



Mémoire de synthèse

En vue de l'obtention de

l'Habilitation à Diriger des Recherches

Délivrée par : *l'Université de Guyane*

Présenté et soutenu publiquement le 14/04/2016 par :

ÉRIC MARCON

Mesurer la Biodiversité et la Structuration Spatiale

JURY

AVNER BAR-HEN

Professeur des Universités

Président du Jury

JÉRÔME CHAVE

Directeur de Recherche

Membre du Jury

JULIE LE GALLO

Professeur des Universités

Rapporteur

OLIVIER HARDY

Professeur des Universités

Membre du Jury

ABDENNEBI OMRANE

Professeur des Universités

Tuteur et Rapporteur

RAPHAËL PÉLISSIER

Directeur de Recherche

Membre du Jury

ALAIN ROUSTEAU

Maître de Conférences

Rapporteur

Rapporteurs :

Julie le Gallo, Abdennebi Omrane et Alain Rousteau

Table des matières

I	Présentation du candidat	1
1	Curriculum vitae	3
1.1	Titres universitaires	3
1.2	Fonctions occupées	3
1.3	Encadrement d'étudiants stagiaires	4
1.4	Encadrement de thèses	5
1.5	Activités d'enseignement	5
1.6	Activités éditoriales	5
2	Liste des publications	7
2.1	Articles dans des revues indexées	7
2.2	Autres Articles	8
2.3	Documents de Travail	8
2.4	Ouvrages et Chapitres d'ouvrages	8
2.5	Communications orales à des colloques et séminaires	8
II	Mémoire de synthèse	11
3	Introduction	13
4	Mesure de l'hétérogénéité spatiale	17
4.1	Analyse statistique des processus ponctuels	17
4.1.1	Processus ponctuels	17
4.1.2	Définition locale	18
4.1.3	Processus de Poisson homogène	21
4.1.4	Processus de Poisson inhomogène	21
4.1.5	Autres Processus	21
4.1.6	Vocabulaire	23
4.1.7	La fonction K de Ripley	24
4.1.8	Exemple	25
4.2	Tests statistiques	26
4.2.1	Test analytique de K	27
4.3	Généralisation de la fonction de Ripley	30
4.3.1	La fonction M	30
4.3.2	Application	32
4.3.3	Test de significativité	33
4.4	Typologie des mesures de structure spatiale	33

5	Mesure de la diversité	35
5.1	La diversité définie comme quantité d'information	35
5.1.1	Entropie et théorie de l'information	35
5.1.2	Application à la biodiversité	36
5.1.3	Biais d'estimation	38
5.1.4	Entropie HCDT	38
5.1.5	Logarithmes déformés	39
5.1.6	Entropie et diversité	40
5.1.7	Profils de diversité	41
5.2	Diversité neutre, phylogénétique et fonctionnelle	43
5.2.1	Phylodiversité	43
5.2.2	Diversité de Leinster et Cobbold	47
5.2.3	Diversité des valeurs propres	55
5.3	Diversité beta et décomposition	57
5.3.1	Définitions de la diversité beta, mesure dérivée	58
5.3.2	Le débat sur la décomposition	58
5.3.3	Décomposition multiplicative de la diversité	59
5.3.4	Définitions de la diversité alpha	60
5.3.5	Décomposition de l'entropie HCDT	61
5.3.6	Normalisation	68
5.3.7	Décomposition de la diversité phylogénétique	70
5.3.8	Partitionnement de la diversité de Leinster et Cobbold	71
5.3.9	Autres approches	72
5.4	Estimation	75
5.4.1	Le biais d'estimation	75
5.4.2	Techniques d'estimation	75
5.4.3	Pratique de l'estimation	77
5.5	Conclusion	80
6	Perspectives	81
6.1	Mesure de la diversité spatialement explicite	82
6.2	Transfert à l'économie géographique des méthodes de la biodiversité	82
6.3	Trajectoires de la diversité	83
6.4	Conclusion	84
	Bibliographie	87
	III Annexes : Publications	95
A	entropart, an R Package to Measure and Partition Diversity	97
B	Tools to Characterize Point Patterns: dbmss for R	125
C	Landscape patterns influence communities of medium- to large-bodied vertebrate in undisturbed terra firme forests of French Guiana	141
D	Decomposing phylodiversity	157
E	Generalization of the Partitioning of Shannon Diversity	165

F	A Statistical Test for Ripley's Function Rejection of Poisson Null Hypothesis	175
G	Testing randomness of spatial point patterns with the Ripley statistic	185
H	The decomposition of Shannon's entropy and a confidence interval for beta diversity	209
I	Characterizing the Relative Spatial Structure of Point Patterns	217
J	Measures of the geographic concentration of industries: improving distance-based methods	229
K	Assessing foliar chlorophyll contents with the SPAD-502 chlorophyll meter: a calibration test with thirteen tree species of tropical rainforest in French Guiana	249
L	Integrating functional diversity into tropical forest plantation designs to study ecosystem processes	255
M	Dynamics of aboveground carbon stocks in a selectively logged tropical forest	267
N	The successional status of tropical rainforest tree species is associated with differences in leaf carbon isotope discrimination and functional traits	277
O	A trait database for Guianan rain forest trees permits intra- and inter-specific contrasts	287
P	Evaluating the geographic concentration of industries using distance-based methods	295

Première partie

Présentation du candidat

CHAPITRE 1

Curriculum vitae

1.1 Titres universitaires

2005-2010 Thèse de doctorat en écologie.

École Doctorale ABIES. *Statistiques spatiales avec applications à l'écologie et à l'économie*. Directeurs de thèse : Gabriel Lang (AgroParisTech), Jean-Pierre Pascal (CNRS).

1997-1999 Ingénieur du Génie Rural, des Eaux et des Forêts.

École Nationale du Génie Rural, des Eaux et des Forêts (Paris), École d'application de l'Institut National Agronomique et de l'École Polytechnique.

1998-1999 DEA d'Économie Internationale et Économie du développement.

Université Paris I Sorbonne. Mémoire : *Le commerce international du bois*.

1987-1989 Ingénieur forestier.

École Nationale des Ingénieurs des Travaux des Eaux et Forêts (Nogent-sur-Vernisson, Loiret).

1.2 Fonctions occupées

Depuis 2010 Directeur l'Unité Mixte de Recherches Écologie des Forêts de Guyane.

Depuis 2006 Directeur du centre de Kourou d'AgroParisTech.

2006-2009 Directeur adjoint de l'Unité Mixte de Recherches Écologie des Forêts de Guyane.

2002-2006 Ingénieur, Enseignant-Chercheur à l'ENGREF de Kourou ; UMR EcoFoG.

Thème de recherche : Analyse spatiale en écologie forestière. Responsable du module d'enseignement Forêts Tropicales Humide.

1999-2002 Responsable informatique de la direction générale du Cemagref.

Institut de recherche pour l'ingénierie de l'agriculture et de l'environnement (Antony, Région parisienne). Conception et mise en place de l'administration du système informatique.

1995-1997 Responsable informatique du centre de Paris de l'EN-GREF.

Chargé de l'administration du réseau informatique. Responsable de l'analyse et du développement d'une base de données de gestion pour les besoins de l'école (relationnel entreprise et formation continue).

1991-1995 Chef de la division de Charleville-Rocroi à l'Office National des Forêts des Ardennes.

1990-1991 Analyste-programmeur à l'École d'Application de l'Artillerie (Draguignan, Var).
Service national.

1.3 Encadrement d'étudiants stagiaires

Mariwenn Ollivier Conception et implémentation d'une base de données regroupant les données de recherches sur les espèces forestières de Guyane.

Stage de césure de l'École Nationale Supérieure d'Agronomie de Montpellier, de septembre 2003 à février 2004. Ce stage a donné lieu à une publication.¹

¹M. OLLIVIER et al. (2007). « A trait database for Guianan rain forest trees permits intra- and inter-specific contrasts ». In : *Annals of Forest Science* 64.7, p. 781–786.

Lisa Cantet Prédiction de l'engorgement des sols forestiers tropicaux par le cortège floristique.

Stage du DEA Forêt, Agronomie et Gestion de l'Environnement de l'Université de Nancy, d'avril à juin 2004, co-encadré avec Bruno Ferry (École Nationale du Génie Rural, des Eaux et des Forêts).

Steve Rodrigues BTS Informatique de Gestion en alternance

De septembre 2005 à juin 2007, j'ai été le maître d'apprentissage de l'étudiant. Cet apprentissage a abouti à la création du système d'information de l'UMR EcoFoG.

Germain Baud Unification des bases de données Mariwenn et Guyafor.

Stage technique de troisième année de l'École des Mines d'Albi, de mai à août 2007. Ce stage a complété le développement de la base de données sur les espèces forestières de Guyane (notamment l'unification de son référentiel botanique avec celui des données d'inventaires forestiers de la base Guyafor) et sa mise en ligne².

²<http://mariwenn.ecofog.gf>

1.4 Encadrement de thèses

À partir du 1er octobre 2015, j'encadre la thèse d'Ariane Mirabel à l'École doctorale de l'Université de Guyane, sur le thème des trajectoires de biodiversité en forêt tropicale exploitée. Les objectifs de la thèse sont présentés dans les perspectives du mémoire de synthèse, page 83.

1.5 Activités d'enseignement

J'assure environ un quart de service d'enseignement, principalement dans les formations suivantes :

École doctorale Formations relatives à la réalisation de la thèse : traitement de texte scientifique (18 heures³), conception et réalisation de bases de données (18 heures⁴). Ces formations sont mises en place tous les trois ans.

³<http://www.ecofog.gf/spip.php?article426>

⁴<http://www.ecofog.gf/spip.php?article428>

Master avancé AgroParisTech Forêt, Nature et Société
Biodiversité : mesure et mécanismes (6 heures) ; Statistiques (4 heures).

Master 2 Écologie Forestière Tropicale Mesure de la Biodiversité (5 heures), Méthodes d'ordination (4 heures).

Master 2 Biologie, Chimie, Environnement (2014) Design expérimental (12 heures).

