modes of Evolution in the Mollusca ». In : *Rates of Evolution*. Sous la dir. de M. Campbell et M. F. Day. London : Allen & Unwin, p. 39–60 (cf. p. 4).

Ryman, N et O Leimar (2009). « G(ST) is still a useful measure of genetic differentiation - a comment on Jost's D ». English. In : *Molecular Ecology* 18.10, p. 2084–2087 (cf. p. 137).

Scheiner, S. M. (2003). « Six types of species-area curves ». In : *Global Ecology and Biogeography* 12.6, p. 441–447 (cf. p. 6).

Scheiner, S. M. (2012). « A metric of biodiversity that integrates abundance, phylogeny, and function ». In : *Oikos* 121.8, p. 1191–1202 (cf. p. 118).

Schleuter, D., M. Daufresne, F. Massol et C. Argillier (2010). « A user's guide to functional diversity indices ». In : *Ecological Monographs* 80.3, p. 469–484 (cf. p. 88, 89).

Schmera, D., T. Erős et J. Podani (2009). « A measure for assessing functional diversity in ecological communities ». In : *Aquatic Ecology* 43.1, p. 157–167 (cf. p. 90).

Scholes, R. J. et R. Biggs (2005). « A Biodiversity Intactness Index ». In : *Nature* 434.7029, p. 45–49 (cf. p. 179).

Schulte, R. P. O., E. A. Lantinga et M. J. Hawkins (2005). « A new family of Fisher-curves estimates Fisher's alpha more accurately ». In : *Journal of Theoretical Biology* 232.3, p. 305–313 (cf. p. 15, 16).

Schürmann, T. (2004). « Bias analysis in entropy estimation ». In : *Journal of Physics A : Mathematical and General* 37.27, p. L295–L301 (cf. p. 51).

Shannon, C. E. (1948). « A Mathematical Theory of Communication ». In : *The Bell System Technical Journal* 27, p. 379–423, 623–656 (cf. p. 47, 57).

Shannon, C. E. et W. Weaver (1963). *The Mathematical Theory of Communication*. University of Illinois Press (cf. p. 47, 57).

Sheldon, A. L. (1969). « Equitability indices : dependence on the species count ». In : *Ecology* 50.3, p. 466–467 (cf. p. 77, 78, 80).

Shen, G., T. Wiegand, X. Mi et F. He (2013). « Quantifying spatial phylogenetic structures of fully stem-mapped plant communities ». In : *Methods in Ecology and Evolution* 4.12, p. 1132–1141 (cf. p. 147).

Shen, T. J., A. Chao et C.-F. Lin (2003). « Predicting the number of new species in a further taxonomic sampling ». In : *Ecology* 84, p. 798–804 (cf. p. 35).

Shimatani, K. (2001a). « Multivariate point processes and spatial variation of species diversity ». In : *Forest Ecology and Management* 142.1-3, p. 215–229 (cf. p. 102).

— (2001b). « On the measurement of species diversity incorporating species differences ». In : *Oikos* 93.1, p. 135–147 (cf. p. 98).

Simpson, E. H. (1949). « Measurement of diversity ». In : *Nature* 163.4148, p. 688 (cf. p. 45).

Simpson, G. G. (1943). « Mammals and the Nature of Continents ». In : *American Journal of Science* 241, p. 1–31 (cf. p. 160).

Slik, J. W. F. et al. (2015). « An estimate of the number of tropical tree species ». In : *Proceedings of the National Academy of Sciences* in press (cf. p. 25, 43).

Smith, B. et J. B. Wilson (1996). « A consumer's guide to evenness indices ». In : *Oikos* 76.1, p. 70–82 (cf. p. 76, 79, 80).

Smith, E. P. et G. V. Belle (1984). « Nonparametric Estimation of Species Richness ». In : *Biometrics* 40.1, p. 119–129 (cf. p. 31).

Smith, W. et J. F. Grassle (1977). « Sampling properties of a family of diversity measures ». In : *Biometrics* 33.2, p. 283–292 (cf. p. 54).

Soberón M., J. et J. Llorente B. (1993). « The Use of Species Accumulation Functions for the Prediction of Species Richness ». In : *Conservation Biology* 7.3, p. 480–488 (cf. p. 41).

Sokal, R. R. et C. D. Michener (1958). « A statistical method for evaluating systematic relationships ». In : *The University of Kansas Science Bulletin* 38.22, p. 1409–1438 (cf. p. 86).