J'ai été membre du conseil doctoral de l'Université des Antilles et de la Guyane de 2010 à 2014. Je fais partie de l'équipe pédagogique du master Écologie Tropicale (Biodiversité, Écologie et Évolution à partir de 2015), j'assure en particulier la coordination des programmes entre les parcours de master portés par AgroParisTech, l'Université des Antilles, l'Université de Guyane, l'Université de Lorraine et l'Université de Montpellier. Enfin, je suis régulièrement membre des jurys de fin de stage du master Écologie Tropicale (jury commun à l'ensemble des parcours) et de restitution de projets du Module Forêt Tropicale Humide (FTH, formation AgroParisTech de quatre semaines regroupant une quarantaine d'étudiants de troisième année d'école d'ingénieur, de master 2 et de master avancé⁵).

⁵<http://www.ecofog.gf/spip.php?rubrique45>

J'ai organisé et encadré le module FTH (50 heures de cours, 50 heures de visites de terrain, une semaine de projet de terrain en forêt pour 10 groupes et une semaine d'analyse, rédaction et présentation) pendant 5 ans, de 2002 à 2006.

1.6 Activités éditoriales

Je suis rapporteur pour des revues d'écologie, d'économie et de statistiques appliquées principalement :

- Acta Biotheoretica
- Annals of Forest Science
- AStA Advances in Statistical Analysis
- Empirical Economics
- Entropy
- Environmental and Ecological Statistics
- Journal of Economic Geography
- Journal of Geographical Systems
- Journal of Vegetation Science
- Methods in Ecology and Evolution
- Molecular Ecology Resources
- Oecologia
- Papers in Regional Science
- Plos ONE
- Planning Practice and Research
- Regional Science and Urban Economics
- Regional Studies
- Stochastic Environmental Research and Risk Assessment
- Urban Studies

J'ai été récemment rapporteur pour une proposition d'ouvrage dans la section Mathématiques et Statistiques de Wiley.

Liste des publications

2.1 Articles dans des revues indexées

1. Marcon, E. et B. Hérault (2015). « entropart, an R Package to Measure and Partition Diversity ». In : Journal of Statistical Software 67.8, p. 1–26.
2. Marcon, E., S. Traissac, F. Puech et G. Lang (2015). « Tools to Characterize Point Patterns : dbmss for R ». In : Journal of Statistical Software 67.3, p. 1–15.
3. Richard-Hansen C., Jaouen G., Denis T., Brunaux O., Marcon E. et Guitet, S. (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.
4. Marcon, E. et B. Hérault (2015). « Decomposing Phylodiversity ». In : Methods in Ecology and Evolution 6.3, p. 333–339.
5. Marcon, E., I. Scotti, B. Hérault, V. Rossi et G. Lang (2014). « Generalization of the Partitioning of Shannon Diversity ». In : Plos One 9.3, e90289.
6. Marcon, E., S. Traissac et G. Lang (2013). « A Statistical Test for Ripley’s Function Rejection of Poisson Null Hypothesis ». In : ISRN Ecology 2013. Article ID 753475.
7. Lang, G. et E. Marcon (2013). « Testing randomness of spatial point patterns with the Ripley statistic ». In : ESAIM : Probability and Statistics 17, p. 767–788.
8. Marcon, E., 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.
9. Marcon, E., F. Puech et S. Traissac (2012). « Characterizing the Relative Spatial Structure of Point Patterns ». In : International Journal of Ecology 2012. Article ID 619281.
10. Marcon, E. et F. Puech (2010). « Measures of the Geographic Concentration of Industries : Improving Distance-Based Methods ». In : Journal of Economic Geography 10.5, p. 745–762.
11. Coste, S., C. Baraloto, C. Leroy, E. Marcon, A. Renaud, A. D. Richardson, J.-C. Roggy, H. Schimann, J. Uddling et B. Hérault (2010). « Assessing foliar chlorophyll contents with the SPAD-502 chlorophyll meter : a calibration test with thirteen tree species of tropical rainforest in French Guiana ». In : Annals of Forest Science 67.6, p. 607.
12. Baraloto, C., E. Marcon, F. Morneau, S. Pavoine et J.-C. Roggy (2010). « Integrating functional diversity into tropical forest plantation designs to study ecosystem processes ». In : Annals of Forest Science 67, p. 303.
13. Blanc, L., M. Echard, B. Hérault, D. Bonal, E. Marcon, J. Chave et C. Baraloto (2009). « Dynamics of aboveground carbon stocks in a selectively logged tropical forest ». In : Ecological Applications 19.6, p. 1397–1404.

14. Bonal, D., C. Born, C. Brechet, S. Coste, E. Marcon, J. C. Roggy et J. M. Guehl (2007). « The successional status of tropical rainforest tree species is associated with differences in leaf carbon isotope discrimination and functional traits ». In : *Annals of Forest Science* 64.2, p. 169–176.
15. Ollivier, M., C. Baraloto et E. Marcon (2007). « A trait database for Guianan rain forest trees permits intra- and inter-specific contrasts ». In : *Annals of Forest Science* 64, p. 781–786.
16. Marcon, E. et F. Puech (2003). « Evaluating the geographic concentration of industries using distance-based methods ». In : *Journal of Economic Geography* 3.4, p. 409–428.

2.2 Autres Articles

1. Marcon, E. et F. Puech (2015). « Mesures de la concentration spatiale en espace continu : théorie et applications ». In : *Économie et Statistique* 474, p. 105–131.
2. Marcon, E. (1999). « Forest surveys on a tree by tree basis : A theoretical and practical approach ». In : *Revue Forestière Française* 51.1, p. 57–69.

2.3 Documents de Travail

1. Marcon, E. (2015). « Practical Estimation of Diversity from Abundance Data ». In : HAL 01212435. version 1, p. 1–27.
2. Marcon, E. et F. Puech (2015). « A Typology of Distance-Based Measures of Spatial Concentration ». In : HAL SHS 00679993. version 4, p. 1–16.
3. Lang, G., E. Marcon et F. Puech (2015). « Distance-Based Measures of Spatial Concentration: Introducing a Relative Density Function ». HAL 01082178. version 2, p. 1–18.
4. Marcon, E., Z. Zhang et B. Hérault (2014). « The Decomposition of Similarity-Based Diversity and its Bias Correction ». In : HAL 00989454. version 2, p. 1–12.

2.4 Ouvrages et Chapitres d'ouvrages

1. Marcon, E. (2015). *Mesures de la Biodiversité*. Kourou, France : UMR EcoFoG. ¹
2. Marcon, E. et F. Puech (2012). « La mesure en économie internationale ». In : *Développements récents en économie et finances internationales*. Sous la dir. de L. Abdelmalki, J.-P. Allegret, F. Puech, M. S. Jallab et A. Silem. Paris : Armand Colin, p. 15–27.
3. Marcon, E., J.-L. Mucchielli et F. Puech (2005). « Concentration géographique de l'emploi industriel et dynamiques territoriales en France de 1993 à 2001 ». In : *Localisation des activités et stratégies de l'état - Rapport du commissariat général du plan Groupe Perroux*. Sous la dir. de E. M. Mouhoud. Paris : L'Action Municipale, p. 99–109.

2.5 Communications orales à des colloques et séminaires

1. E. Marcon (2015). La dualité entropie-diversité pour mesurer la biodiversité. Journées du GDR Écologie Statistique (EcoStat), Lyon (France), Mars 2015.
2. E. Marcon (2015). Mesure de la biodiversité : avancées récentes. Première réunion du Groupe de travail AnaEE France « Biodiversité », Moulis (France), Février 2015.

¹http://www.ecofog.gf/IMG/pdf/mesures_de_la_biodiversite.pdf

3. G. Lang, E. Marcon et F. Puech (2014). Distance-based measures of spatial concentration: introducing a relative density function. 61^e Conférence Annuelle de l'Association Nord-Américaine de Sciences Régionales (Annual North American Meeting of the Regional Science Association International – RSAI), Washington D.C. (États-Unis), Novembre 2014.
4. R. Pélissier, P. Couteron, O. Hardy, E. Marcon, S. Pavoine (2014). Quantifying spatial patterns of species diversity : integrating methods of spatial and diversity analyses. International Statistical Ecology Conference 2014, Montpellier (France), Juillet 2014.
5. G. Lang, E. Marcon et F. Puech (2014). Distance-based measures of spatial concentration: introducing a relative density function. Journée de recherche pluridisciplinaire en statistiques spatiales, Sceaux (France), Juin 2014.
6. G. Lang, E. Marcon et F. Puech (2014) Distance-based measures of spatial concentration: introducing a relative density function. 13th International Workshop Spatial Econometrics and Statistics, Toulon (France), Avril 2014.
7. E. Marcon et F. Puech (2013). A Typology of Distance-Based Measures of Spatial Concentration. 12th International Workshop Spatial Econometrics and Statistics, Orléans (France), Juin 2013.
8. E. Marcon et F. Puech (2013). A Typology of Distance-Based Measures of Spatial Concentration. Séminaire Hotelling (CES – ENS Cachan – ADIS), Sceaux (France), Mars 2013.
9. E. Marcon et F. Puech (2012). A Typology of Distance-Based Measures of Spatial Concentration. LXI^e Congrès Annuel de l'Association Française de Sciences Économiques (AFSE), Paris (France), Juillet 2012.
10. E. Marcon, F. Puech (2004). Characterizing spatial structures: towards a consistent approach? Journées interdisciplinaires de statistiques spatiales. Paris (France), Décembre 2004.
11. E. Marcon, J.-L. Mucchielli et F. Puech (2004). Concentration géographique de l'emploi industriel et dynamiques territoriales en France de 1993 à 2001. Présentation de l'étude complémentaire pour le Groupe de réflexion François Perroux du Commissariat Général du Plan « Prospective de localisation des activités pour les régions françaises dans une Union européenne élargie » (2003-2004), Paris (France), Juin 2004.
12. E. Marcon et F. Puech (2003). The Determinants of Agglomeration in a Continuous-Space Framework. Séminaire à l'INRA – ENESAD, Dijon (France). Mars 2004.
13. E. Marcon et F. Puech (2003). The Determinants of Agglomeration in a Continuous-Space Framework. 50^e Conférence Annuelle de l'Association Nord-Américaine de Sciences Régionales (Annual North American Meeting of the Regional Science Association International – RSAI), Philadelphie (États-Unis), Novembre 2003.
14. E. Marcon et F. Puech (2003). Measures of the geographic concentration of industries: improving distance-based methods. LII^e Congrès Annuel de l'Association Française de Sciences Économiques (AFSE), Paris (France), Septembre 2003. ²
15. E. Marcon et F. Puech (2003). The Determinants of Agglomeration in a Continuous-Space Framework. Séminaire du laboratoire TEAM de l'Université de Paris I, Paris (France), Juin 2003.
16. E. Marcon et F. Puech (2003). Measures of the geographic concentration of industries: improving distance-based methods. Meeting CEPR : The Economics of Cities: Technology, Integration and Local Labour Markets, Londres (Grande-Bretagne), Juin 2003.
17. E. Marcon et F. Puech (2003). Measures of the geographic concentration of industries: improving distance-based methods. Deuxième Journée d'Économétrie Spatiale (2nd Spatial

²http://www.afse.fr/docs/AFSE_56.pdf

- Econometrics Workshop), Dijon (France), Mai 2003.
18. E. Marcon et F. Puech (2003). Measures of the geographic concentration of industries: improving distance-based methods. Séminaire du laboratoire TEAM de l'Université de Paris I, Paris (France), Janvier 2003.
 19. E. Marcon et F. Puech (2002). Measures of the geographic concentration of industries: improving distance-based methods Groupement de Recherche en Économie et Finances Internationales Quantitatives (EFIQ), Nanterre (France), Décembre 2002.
 20. E. Marcon et F. Puech (2002). Measures of the geographic concentration of industries: improving distance-based methods. Douzièmes Journées du Séminaire d'Études et de Statistiques Appliquées à la Modélisation en Économie (SESAME), Aix-en-Provence (France), Septembre 2002.
 21. E. Marcon et F. Puech (2002). A New Method to Evaluate Spatial Economic Activity and its Application to Two French Areas. LI^e Congrès Annuel de l'Association Française de Sciences Économiques (AFSE), Paris (France), Septembre 2002.
 22. E. Marcon et F. Puech (2002). A New Method to Evaluate Spatial Economic Activity and its Application to Two French Areas. XVII^e Congrès Annuel de l'Association Économique Européenne (European Economic Association – EEA), Venise (Italie), Août 2002. ³
 23. E. Marcon et F. Puech (2002). A New Method to Evaluate Spatial Economic Activity and its Application to Two French Areas. Spring Workshop on Economic Geography and Multinationals' Location, Paris (France), Mai 2002.
 24. E. Marcon et F. Puech (2002). A New Method to Evaluate Spatial Economic Activity and its Application to Two French Areas. 7^e Rencontre des Jeunes Économistes (VIIth Spring Meeting of Young Economists), Paris (France), Avril 2002.
 25. E. Marcon et F. Puech (2002). A New Method to Evaluate Spatial Economic Activity and its Application to Two French Areas. Séminaire du laboratoire TEAM de l'Université de Paris I, Paris (France), Février 2002.
 26. E. Marcon (2001). The Structure of International Timber Trade. 76th Western Economic Association International Annual Conference, San Francisco (États-Unis), Juillet 2001.

³<http://www.eea-esem.com/eea-esem/eea2002/prog/viewpaper.asp?pid=1611>

Deuxième partie

Mémoire de synthèse

CHAPITRE 3

Introduction

JE suis entré dans le monde de la recherche par son administration. En tant qu'ingénieur des corps de l'État, j'ai été responsable de l'ingénierie de formation au centre d'AgroParis-Tech de Kourou de 2002 à 2005, en support de l'UMR Écologie des Forêts de Guyane nouvellement créée. J'ai pris la direction du centre en 2006 et je suis devenu en même temps directeur adjoint de l'UMR. J'ai vraiment commencé une activité de recherche classique, c'est-à-dire un doctorat, à ce moment.

Avant cela, j'ai été ingénieur forestier à l'Office National des Forêts, en poste dans les Ardennes de 1991 à 1995. La géographie a fait que de nombreuses lignes électriques ont été créées dans la région, pour rejoindre la Belgique et surtout une centrale nucléaire frontalière. L'évaluation de l'indemnisation des propriétaires forestiers concernés manquait clairement de support théorique. J'ai développé une méthode d'évaluation de la valeur économique d'une forêt publiée dans la Revue Forestière Française,¹ revue nationale à comité de lecture indexée par Scopus.

J'ai été responsable informatique du centre de Paris de l'EN-GREF de 1995 à 1997 et de la direction générale du Cemagref de 1999 à 2002, avec une interruption de 1997 à 1999 pour reprendre des études et devenir ingénieur du génie rural, des eaux et des forêts. J'ai obtenu un DEA en économie internationale en 1999 dans le cadre de cette formation. J'ai commencé à cette époque à m'intéresser aux questions de structuration spatiale, appliquées aux arbres comme aux entreprises, mais sans disposer du temps ni du cadre professionnel nécessaire pour approfondir. J'ai tout de même établi à cette occasion une collaboration avec Florence Puech, doctorante à l'Université de Paris I, Panthéon-Sorbonne, puis Maître de Conférence à l'Université de Lyon, qui a débouché sur une publication² qui a contribué significativement³ à l'introduction de l'analyse statistique des processus ponctuels en économie géographique.

En tant qu'ingénieur à l'UMR EcoFoG à partir de 2002, j'ai participé à plusieurs programmes de recherche concernant le cycle

¹É. MARCON (1999). « Forest surveys on a tree by tree basis : A theoretical and practical approach ». In : *Revue Forestière Française* 51.1, p. 57–69.

²É. MARCON et F. PUECH (2003). « Evaluating the geographic concentration of industries using distance-based methods ». In : *Journal of Economic Geography* 3.4, p. 409–428.

³P.-P. COMBES et al. (2008). *Economic Geography*. Princeton, New Jersey : Princeton University Press, p. 1–416, Chapitre 10.

⁴OLLIVIER et al. (2007). « A trait database for Guianan rain forest trees permits intra- and inter-specific contrasts », cf. note 1, p. 4; D. BONAL et al. (2007). « The successional status of tropical rainforest tree species is associated with differences in leaf carbon isotope discrimination and functional traits ». In : *Annals of Forest Science* 64.2, p. 169–176; L. BLANC et al. (2009). « Dynamics of aboveground carbon stocks in a selectively logged tropical forest ». In : *Ecological Applications* 19.6, p. 1397–1404; C. BARALOTO et al. (2010b). « Integrating functional diversity into tropical forest plantation designs to study ecosystem processes ». In : *Annals of Forest Science* 67.3, p. 303; S. COSTE et al. (2010). « Assessing foliar chlorophyll contents with the SPAD-502 chlorophyll meter : a calibration test with thirteen tree species of tropical rainforest in French Guiana ». In : *Annals of Forest Science* 67.6, p. 607.

⁵F. GOREAUD (2000). « Apports de l'analyse de la structure spatiale en forêt tempérée à l'étude de la modélisation des peuplements complexes ». Thèse de doct. Nancy : ENGREF.

⁶É. MARCON et F. PUECH (2010). « Measures of the Geographic Concentration of Industries : Improving Distance-Based Methods ». In : *Journal of Economic Geography* 10.5, p. 745–762; É. MARCON et al. (2012a). « Characterizing the relative spatial structure of point patterns ». In : *International Journal of Ecology* 2012.Article ID 619281, p. 11.

⁷B. D. RIPLEY (1976). « The Foundations of Stochastic Geometry ». In : *Annals of Probability* 4.6, p. 995–998; B. D. RIPLEY (1977). « Modelling Spatial Patterns ». In : *Journal of the Royal Statistical Society B* 39.2, p. 172–212.

⁸J. E. BESAG (1977). « Comments on Ripley's paper ». In : *Journal of the Royal Statistical Society B* 39.2, p. 193–195.

⁹É. MARCON et F. PUECH (2015a). « A Typology of Distance-Based Measures of Spatial Concentration ». In : *HAL SHS* 00679993.version 4, p. 1–16; É. MARCON et F. PUECH (2015b). « Mesures de la concentration spatiale en espace continu : théorie et applications ». In : *Economie et Statistique* 474, p. 105–131.

¹⁰R. LAW et al. (2009). « Ecological information from spatial patterns of plants : insights from point process theory ». English. In : *Journal of Ecology* 97.4, p. 616–628.

¹¹R. PÉLISSIER et F. GOREAUD (2001). « A practical approach to the study of spatial structure in simple cases of heterogeneous vegetation ». In : *Journal of Vegetation Science* 12.1, p. 99–108.

du carbone et l'écologie des communautés abordée sous l'angle de l'assemblage des traits fonctionnels. Ces travaux ont donné lieu à 5 publications de 2007 à 2010.⁴

Mes travaux de thèse commencés en 2006 ont porté sur la mesure de la structuration spatiale, appliquée à la forêt tropicale, sous la direction de Jean-Pierre Pascal, Directeur de Recherche au CNRS, en écologie et de Gabriel Lang, Ingénieur du Génie Rural, des Eaux et des Forêts, en mathématiques. L'élément déclencheur a été la thèse de François Goreaud⁵ qui m'avait initié aux statistiques non paramétriques appliquées aux processus ponctuels, qui permettaient de traiter des questions très similaires à celles de l'économie géographique, mais avec des outils plus puissants.

Ces questions sont celles de l'hétérogénéité spatiale, qui peut être de premier ordre (la variabilité de l'intensité du processus responsable de la distribution spatiale des objets) et de second ordre (la dépendance entre les objets, qui peuvent s'attirer ou se repousser). La variabilité de second ordre est le sujet qui m'a intéressé particulièrement.