Sokal, R. R. et F. J. Rohlf (1962). « The Comparison of Dendrograms by Objective Methods ». In : *Taxon* 11.2, p. 33–40 (cf. p. 86).

Sokal, R. R. et P. H. A. Sneath (1963). *Principles of numerical taxonomy*. San Francisco : W. H. Freeman & Co. (cf. p. 157).

Solow, A. R. et S. Polasky (1994). « Measuring biological diversity ». In : *Environmental and Ecological Statistics* 1.2, p. 95–103 (cf. p. 87).

Sørensen, T. (1948). *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*. T. 5. Biologiske Skrifter 4. København : I kommission hos E. Munksgaard, p. 1–34 (cf. p. 156).

Speth, J., M. Holdgate et M. Tolba (1992). « Foreword ». In : *Global Biodiversity Strategy*. Sous la dir. de K. Courrier. Washington, D.C. : WRI, IUCN, UNEP, p. v–vi (cf. p. 1).

Steege, H. ter, 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. 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. d. A. 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 C, T. R. 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. d. A. 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. Feeley, G. Lopez-Gonzalez et M. R. Silman (2013). « Hyperdominance in the Amazonian Tree Flora ». In : *Science* 342.6156, p. 1243092 (cf. p. 25, 43).

Steel, G. D. (1953). « Relation Between Poisson and Multinomial Distributions ». In : *Biometrics Unit Technical Reports*. Cornell University. Biometrics Unit. Chap. BU-39-M, p. 1–2 (cf. p. 13).

Stegen, J. C. et A. H. Hurlbert (2011). « Inferring ecological processes from taxonomic, phylogenetic and functional trait β-diversity. » In : *PloS one* 6.6, e20906 (cf. p. 9).

Stirling, A. (2007). « A general framework for analysing diversity in science, technology and society. » In : *Journal of the Royal Society, Interface / the Royal Society* 4.15, p. 707–719 (cf. p. 4).

Stoddart, J. A. (1983). « A genotypic diversity measure ». In : *Journal of Heredity* 74, p. 489–490 (cf. p. 61).

Stoyan, D. et H. Stoyan (1994). *Fractals, Random Shapes and Point Fields : Methods of Geometrical Statistics*. Chichester : John Wiley & Sons, 406 p. (Cf. p. 22).

Studeny, A. C., S. T. Buckland, J. B. Illian, A. Johnston et A. E. Magurran (2011). « Goodness-of-fit measures of evenness : a new tool for exploring changes in community structure ». In : *Ecosphere* 2.2, art.15 (cf. p. 75).

Swenson, N. G. et B. J. Enquist (2009). « Opposing assembly mechanisms in a Neotropical dry forest : Implications for phylogenetic and functional community ecology ». In : *Ecology* 90.8, p. 2161–2170 (cf. p. 87).

Taillie, C. (1979). « Species equitability : a comparative approach ». In : *Ecological Diversity in Theory and Practice*. Sous la dir. de J. F. Grassle, G. P. Patil, W. K. Smith et C. Taillie. Fairland, MD : International Cooperative Publishing House, p. 51–62 (cf. p. 77).

Theil, H. (1967). *Economics and Information Theory*. Chicago : Rand McNally et Company (cf. p. 48, 57, 59, 75).

Tilman, D., J. Knops, D. Wedin, P. Reich, M. Ritchie et E. Siemann (1997). « The Influence of Functional Diversity and Composition on Ecosystem Processes ». In : *Science* 277.5330, p. 1300–1302 (cf. p. 83).

Tothmeresz, B. (1995). « Comparison of different methods for diversity ordering ». In : *Journal of Vegetation Science* 6.2, p. 283–290 (cf. p. 66).

Triantis, K. A., F. Guilhaumon et R. J. Whittaker (2012). « The island species–area relationship : biology and statistics ». In : *Journal of Biogeography* 39.2, p. 215–231 (cf. p. 24).

Tsallis, C. (1988). « Possible generalization of Boltzmann-Gibbs statistics ». In : *Journal of Statistical Physics* 52.1, p. 479–487 (cf. p. 62).

— (1994). « What are the numbers that experiments provide ? » In : *Química Nova* 17.6, p. 468–471 (cf. p. 62).