J'ai travaillé à l'amélioration de la caractérisation de cette variabilité en développant des méthodes prenant en compte l'hétérogénéité de premier ordre, qui ont donné lieu à la publication d'une nouvelle fonction non paramétrique appelée M^6 (à la suite des fonctions K de Ripley⁷ et de sa variante L de Besag⁸). Les applications de ce nouvel outil en écologie sont restées confidentielles, alors que l'impact a été important en économie (plus de 100 citations en juin 2015 sur Google Scholar pour l'article dans *Journal of Economic Geography* contre 4 pour celui dans *International Journal of Ecology*). La raison principale est que le cadre théorique en économie correspond parfaitement à l'approche développée pour la fonction M et théorisée plus clairement ensuite,⁹ alors que l'homogénéité de premier ordre est souvent une approximation acceptable en écologie¹⁰ où la fonction K de Ripley est souvent applicable, après un découpage éventuel de l'espace en zones homogènes.¹¹

J'ai développé en parallèle, sous la direction de Gabriel Lang, une approche mathématique « dure » de la théorie des processus ponctuels, qui a abouti à la publication du premier test statistique non asymptotique de la fonction K contre une distribution complètement aléatoire.¹²

Pour permettre l'utilisation large de ces méthodes, j'ai publié un package pour R¹³ qui les rassemble.¹⁴

Ces résultats sont présentés dans la première partie de ce mémoire.

Les mesures de concentration spatiale sont des mesures d'inégalité. En économie, le lien entre la concentration spatiale (des secteurs d'activités) et la spécialisation (des régions) sont considé-

rées comme deux aspects du même phénomène.¹⁵ La spécialisation est l'équivalent de la diversité, selon un point de vue opposé : pour l'illustrer, on peut noter que l'indice de Theil¹⁶ utilisé en économie pour mesurer l'inégalité ou la concentration est le complément au logarithme du nombre de catégories (secteurs économiques, espèces) de l'indice de Shannon¹⁷ utilisé pour mesurer la biodiversité. Les littératures traitant de l'inégalité économique, de la concentration spatiale et de la diversité partagent de nombreux concepts et outils, même si les objets sont très différents. Le glissement thématique de mes recherches de la structure spatiale à la biodiversité a donc été assez naturel.

J'ai commencé à travailler sur l'entropie comme mesure de la diversité après l'avoir utilisée en économie comme mesure de la concentration. L'article¹⁸ de Lou Jost sur l'opposition entre l'entropie et la diversité et un désaccord sur la notion de diversité β ¹⁹ m'ont motivé à aller plus loin. Après un travail préliminaire sur la décomposition de la diversité de Shannon,²⁰ j'ai généralisé les concepts qui avaient été formulés par Patil et Taillie²¹ : la dualité entre l'entropie, mesure de surprise moyenne, et la diversité, mesure d'un nombre effectif de classes ; la définition de l'entropie β comme information supplémentaire apportée par la connaissance des distributions locales en plus de la distribution globale et de la diversité β comme un nombre effectif de communautés. Cette approche m'a permis de montrer que l'entropie généralisée²² pouvait être utilisée au-delà de l'entropie de Shannon pour mesurer la diversité dans des formes très similaires.²³ Une nouvelle généralisation permet de traiter la diversité phylogénétique de la même façon.²⁴

J'ai appliqué la même approche à la diversité fondée sur la similarité²⁵ pour définir sa diversité β .²⁶

L'estimation de la diversité à partir de données réelles est un problème particulièrement intéressant du point de vue mathématique. J'ai développé des estimateurs nouveaux pour toutes les mesures de diversité que j'ai traitées.

Pour permettre l'utilisation de ces méthodes par les écologues, je les ai intégrées dans un package²⁷ pour R et je tiens à jour un support de cours (en Français) accessible en ligne qui ressemble de plus en plus à un livre sur le sujet²⁸. Ce document a été téléchargé plus de 1000 fois sur ResearchGate entre mai et juin 2015.

Ces travaux sont présentés en deuxième partie du mémoire de synthèse. La troisième partie présente mes perspectives.

¹²G. LANG et É. MARCON (2013). « Testing randomness of spatial point patterns with the Ripley statistic ». In : *ESAIM : Probability and Statistics* 17, p. 767–788. arXiv : 1006.1567 ; É. MARCON et al. (2013). « A Statistical Test for Ripley's Function Rejection of Poisson Null Hypothesis ». In : *ISRN Ecology* 2013.Article ID 753475, p. 9.

¹³R CORE TEAM (2015). *R : A Language and Environment for Statistical Computing*. Vienna, Austria : R Foundation for Statistical Computing.

¹⁴É. MARCON et al. (2015). « Tools to Characterize Point Patterns : dbmss for R ». In : *Journal of Statistical Software* 67.3, p. 1–15.

¹⁵E. CUTRINI (2010). « Specialization and Concentration from a Two-fold Geographical Perspective : Evidence from Europe ». In : *Regional Studies* 44.3, p. 315–336.

¹⁶H. THEIL (1967). *Economics and Information Theory*. Chicago : Rand McNally et Company.

¹⁷C. E. SHANNON (1948). « A Mathematical Theory of Communication ». In : *The Bell System Technical Journal* 27, p. 379–423, 623–656.

¹⁸L. JOST (2006). « Entropy and diversity ». In : *Oikos* 113.2, p. 363–375.

¹⁹L. JOST (2007). « Partitioning diversity into independent alpha and beta components ». In : *Ecology* 88.10, p. 2427–2439.

²⁰É. MARCON et al. (2012b). « The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity ». In : *Oikos* 121.4, p. 516–522.

²¹G. 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.

²²C. TSALLIS et E. BRIGATTI (2004). « Nonextensive statistical mechanics : A brief introduction ». English. In : *Continuum Mechanics and Thermodynamics* 16.3, p. 223–235.

²³É. MARCON et al. (2014a). « Generalization of the partitioning of Shannon diversity ». In : *Plos One* 9.3, e90289.

²⁴É. MARCON et B. HÉRAULT (2015a). « Decomposing Phylodiversity ». In : *Methods in Ecology and Evolution* 6.3, p. 333–339.

²⁵T. LEINSTER et C. COBBOLD (2012). « Measuring diversity : the importance of species similarity ». In : *Ecology* 93.3, p. 477–489.

²⁶É. MARCON et al. (2014b). « The Decomposition of Similarity-Based Diversity and its Bias Correction ». In : *HAL* 00989454.version 3, p. 1–12.

²⁷É. MARCON et B. HÉRAULT
(2015b). « entropart, an R Package
to Measure and Partition Diversity ». *In : Journal of Statistical Software*
67.8, p. 1–26.

²⁸[http://www.ecofog.gf/IMG/pdf/
mesures_de_la_biodiversite.pdf](http://www.ecofog.gf/IMG/pdf/mesures_de_la_biodiversite.pdf)

CHAPITRE 4

Mesure de l'hétérogénéité spatiale

JE résume dans cette section mes travaux en statistiques spatiales. La première section (4.1) introduit les notions nécessaires. Les deux suivantes présentent mes apports : le test statistique de la fonction K de Ripley (4.2) et les nouvelles fonctions de second ordre non paramétriques (4.3).

4.1 Analyse statistique des processus ponctuels

4.1.1 Processus ponctuels

Les processus ponctuels fournissent le cadre mathématique nécessaire à l'étude des structures spatiales. L'approche utilisée classiquement par les non-mathématiciens est locale : les propriétés d'un processus sont définies autour de chaque point. Elle a l'avantage d'être concrète et facilement compréhensible. Son inconvénient est de laisser un certain flou sur le comportement global du processus. Une définition globale est nécessaire pour étudier rigoureusement les processus : elle est présentée ici avant son équivalent local.

Définition

Un processus ponctuel¹ est un sous-ensemble aléatoire dénombrable d'un espace $S \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$.

Les définitions données ici et la plupart des résultats sont valables dans un espace de dimension quelconque finie (la formulation des statistiques descriptives peut dépendre du nombre de dimensions) mais on se limitera en pratique à un espace à deux dimensions.

Nous nous intéresserons à des ensembles dénombrables de points (on dira aussi « semis de points ») notés X . Les points seront notés en minuscules, les ensembles en majuscules : $X \subset \mathbb{R}^2$,

¹ J. MØLLER et R. P. WAAGEPETERSEN (2004). *Statistical Inference and Simulation for Spatial Point Processes*. T. 100. Chapman et Hall, p. 1–300.

$x \in X$. Le semis de point X sera généralement défini sur tout le plan, et son nombre de points sera infini.

À l'intérieur de l'aire d'étude A (aussi appelée fenêtre), un sous ensemble de X noté X_A sera observé et cartographié : $X_A = X \cap A$. Nous ne nous intéresserons qu'à des ensembles de points localement finis, c'est-à-dire tels que leur nombre de points dans A soit fini : $n(A) < \infty$ pour A borné. Cette restriction n'a pas de conséquences pratiques. Il est impossible de définir directement une fonction qui attribuerait à chaque semis la probabilité de le tirer, parce que la probabilité de chaque semis est nulle. On passe donc par des ensembles de semis de points, dont la probabilité n'est pas nulle.

Les processus sont notés en lettres grecques majuscules, par exemple Ξ . Un semis de points X est une réalisation de Ξ . On note $P(X \in F)$ la probabilité que le tirage de Ξ soit un élément d'un ensemble de semis de points F particulier, par exemple défini par son nombre de points.

Propriété de premier ordre Soit S une partie de A . La propriété de premier ordre $\mu(S)$, appelée également mesure d'intensité du processus Ξ , est l'espérance du nombre de points dans S :

$$\mu(S) = \mathbb{E}(N(S)) \quad (4.1)$$

Dans tous les cas que nous traiterons, la mesure d'intensité pourra être écrite comme l'intégrale d'une fonction d'intensité λ :

$$\mu(S) = \int_S \lambda(x) dx \quad (4.2)$$

Propriété de second ordre La mesure du moment factoriel de second ordre de deux parties de A , S_1 et S_2 , est l'espérance du nombre de paires de points du processus Ξ se trouvant respectivement dans S_1 et S_2 :

$$\mu_2(S_1, S_2) = \mathbb{E} \left(\sum_{x_1 \neq x_2 \in X} \mathbf{1}(x_1 \in S_1, x_2 \in S_2) \right) \quad (4.3)$$

La fonction indicatrice $\mathbf{1}(T)$ vaut 1 si T est vrai, 0 sinon.

De même, cette mesure pourra être écrite comme l'intégrale de λ_2 , appelée densité du produit de second ordre :

$$\mu(S_1, S_2) = \iint_{(\mathbb{R}^2)^2} \mathbf{1}(x_1 \in S_1, x_2 \in S_2) \lambda_2(x_1, x_2) dx_1 dx_2 \quad (4.4)$$

4.1.2 Définition locale

Un processus ponctuel est l'équivalent d'une variable aléatoire dont le résultat est un ensemble de points noté X_A , dans un ensemble

de réalisations possibles, qui sera toujours ici une surface connue et délimitée notée A .

On utilise les processus ponctuels comme outils mathématiques pour caractériser et éventuellement modéliser des événements dont on connaît la répartition spatiale, par exemple les arbres dans une forêt.

Une façon intéressante de décrire un processus ponctuel dont on ne connaît pas la loi consiste à utiliser ses propriétés de premier ordre et de second ordre.

Propriété de premier ordre Considérons une surface A dans laquelle on observe une réalisation d'un processus ponctuel. Chaque point est noté x . On note $N(S)$ le nombre (aléatoire) de points situés dans une surface S donnée. La propriété de premier ordre du processus ponctuel est son intensité, notée $\lambda(x)$. Elle est définie par :

$$\lambda(x) = \lim_{dx \rightarrow 0} \frac{\mathbb{E}(N(dx))}{dx} \quad (4.5)$$

dx est la surface élémentaire définie autour du point x . Si $\lambda(x)$ est constante, on parlera de processus ponctuel homogène et on notera l'intensité simplement λ . Un processus est stationnaire s'il est invariant par translation et isotrope s'il est invariant par rotation. Un processus homogène est donc à la fois stationnaire et isotrope.

Probabilité de trouver un point dans une surface élémentaire On ne s'intéressera ici qu'à des processus ponctuels ordonnés,² c'est-à-dire dont la probabilité de trouver plusieurs points sur une surface élémentaire dx est d'un ordre de grandeur plus petit que dx . Cette hypothèse n'est pas contraignante : elle élimine les processus coalescents (dans lesquels tous les points se trouveraient superposés parce qu'ils s'attirent entre eux) ou des processus qui généreraient par exemple pour chaque point un deuxième superposé. En pratique, tous les processus ponctuels que l'on rencontrera seront ordonnés.

Cette propriété permet de lier la probabilité à l'intensité. L'existence d'un point dans la surface dx suit une loi de Bernoulli de paramètre $P(dx)$, qui est à la fois sa probabilité de succès et son espérance. Cette espérance est, d'après l'équation (4.5), $\lambda(x)dx$.

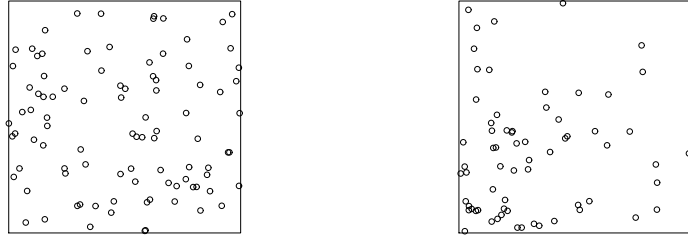
La probabilité de trouver un point dans la surface élémentaire dx autour du point x est par conséquent :

$$P(N(dx) = 1) = \lambda(x)dx \quad (4.6)$$

Cette relation est vérifiée tant que dx est suffisamment petite pour que la probabilité d'apparition de deux points reste négligeable.

²P. J. DIGGLE (1983). *Statistical analysis of spatial point patterns*. London : Academic Press, p. 1-148, page 47, Annexe C.

FIGURE 4.1 – Processus de Poisson



(a) Réalisation d'un processus de Poisson homogène dont l'intensité est 500 points sur la surface de la fenêtre.

(b) Réalisation d'un processus de Poisson inhomogène dont l'intensité décroît en s'éloignant du coin inférieur gauche de la fenêtre. La fenêtre est un carré de côté 1 et l'intensité $\lambda(x) = 500e^{-3\|x\|}$

Propriété de second ordre La propriété de second ordre d'un processus ponctuel, notée $\lambda_2(x_1, x_2)$ est définie par :

$$\lambda_2(x_1, x_2) = \lim_{dx_1 \rightarrow 0, dx_2 \rightarrow 0} \frac{\mathbb{E}(N(dx_1)N(dx_2))}{dx_1 dx_2} \quad (4.7)$$

λ_2 est aussi appelée densité de paires de points.³

Probabilité de trouver deux points dans deux surfaces élémentaires La probabilité jointe de la présence d'au moins un point dans chaque surface élémentaire centrée sur x_1 et x_2 est notée $P(dx_1 dx_2)$. Ici encore, la probabilité de trouver plus d'un point dans une surface élémentaire est négligeable. L'événement « trouver à la fois un point dans dx_1 et dans dx_2 » réalise une épreuve de Bernoulli de paramètre $P(dx_1 dx_2)$. Selon le même raisonnement que précédemment, son espérance est $P(dx_1 dx_2)$. Or cette espérance est connue (4.7), d'où :

$$P(N(dx_1)N(dx_2) = 1) = \lambda_2(x_1, x_2) dS_1 dS_2 \quad (4.8)$$

On peut rapporter dS_1 et dS_2 à la propriété de premier ordre pour obtenir :

$$\begin{aligned} P(N(dx_1)N(dx_2) = 1) \\ = P(N(dx_1) = 1)P(N(dx_2) = 1) \frac{\lambda_2(x_1, x_2)}{\lambda(x_1)\lambda(x_2)} \end{aligned} \quad (4.9)$$

La grandeur $\lambda_2(x_1, x_2)/\lambda(x_1)\lambda(x_2)$, rapport de la propriété de second ordre sur la propriété de premier ordre, est appelée fonction de distribution radiale⁴ ou fonction de corrélation des paires de points.⁵ Nous suivrons Ripley⁶ et toute la littérature en découlant en la notant $g(x_1, x_2)$. L'usage a imposé g plutôt que λ_2 comme mesure de la propriété de second ordre des processus ponctuels⁷ :

$$g(x_1, x_2) = \frac{P(N(dx_1)N(dx_2) = 1)}{P(N(dx_1) = 1)P(N(dx_2) = 1)} \quad (4.10)$$

³LAW et al. (2009). « Ecological information from spatial patterns of plants : insights from point process theory », cf. note 10, p. 14.

⁴DIGGLE (1983). *Statistical analysis of spatial point patterns*, cf. note 2, p. 19.

⁵N. A. CRESSIE (1993). *Statistics for spatial data*. New York : John Wiley & Sons, p. 1-900.

⁶RIPLEY (1977). « Modelling Spatial Patterns », cf. note 7, p. 14.

⁷Par exemple : RIPLEY (1977). « Modelling Spatial Patterns », cf. note 7, p. 14 ; GOREAUD (2000). « Apports de l'analyse de la structure spatiale en forêt tempérée à l'étude de la modélisation des peuplements complexes », cf. note 5, p. 14.

Si le processus est isotrope, c'est-à-dire que ses propriétés sont les mêmes dans toutes les directions, et que sa propriété de second ordre est stationnaire, $g(\cdot)$ ne dépend que de la distance entre les deux points et on la notera simplement $g(r)$.

On voit immédiatement que dans le cas d'une distribution de points indépendants, la probabilité jointe est égale au produit des probabilités, et par conséquent $g(\cdot) = 1$.

4.1.3 Processus de Poisson homogène

Le processus de Poisson est un processus stationnaire et isotrope, dont la réalisation donne des points à la position complètement aléatoire. Inversement, un processus ponctuel complètement aléatoire est un processus de Poisson.⁸

Les points sont distribués indépendamment les uns des autres, avec une intensité constante (Figure 4.1a). Ce processus est souvent appelé « distribution complètement aléatoire », ou CSR (*Complete Spatial Randomness*).

Le processus de Poisson joue un rôle central en statistiques spatiales, à la fois parce que c'est le plus simple donc celui dont les propriétés ont été le mieux étudiées, et aussi parce qu'il constitue généralement le modèle nul contre lequel des semis de points peuvent être testés. C'est un processus à accroissements indépendants, il joue le rôle des marches aléatoires pour les séries temporelles à temps discret et du mouvement Brownien pour les séries à temps continu. Ces propriétés sont utilisées pour le calcul de l'intervalle de confiance de la fonction K de Ripley.

L'espérance du nombre de points dans A est $\lambda|A|$ et suit une loi de Poisson. Un processus homogène et indépendant dont le nombre de points est fixé dans A est un processus binomial.⁹

⁸DIGGLE (1983). *Statistical analysis of spatial point patterns*, cf. note 2, p. 19, pages 51-52.

⁹D. STOYAN et al. (1987). *Stochastic Geometry and its Applications*. New York : John Wiley & Sons, 345 p. page 36.

4.1.4 Processus de Poisson inhomogène

Le processus de Poisson hétérogène ou inhomogène est une extension du processus de Poisson homogène dans laquelle l'intensité n'est pas constante (Figure 4.1b).

Tout processus dont les points sont indépendants et dont l'intensité est une fonction de chaque point $\lambda(x)$ est un processus de Poisson hétérogène.¹⁰

¹⁰DIGGLE (1983). *Statistical analysis of spatial point patterns*, cf. note 2, p. 19.

4.1.5 Autres Processus

Les processus classiques permettent de générer des semis de points dont les propriétés de premier et de second ordre varient. Ils sont seulement rapidement résumés ici pour mémoire. Une présentation détaillée est disponible dans tous les manuels de statistiques spatiales.

¹¹D. R. COX (1955). « Some Statistical Methods Connected with Series of Events ». In : *Journal of the Royal Statistical Society B* 17.2, p. 129–164.

¹²B. MATÉRN (1960). « Spatial variation ». In : *Meddelanden från Statens Skogsforskningsinstitut* 49.5, p. 1–144.

¹³M. THOMAS (1949). « A Generalization of Poisson's Binomial Limit for Use in Ecology ». In : *Biometrika* 36.1/2, p. 18–25.

FIGURE 4.2 – Processus de Gibbs.^a Trois semis de points (en haut) obtenus par trois fonctions différentes $u(r)$ – notées $f(r)$ en bas : (i) un semis de points régulier, (ii) un semis attractif à courte distance, puis répulsif, (iii) un semis constitué d'agrégats régulièrement répartis.