Tsallis, C., R. S. Mendes et A. R. Plastino (1998). « The role of constraints within generalized nonextensive statistics ». In : *Physica A* 261.3, p. 534–554 (cf. p. 134).

Tucker, C. M., L. G. Shoemaker, K. F. Davies, D. Nemergut et B. A. Melbourne (2015). « Differentiating between niche and neutral assembly in metacommunities using null models of β-diversity ». In : *Oikos* in press (cf. p. 167).

Tuomisto, H. (2010a). « A diversity of beta diversities : straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity ». In : *Ecography* 33.1, p. 2–22 (cf. p. 123, 127).

— (2010b). « A diversity of beta diversities : straightening up a concept gone awry. Part 2. Quantifying beta diversity and related phenomena ». In : *Ecography* 33.1, p. 23–45 (cf. p. 123).

— (2011). « Commentary : do we have a consistent terminology for species diversity ? Yes, if we choose to use it ». In : *Oecologia* 167.4, p. 903–911 (cf. p. 127).

Tuomisto, H. (2012). « An updated consumer's guide to evenness and related indices ». In : *Oikos* 121.8, p. 1203–1218 (cf. p. 76, 80).

Ugland, K. I., J. S. Gray et K. E. Ellingsen (2003). « The species-accumulation curve and estimation of species richness ». In : *Journal of Animal Ecology* 72, p. 888–897 (cf. p. 12).

Ulanowicz, R. E. (2001). « Information theory in ecology ». In : *Computers & Chemistry* 25.4, p. 393–399 (cf. p. 60).

Van Valen, L. (1976). « Ecological Species, Multispecies, and Oaks ». In : *Taxon* 25.2/3, p. 233–239 (cf. p. 45).

Vane-Wright, R., C. Humphries et P. Williams (1991). « What to protect ?—Systematics and the agony of choice ». In : *Biological Conservation* 55.3, p. 235–254 (cf. p. 83, 92).

Veech, J. A. et T. O. Crist (2010). « Diversity partitioning without statistical independence of alpha and beta ». In : *Ecology* 91.7, p. 1964–1969 (cf. p. 123, 128).

Veech, J. A., K. S. Summerville, T. O. Crist et J. C. Gering (2002). « The additive partitioning of species diversity : recent revival of an old idea ». In : *Oikos* 99.1, p. 3–9 (cf. p. 128).

Vellend, M. (2001). « Do commonly used indices of f-diversity measure species turnover ? » In : *Journal of Vegetation Science* 12, p. 545–552 (cf. p. 124).

Vellend, M., W. K. Cornwell, K. Magnuson-Ford et A. Ø. Mooers (2010). « Measuring phylogenetic biodiversity ». In : *Biological diversity : frontiers in measurement and assessment*. Sous la dir. d'A. E. Magurran et B. J. McGill. Oxford : Oxford University Press, p. 194–207 (cf. p. 83).

Villéger, S. et D. Mouillot (2008). « Additive partitioning of diversity including species differences : a comment on Hardy & Senterre (2007) ». In : *Journal of Ecology* 96.5, p. 845–848 (cf. p. 139).

Villéger, S., N. W. H. Mason et D. Mouillot (2008). « New multidimensional functional diversity indices for a multifaceted framework in functional ecology ». In : *Ecology* 89.8, p. 2290–2301 (cf. p. 91).

Vinck, M., F. P. Battaglia, V. B. Balakirsky, A. J. H. Vinck et C. M. A. Pennartz (2012). « Estimation of the entropy based on its polynomial representation ». In : *Physical Review E* 85.5 (cf. p. 50, 68).

Volkov, I., J. R. Banavar, S. P. Hubbell et A. Maritan (2003). « Neutral theory and relative species abundance in ecology ». In : *Nature* 424.6952, p. 1035–1037 (cf. p. 24, 25).

Vu, V. Q., B. Yu et R. E. Kass (2007). « Coverage-adjusted entropy estimation ». In : *Statistics in Medicine* 26.21, p. 4039–4060 (cf. p. 49).

Wagner, H. H. (2003). « Spatial covariance in plant communities : Integrating ordination, geostatistics, and variance testing ». In : *Ecology* 84.4, p. 1045–1057 (cf. p. 148).

Walker, B., A. Kinzig et J. Langridge (1999). « Plant Attribute Diversity, Resilience, and Ecosystem Function : The Nature and Significance of Dominant and Minor Species ». In : *Ecosystems* 2.2, p. 95–113 (cf. p. 90).