^aFigure in F. GOREAUD (2000). « Apports de l'analyse de la structure spatiale en forêt tempérée à l'étude de la modélisation des peuplements complexes ». Thèse de doct. Nancy : EN-GREF, page 42.

¹⁴MÖLLER et WAAGEPETERSEN (2004). *Statistical Inference and Simulation for Spatial Point Processes*, cf. note 1, p. 17.

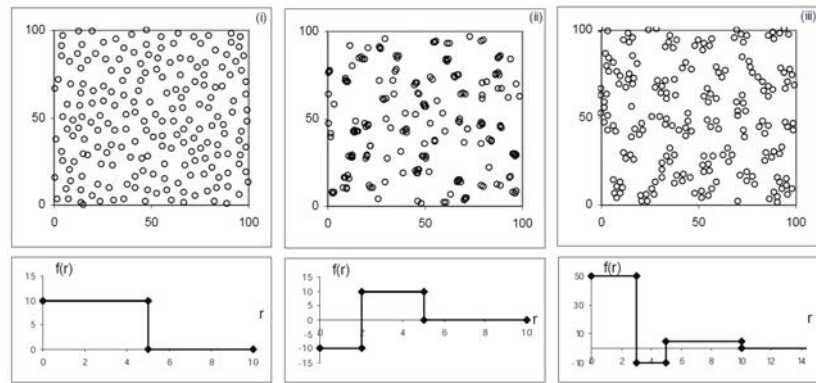
¹⁵E. TOMPPA (1986). *Models and methods for analysing spatial patterns of trees*. T. 138. Helsinki, Finland : The Finnish forest research institute, p. 1–65.

¹⁶É. MARCON (2010). « Statistiques spatiales avec applications à l'écologie et à l'économie ». Thèse de doct. AgroParisTech.

Processus de Cox Ce sont des processus de Poisson hétérogènes introduits par Cox¹¹ dont la densité est une variable aléatoire $\Lambda(x)$ et non une fonction déterministe.

Parmi eux, le processus Matérn¹² est obtenu en tirant dans un premier temps un ensemble de points intermédiaires (appelés centres) dans un processus de Poisson. Ensuite, un second processus de Poisson génère les points définitifs autour de chaque centre, dans un cercle de rayon fixé. Le résultat obtenu est un ensemble d'agrégats de points de même intensité autour des centres distribués totalement aléatoirement.

Le processus de Thomas¹³ est similaire, mais les points définitifs sont tirés dans un processus de Poisson inhomogène dont l'intensité suit une loi normale à deux dimensions autour de chaque centre.



Processus de Gibbs Les processus de Gibbs (ou de Markov) sont des processus dans lesquels les points ne sont pas indépendants, contrairement aux deux cas précédents. Ils peuvent être définis de façon globale, rigoureuse mais peu intuitive,¹⁴ ou de façon locale,¹⁵ plus simple à comprendre. La présentation locale est donnée ici ; le lien entre les deux définitions est détaillé dans ma thèse.¹⁶

On mesure pour un semis de points dont le nombre est fixé son énergie totale définie par :

$$U(X_A) = \sum_{x_i \neq x_j \in X_A} u(\|x_i - x_j\|) \quad (4.11)$$

$u(\|x_i - x_j\|)$ est une fonction dépendant seulement de la distance entre les paires de points, appelée *potentiel de paire* :

- Au-delà d'un seuil R , $u(r) = 0$. Seuls les points voisins interagissent ;
- $u(r) = \infty$ pour $r \leq R$ interdit l'existence de paires de points à distance inférieure au seuil choisi R (processus hard-core) ;
- $u(r) = \beta \in \mathbb{R}_+^*$ pour $r \leq R$ crée une répulsion entre les points voisins. Au final, le semis de point est régulier ;

- $u(r) = \beta \in \mathbb{R}_*$ pour $r \leq R$ crée une attraction entre les points voisins. Sans précaution, tous les points vont se superposer ;
- La fonction $u(r)$ peut prendre diverses valeurs qui changent à plusieurs reprises, définissant plusieurs seuils R_i .

Les processus de Gibbs sont très versatiles et permettent de simuler diverses configurations spatiales (Figure 4.2).

4.1.6 Vocabulaire

Un semis de point observable, ou distribution, est la réalisation d'un processus ponctuel.

Une distribution est dite *homogène* si son intensité est constante sur l'aire d'étude et si les relations de dépendance entre les points le sont aussi. Elle est *indépendante* si la position d'un point ne dépend pas de la position des autres. En cas de dépendance, celle-ci est toujours stationnaire dans les cas traités par la littérature.

Une distribution peut être *complètement aléatoire*, c'est-à-dire homogène (propriété de premier ordre) et indépendante (propriété de second ordre), si chaque point est distribué avec une probabilité indépendante du lieu et indépendamment des autres. En d'autres termes, il s'agit d'une distribution de Poisson homogène (Figure 4.1a).

Les points peuvent s'attirer, donnant des agrégats. On pourra parler de *concentration spatiale* ou d'*agglomération*. L'intensité de points sera localement plus grande, sans que ce ne soit contradictoire avec l'hypothèse éventuelle d'homogénéité : une autre réalisation du même processus aurait donné des agrégats à d'autres emplacements, et l'intensité locale mesurée sur plusieurs réalisations du processus aurait été constante.

Les points peuvent se repousser, générant la dispersion. Les forces de dispersion créent des distributions *régulières* dans lesquelles les points ont tendance à se situer à égale distance les uns des autres.

Les irrégularités dans une distribution (par exemple des agrégats) peuvent être dues à sa propriété de premier ordre (les agrégats sont la réalisation d'un processus ponctuel dont l'intensité est plus grande) ou de deuxième ordre (les points s'attirent). Il est impossible de trancher à partir d'un jeu de données¹⁷ : plusieurs réalisations du processus ponctuel sont nécessaires, mais rarement disponibles dans la réalité.

Les processus peuvent être *marqués*, ce qui signifie qu'à chaque point est associée une variable aléatoire. Si la variable est discrète (par exemple l'espèce pour des points représentant des arbres), on l'appelle souvent *étiquette*. Le processus peut alors être partitionné en sous-ensembles dont chacun correspond à une valeur de l'étiquette. La variable peut aussi être continue (par exemple, la

¹⁷W. FELLER (1943). « On a general class of contagious distributions ». In : *The Annals of Mathematical Statistics* 14, p. 389–400 ; G. ELLISON et E. L. GLAESER (1997). « Geographic Concentration in U.S. Manufacturing Industries : A Dartboard Approach ». In : *Journal of Political Economy* 105.5, p. 889–927.

surface terrière de chaque arbre).

4.1.7 La fonction K de Ripley

Les travaux théoriques fondamentaux sur la structure des processus ponctuels¹⁸ ont mis en place un cadre théorique clair quoique limité :

- Les semis de points observés peuvent être considérés comme la réalisation de processus ponctuels ;
- La référence est un semis de points complètement aléatoire : tous les points sont distribués avec une probabilité égale partout, indépendamment les uns des autres ;
- La structure spatiale est l'écart à l'indépendance entre les points : la concentration, due à l'attraction entre les points, ou la dispersion, due à leur répulsion ;
- La significativité statistique des valeurs de concentration ou de répulsion observées est classiquement testée par la méthode de Monte-Carlo¹⁹ parce que la distribution de la statistique est inconnue.

¹⁸RIPLEY (1976). « The Foundations of Stochastic Geometry », cf. note 7, p. 14; RIPLEY (1977). « Modelling Spatial Patterns », cf. note 7, p. 14.

¹⁹J. E. BESAG et P. J. DIGGLE (1977). « Simple Monte Carlo Tests for Spatial Pattern ». In : *Applied Statistics* 26.3, p. 327-333.

Définition

Ripley²⁰ a défini la fonction K :

$$K(r) = \int_0^r g(\rho) 2\pi\rho d\rho \quad (4.12)$$

Si les points sont distribués indépendamment les uns des autres (processus de Poisson homogène), $g(\rho)$ vaut 1 pour toutes les valeurs de ρ , et $K(r) = \pi r^2$. Cette valeur sert de référence :

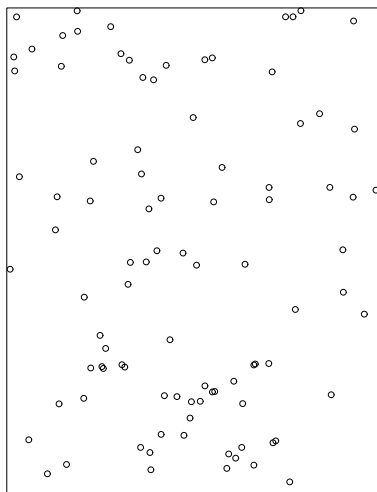
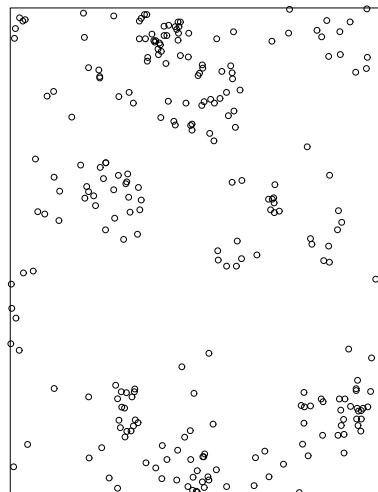
- $K(r) > \pi r^2$ indique qu'en moyenne $g(\rho)$ est supérieur à 1. La probabilité de trouver un voisin à la distance ρ est donc supérieure à la probabilité de trouver un point dans un lieu quelconque du domaine d'étude : les points sont agglomérés ;
- Inversement, $K(r) < \pi r^2$ indique que la densité de voisins autour des points est moins grande que la densité moyenne sur l'ensemble du domaine d'étude. Les points se repoussent.

Les fonctions g et K peuvent être bivariées (les économistes disent aussi intertypes) si le processus ponctuel est marqué. Dans ce cas, une partie des points appartient au type de référence noté R , dont le sous-ensemble des points dans la fenêtre A est noté R_A , une autre au type des voisins N , $N_A = N \cap A$.

Estimation

$\lambda K(r)$ est aussi l'espérance du nombre de voisins situés à distance inférieure ou égale à r d'un point quelconque. Un estimateur de K sans biais est obtenu en calculant le nombre moyen de voisins de chaque point et en le normalisant par l'estimateur de l'intensité (le nombre de points observés N divisé par la surface de A).²¹ Pour

²¹Par exemple : LANG et MARCON (2013). « Testing randomness of spatial point patterns with the Ripley statistic », cf. note 12, p. 15.

(a) Carte des *Tachigali melinonii* (94 arbres).(b) Carte des *Dicorynia guianensis* (254).FIGURE 4.3 – Semis de points. Carte des arbres de plus de 10 centimètres de diamètre d'une parcelle de Paracou de $400,6 \times 522,3$ mètres.

chaque point de référence x_i , un autre point x_j est par définition un voisin si la distance $\|x_i - x_j\|$ est inférieure ou égale à r . Les points dont la distance au bord de la fenêtre est inférieure à r posent des problèmes d'effets de bord : une partie de leurs voisins n'est pas détectée. Les méthodes de correction des effets de bord sont nombreuses²² et sans intérêt particulier ici. Pour chaque paire de points x_i et x_j , la valeur de la correction est notée $c(x_i, x_j, r)$. Une correction possible est le rapport entre la surface du disque de rayon r autour du point x_i , notée $\|b(x_i, r)\|$ et celle de sa partie incluse dans la fenêtre $\|b(x_i, r) \cap A\|$, qui ne dépend dans ce cas pas de x_j .

L'estimateur de K est finalement :

$$\hat{K}(r) = \frac{\|A\|}{N(N-1)} \sum_{x_i \neq x_j \in X_A} \mathbf{1}(\|x_i - x_j\| \leq r) c(x_i, x_j, r) \quad (4.13)$$

L'estimateur de K bivarié est :

$$\hat{K}_{R,N}(r) = \frac{\|A\|}{N_R N_N} \sum_{x_i \in R_A, x_j \in N_A} \mathbf{1}(\|x_i - x_j\| \leq r) c(x_i, x_j, r) \quad (4.14)$$

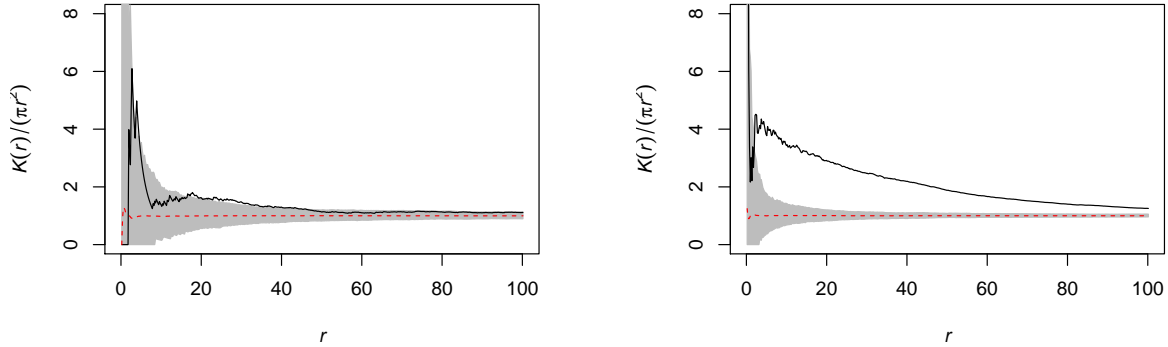
N_R et N_N sont les nombres de points du type de référence et du type des voisins.

4.1.8 Exemple

L'exemple traité est une parcelle du dispositif forestier permanent de Paracou.²³ Les arbres de plus de 10 centimètres de diamètre à hauteur de poitrine y sont cartographiés avec une incertitude

²²J. ILLIAN et al. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Chichester : Wiley-Interscience, p. 1–534, Section 4.2.2.

²³S. GOURLET-FLEURY et al. (2004). *Ecology & management of a neotropical rainforest. Lessons drawn from Paracou, a long-term experimental research site in French Guiana*. Paris.



(a) *Tachigali melinonii*. Les Tachigalis semblent avoir une structure agrégative : la valeur de $K(r)/\pi r^2$ est supérieure à 1 à toutes les distances. Mais la courbe sort difficilement de l'enveloppe correspondant à 99% des réalisations d'un processus de Poisson homogène.

(b) *Dicorynia guianensis*. La courbe de K est nettement au dessus de la valeur 1 correspondant à l'hypothèse nulle, hors de l'intervalle de confiance. On peut conclure sans risque que les Angéliques sont agrégées.

FIGURE 4.4 – Valeur de la fonction K de Ripley, normalisée par πr^2 pour les deux semis de points de la figure 4.3. Les distances sont en mètres. La courbe noire pleine représente la valeur de $K(r)$. L'enveloppe grisée représente les intervalles de confiance de $K(r)$ à 99% à chaque distance, sous l'hypothèse nulle d'une distribution complètement aléatoire. Les intervalles de confiance sont calculés au seuil de 1% par la méthode de Monte-Carlo, à partir de 1000 simulations d'un processus de Poisson homogène.

inférieure à 50 cm. Les Tachigalis (*Tachigali melinonii*) et les Angéliques (*Dicorynia guianensis*) sont représentés en figure 4.3.

La fonction K est calculée sur les deux semis de points (Figure 4.4). La normalisation par πr^2 n'est pas la présentation classique de la littérature (la fonction $L(r) = \sqrt{K/\pi}$ de Besag²⁴ est généralement utilisée) mais a l'avantage de rendre les valeurs de K interprétables : $K(r)/\pi r^2$ est le rapport entre le nombre de voisins observés et le nombre de voisins attendus. Cette normalisation entre dans un cadre plus général développé plus loin.²⁵

L'agrégation des Angéliques ne fait aucun doute : jusqu'au delà de 100 m, la courbe de K se trouve au dessus de la borne supérieure de l'intervalle de confiance de l'hypothèse nulle. A très petite distance, la valeur de K est supérieure à 1 mais le faible nombre de voisins diminue la puissance du test et empêche de conclure.

Le tableau est moins net pour les Tachigalis dont la courbe ne sort de l'enveloppe que légèrement entre 18 et 40 m.

4.2 Tests statistiques

La méthode classique, dite de Monte-Carlo,²⁶ pour tester un jeu de points contre l'hypothèse nulle qu'il soit la réalisation d'un processus connu (souvent, un processus de Poisson homogène),

²⁴BESAG (1977). « Comments on Ripley's paper », cf. note 8, p. 14.

²⁵MARCON et PUECH (2015a). « A Typology of Distance-Based Measures of Spatial Concentration », cf. note 9, p. 14.

²⁶BESAG et DIGGLE (1977). « Simple Monte Carlo Tests for Spatial Pattern », cf. note 19, p. 24.

consiste à simuler un grand nombre de réalisations de ce processus et à calculer la statistique d'intérêt (par exemple la fonction K) à toutes les distances choisies. À chaque distance, la valeur de la statistique observée est comparée aux quantiles des valeurs simulées pour fournir une probabilité de rejeter l'hypothèse nulle par erreur. Au seuil de risque α (par exemple 5%), si une valeur de $K(r)$ sort de l'intervalle de confiance de l'hypothèse nulle fourni par les quantiles $\alpha/2$ et $1 - \alpha/2$ des simulations, l'hypothèse nulle est rejetée.

Ce test augmente en réalité le risque de rejeter l'hypothèse nulle par erreur²⁷ (erreur de type I) parce qu'il est répété à chaque distance utilisée pour le calcul. Le résultat peu tranché de la figure 4.4a peut être mis en doute pour cette raison. Ce problème est atténué par la très forte dépendance des valeurs de K entre elle puisque K est une fonction cumulative : la valeur de $\lambda K(r + dr)$ est essentiellement la valeur de $\lambda K(r)$, à laquelle ne s'ajoutent que le nombre de voisins nouveaux apportés par l'augmentation du rayon du cercle. L'inflation de l'erreur de type I apportée par ce test a été étudiée quantitativement par Loosmore et Ford.²⁸

4.2.1 Test analytique de K

Après plus de trente ans d'utilisation intensive de la fonction K de Ripley, devenue le fondement de l'analyse statistique des processus ponctuels,²⁹ aucun test statistique d'un jeu de point contre l'hypothèse nulle d'une distribution complètement aléatoire n'était disponible. Un intervalle de confiance approximatif de K pour un processus de Poisson avait été proposé par Ripley,³⁰ réfuté par Koen³¹ avec des erreurs rectifiées par Chiu.³² Ward et Ferrandino³³ avaient proposé un estimateur erroné de la variance de K en négligeant la dépendance entre les paires de points (dans un processus de Poisson, les points sont distribués indépendamment les uns des autres, mais chaque point intervient dans plusieurs paires). Heinrich³⁴ avait également proposé des tests de qualité d'ajustement fondés sur la variance asymptotique de K , mais inutilisables sur des données réelles pour lesquelles la variance exacte n'était pas connue.

J'ai développé ce test avec Gabriel Lang.³⁵ Il est présenté ici, appliqué à un exemple forestier.

Cadre mathématique

L'approche originale qui a permis de mettre au point ce test a consisté à ne pas corriger l'estimation de K pour les effets de bord. La distribution des effets de bord de K n'est calculable que pour un processus connu (le processus de Poisson de l'hypothèse nulle, pas celui qui a généré le jeu de points observé). De plus, la correction des effets de bord appliquée à chaque point observé

²⁷G. DURANTON et H. G. OVERMAN (2005). « Testing for Localisation Using Micro-Geographic Data ». In : *Review of Economic Studies* 72.4, p. 1077–1106.

²⁸N. B. LOOSMORE et E. D. FORD (2006). « Statistical inference using the G or K point pattern spatial statistics ». In : *Ecology* 87.8, p. 1925–1931.

²⁹B. D. RIPLEY (1981). *Spatial statistics*. New York : John Wiley & Sons, p. 1–255 ; DIGGLE (1983). *Statistical analysis of spatial point patterns*, cf. note 2, p. 19 ; STOYAN et al. (1987). *Stochastic Geometry and its Applications*, cf. note 9, p. 21 ; CRESSIE (1993). *Statistics for spatial data*, cf. note 5, p. 20 ; ILLIAN et al. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*, cf. note 22, p. 25.