Walker, B. H. (1992). « Biodiversity and Ecological Redundancy ». In : *Conservation biology* 6.1, p. 18–23 (cf. p. 83).

Wang, J.-P. (2010). « Estimating species richness by a Poisson-compound gamma model ». In : *Biometrika* 97.3, p. 727–740 (cf. p. 38).

— (2011). « SPECIES : An R Package for Species Richness Estimation ». In : *Journal of Statistical Software* 40.9, p. 1–15 (cf. p. 32).

Wang, J.-P., B. Lindsay, L. Cui, P. K. Wall, J. Marion, J. Zhang et C. DePamphilis (2005). « Gene capture prediction and overlap estimation in EST sequencing from one or multiple libraries ». In : *BMC Bioinformatics* 6.1, p. 300 (cf. p. 38).

Warwick, R. M. et K. R. Clarke (1995). « New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress ». In : *Marine Ecology Progress Series* 129, p. 301–305 (cf. p. 83, 94).

— (2001). « Practical Measures of Marine Biodiversity Based on Relatedness of Species ». In : *Oceanography and Marine Biology. An Annual Review*. Sous la dir. de R. Gibson, M. Barnes et R. J. A. Atkinson. T. 39. New York : Taylor & Francis, p. 207–232 (cf. p. 84).

Webb, C. O. (2000). « Exploring the phylogenetic structure of ecological communities : An example for rain forest trees ». In : *American Naturalist* 156.2, p. 145–155 (cf. p. 86).

Webb, C. O., J. B. Losos et A. A. Agrawal (2006). « Integrating Phylogenies into Community Ecology ». In : *Ecology* 87.sp7, S1–S2 (cf. p. 83).

Weikard, H.-P., M. Punt et J. Wesseler (2006). « Diversity measurement combining relative abundances and taxonomic distinctiveness of species ». In : *Diversity & Distributions* 12.2, p. 215–217 (cf. p. 93).

Weitzman, M. L. (1992). « On Diversity ». In : *The Quarterly Journal of Economics* 107.2, p. 363–405 (cf. p. 88, 97).

— (1993). « What to Preserve ? An Application of Diversity Theory to Crane Conservation ». In : *The Quarterly Journal of Economics* 108.1, p. 157–183 (cf. p. 97).

Whitlock, M. C. (2011). « G'ST and D do not replace FST ». In : *Molecular Ecology* 20.6, p. 1083–1091 (cf. p. 137).

Whittaker, R. H. (1960). « Vegetation of the Siskiyou Mountains, Oregon and California ». In : *Ecological Monographs* 30.3, p. 279–338 (cf. p. 4, 123, 127, 136).

— (1965). « Dominance and diversity in land plant communities ». In : *Science* 147.3655, p. 250–260 (cf. p. 2, 8).

— (1972). « Evolution and Measurement of Species Diversity ». In : *Taxon* 21.2/3, p. 213–251 (cf. p. 8, 127).

— (1977). « Evolution of species diversity in land communities ». In : *Evolutionary Biology* 10, p. 1–67 (cf. p. 5).

Williams, C. B. (1964). *Patterns in the balance of nature*. Academic Press (cf. p. 11).

Williams, P. et K. J. Gaston (1994). « Measuring more of biodiversity : can higher-taxon richness predict wholesale species richness ? » In : *Biological Conservation* 67, p. 211–217 (cf. p. 43).

Williamson, M. et K. J. Gaston (2005). « The lognormal distribution is not an appropriate null hypothesis for the species-abundance distribution ». In : *Journal of Animal Ecology* 74.2001, p. 409–422 (cf. p. 37).

Williamson, M., K. J. Gaston et W. M. Lonsdale (2001). « The species–area relationship does not have an asymptote ! » In : *Journal of Biogeography* 28.7, p. 827–830 (cf. p. 10, 11).

Wilson, E. O. et F. M. Peter, éds. (1988). *Biodiversity*. Washington, D.C. : The National Academies Press (cf. p. 1).

Wilson, J. B., R. K. Peet, J. Dengler et M. Pärtel (2012). « Plant species richness : The world records ». In : *Journal of Vegetation Science* 23.4, p. 796–802 (cf. p. 44).

Wilson, M. V. et a. Shmida (1984). « Measuring Beta Diversity with Presence-Absence Data ». In : *The Journal of Ecology* 72.3, p. 1055 (cf. p. 129).

Witting, L. et V. Loeschcke (1995). « The optimization of biodiversity conservation ». In : *Biological Conservation* 71.2, p. 205–207 (cf. p. 97).

Wright, I. J., P. B. Reich, M. Westoby, D. D. Ackerly, Z. Baruch, F. Bongers, J. Cavender-Bares, T. Chapin, J. H. C. Cornelissen, M. Diemer, J. Flexas, E. Garnier, P. K. Groom, J. Gulias, K. Hikosaka, B. B. Lamont, T. Lee, W. Lee, C. Lusk, J. J. Midgley, M.-L. Navas, Ü. Niinemets, J. Oleksyn, N. Osada, H. Poorter, P. Poot, L. Prior, V. I. Pyankov, C. Roumet, S. C. Thomas, M. G. Tjoelker, E. J. Veneklaas et R. Villar (2004). « The worldwide leaf economics spectrum ». In : *Nature* 428, p. 821–827 (cf. p. 85).

Wright, S. (1931). « Evolution in Mendelian Populations ». In : *Genetics* 16.2, p. 97–159 (cf. p. 61).

Würtz, P. et A. Annila (2008). « Roots of Diversity Relations ». In : *Journal of Biophysics* 2008, Article ID 654672 (cf. p. 24).

Yager, R. R. (1992). « On the specificity of a possibility distribution ». In : *Fuzzy Sets and Systems* 50.3, p. 279–292 (cf. p. 78).

Zacharias, M. A. et J. C. Roff (2001). « Use of focal species in marine conservation and management : a review and critique ». In : *Aquatic Conservation : Marine and Freshwater Ecosystems* 11.1, p. 59–76 (cf. p. 178).

Zadeh, L. (1965). « Fuzzy sets ». In : *Information and Control* 8.3, p. 338–353 (cf. p. 89).

Zanne, A. E., D. C. Tank, W. K. Cornwell, J. M. Eastman, S. A. Smith, R. G. FitzJohn, D. J. McGlinn, B. C. O'Meara, A. T. Moles, P. B. Reich, D. L. Royer, D. E. Soltis, P. F. Stevens, M. Westoby, I. J. Wright, L. Aarssen, R. I. Bertin, A. Calaminus, R. Govaerts, F. Hemmings, M. R. Leishman, J. Oleksyn, P. S. Soltis, N. G. Swenson, L. Warman et J. M. Beaulieu (2014). « Three keys to the radiation of angiosperms into freezing environments ». In : *Nature* 506.7486, p. 89–92 (cf. p. 85).

Zhang, C.-H. et Z. Zhang (2009). « Asymptotic Normality of a Nonparametric Estimator of Sample Coverage ». English. In : *Annals of Statistics* 37.5A, p. 2582–2595 (cf. p. 27).

Zhang, J., T. O. Crist et P. Hou (2014). « Partitioning of α and β diversity using hierarchical Bayesian modeling of species distribution and abundance ». In : *Environmental and Ecological Statistics* 21.4, p. 611–625 (cf. p. 74).

Zhang, Z. (2012). « Entropy Estimation in Turing's Perspective ». In : *Neural Computation* 24.5, p. 1368–1389 (cf. p. 49, 68).

— (2013). « Asymptotic normality of an entropy estimator with exponentially decaying bias ». In : *IEEE Transactions on Information Theory* 59.1, p. 504–508 (cf. p. 50).

Zhang, Z. et M. Grabchak (2013). « Bias adjustment for a nonparametric entropy estimator ». In : *Entropy* 15.6, p. 1999–2011 (cf. p. 50).

— (2014a). « Entropic Representation and Estimation of Diversity Indices ». In : *arXiv* 1403.3031.version 2 (cf. p. 67, 111).

— (2014b). « Nonparametric Estimation of Kullback-Leibler Divergence ». In : *Neural computation* 26.11, p. 2570–2593 (cf. p. 132).

Zhang, Z. et H. Huang (2007). « Turing's formula revisited ». In : *Journal of Quantitative Linguistics* 14.2-3, p. 222–241 (cf. p. 27).

Zhang, Z. et J. Zhou (2010). « Re-parameterization of multinomial distributions and diversity indices ». In : *Journal of Statistical Planning and Inference* 140.7, p. 1731–1738 (cf. p. 13, 67).