³⁰B. D. RIPLEY (1979). « Tests of 'randomness' for spatial point patterns ». In : *Journal of the Royal Statistical Society B* 41.3, p. 368–374.

³¹C. KOEN (1991). « Approximate confidence bounds for Ripley's statistic for random points in a square ». In : *Biometrical Journal* 33, p. 173–177.

³²S. N. CHIU (2007). « Correction to Koen's critical values in testing spatial randomness ». In : *Journal of Statistical Computation and Simulation* 77.11–12, p. 1001–1004.

³³J. S. WARD et F. J. FERRANDINO (1999). « New derivation reduces bias and increases power of Ripley's L index ». In : *Ecological Modelling* 116.2–3, p. 225–236.

³⁴L. HEINRICH (1991). « Goodness-of-fit tests for the second moment function of a stationary multidimensional poisson process ». In : *Statistics : A Journal of Theoretical and Applied Statistics* 22.2, p. 245–268.

³⁵LANG et MARCON (2013). « Testing randomness of spatial point patterns with the Ripley statistic », cf. note 12, p. 15 ; MARCON et al. (2013). « A Statistical Test for Ripley's Function Rejection of Poisson Null Hypothesis », cf. note 12, p. 15.

³⁶J. GIGNOUX et al. (1999). « Comparing the performances of Diggle's test of spatial randomness for small samples with or without edge effect correction : application to ecological data ». In : *Biometrics* 55.1, p. 156–164.

revient à ajouter des voisins distribués uniformément, ce qui réduit la puissance du test, problème déjà montré pour d'autres méthodes non paramétriques appliquées aux processus ponctuels.³⁶ La valeur estimée de K à partir des données est comparée à la distribution de l'estimateur de K non corrigé des effets de bord pour un processus de Poisson homogène. L'espérance et la variance de cet estimateur sont calculées, et il est montré que la distribution est asymptotiquement normale, ce qui permet l'application de tests statistiques classiques. Les effets de bord interviennent dans le calcul de l'espérance et de la variance de l'estimateur : le calcul de la variance est très lourd même pour une forme de fenêtre simple, mais possible pour le processus de Poisson.

Formellement, la fonction K est estimée pour le jeu de points observé dans une fenêtre rectangulaire de taille $l \times w$, sans correction des effets de bord. À la distance r :

$$\hat{K}(r) = \frac{lw}{N(N-1)} \sum_{x_i \neq x_j \in X_A} \mathbf{1}(\|x_i - x_j\| \leq r) \quad (4.15)$$

L'espérance de K sous l'hypothèse nulle est πr^2 . Les effets de bord font que l'estimateur non corrigé est biaisé. Pour le processus de Poisson, le biais est :

$$B(r) = \frac{4r^3(l+w)}{3lw} + \frac{r^4}{2l^2w^2} \quad (4.16)$$

L'estimateur de K peut être corrigé de ce biais pour obtenir un vecteur de longueur d , pour toutes les valeurs de r choisies :

$$\hat{\mathbf{K}} = \left(\hat{K}(r_1) - B(r_1), \hat{K}(r_2) - B(r_2), \dots, \hat{K}(r_d) - B(r_d) \right) \quad (4.17)$$

Pour un processus de Poisson homogène, $\hat{\mathbf{K}}$ est asymptotiquement normal, d'espérance nulle et sa matrice de variance est $\hat{\Sigma}$:

$$\hat{\Sigma} = \begin{pmatrix} \text{Var}(\hat{K}(r_1)) & \cdots & \text{cov}(\hat{K}(r_1), \hat{K}(r_d)) \\ \vdots & \ddots & \vdots \\ \text{cov}(\hat{K}(r_1), \hat{K}(r_d)) & \cdots & \text{Var}(\hat{K}(r_d)) \end{pmatrix} \quad (4.18)$$

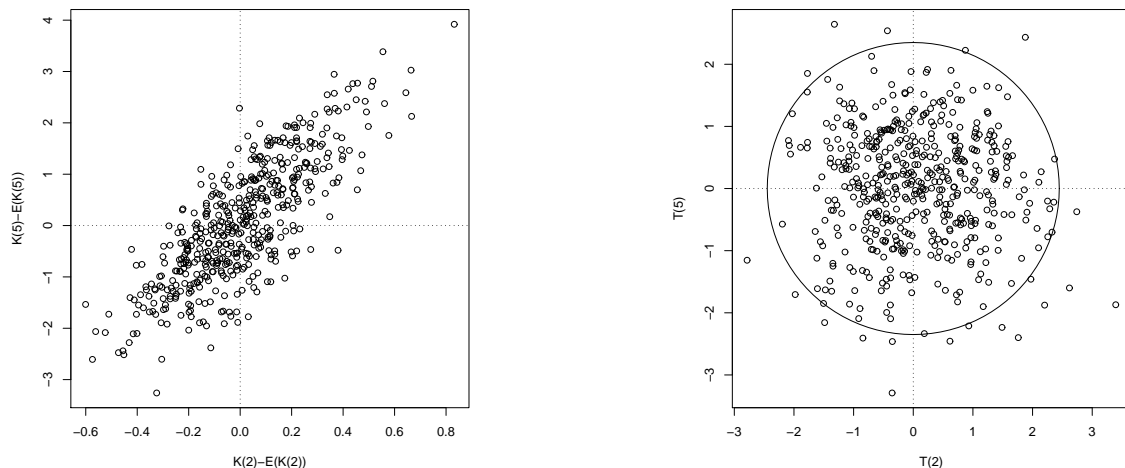
Par conséquent, $T^2 = \|\Sigma^{-1/2}(\hat{\mathbf{K}} - \pi r^2)\|$ suit une loi de χ^2 à d degrés de liberté. La normalité est atteinte dès quelques dizaines de points. Les valeurs exactes de la variance et de la covariance doivent être calculées, selon les formules fournies³⁷ implémentées dans le package *dbmss*³⁸ pour R.

Interprétation graphique

Les valeurs de $K(r)$ sont très corrélées entre elles. La figure 4.5a représente $\hat{K}(5)$ en fonction de $\hat{K}(2)$ après déduction de leur espérance (πr^2) pour 500 tirages d'un processus de Poisson d'intensité

³⁷MARCON et al. (2013). « A Statistical Test for Ripley's Function Rejection of Poisson Null Hypothesis », cf. note 12, p. 15, annexe 1.

³⁸MARCON et al. (2015). « Tools to Characterize Point Patterns : dbmss for R », cf. note 14, p. 15.



(a) Valeurs de $\hat{\mathbf{K}} - \pi r^2$ en deux dimensions ($r = 2$ et $r = 5$) pour 500 simulations d'un processus de Poisson homogène d'intensité $\rho = 5$ dans une fenêtre carrée de côté 10 (500 points en moyenne). Les valeurs de $\hat{\mathbf{K}}$ sont très corrélées entre elles.

(b) Comparaison des valeurs de $T(2)$ et $T(5)$ après transformation. Environ 25 simulations sur 500 ont une valeur de T^2 supérieure à la valeur critique $\chi^2_{5\%}(2)$ et sont donc rejetées par le test.

FIGURE 4.5 – Interprétation graphique du test de K

10. En pratique, K est estimé pour d valeurs de r et le nuage de points serait très similaire en d dimensions. Le test consiste à centrer les valeurs de K et à les décorréliser en multipliant $\hat{\mathbf{K}} - \pi r^2$ par la matrice de variance à la puissance $-1/2$ pour obtenir un vecteur \mathbf{T} dont les valeurs sont normales, centrées, réduites et indépendantes (Figure 4.5b). Le test revient à tester \mathbf{T} contre le vecteur nul. Sa norme suit une loi de χ^2 .

Application

Le test statistique est appliqué aux Angéliques et Tachgalis de la figure 4.4.

L'application du test aux données nécessite de choisir un vecteur de distances pour les valeurs de r . Aucune interaction entre les arbres n'est attendue au-delà de 150 m. 15 valeurs régulièrement réparties de 10 à 150 m sont un bon choix : trop peu de valeurs masquent des détails et trop de valeurs entraînent des problèmes numériques pour l'inversion de la matrice de variance. Le test retourne une p-value de 0 pour les Angéliques (c'est-à-dire une p-value plus faible que la plus petite valeur numérique supportée par R) et 2,5% pour les Tachigalis, qui précise les résultats incertains obtenus par la méthode de Monte-Carlo. L'hypothèse nulle d'une distribution complètement aléatoire est donc rejetée pour les deux jeux de points, au seuil de 2,5% seulement pour les Tachigalis.

4.3 Généralisation de la fonction de Ripley

4.3.1 La fonction M

La fonction K de Ripley a été définie comme l'intégrale sur un disque de la fonction g , interprétable intuitivement comme la moyenne de la propriété de second ordre du processus ponctuel. Cette interprétation est plus claire en normalisant K par la surface du disque. La valeur de référence de $K(r)/\pi r^2$ est 1 : en moyenne, on trouve autant de voisins autour d'un point que n'importe où dans la fenêtre.

La fonction K est limitée aux processus homogènes pour des raisons techniques : l'intégration de g se fait sans pondération. La fonction K_{inhom} ³⁹ a réglé ce problème en donnant à chaque point un poids inversement proportionnel à l'intensité du processus autour de lui. Lorsque le processus est localement plus intense, plus de points contribuent au calcul mais avec un poids plus faible qui permet, heuristiquement parlant, de donner le même poids à chaque portion de l'espace. La question du contrôle de l'hétérogénéité de premier ordre lors de l'estimation de l'attraction ou de la répulsion des points est donc réglée sur le plan théorique.

Sur le plan pratique, un nouveau problème apparaît : l'estimation locale de la densité nécessite l'utilisation de noyaux. Le choix de la taille du noyau influe radicalement sur les résultats obtenus : un noyau étroit estime la densité très localement, avec une forte variabilité. La distribution du jeu de points est alors largement décrite par la variation de la densité sans prise en compte des interactions entre points. Inversement, un noyau très large lisse la densité, et laisse aux interactions de second ordre la responsabilité de la variabilité. Le choix du noyau doit donc être guidé par des considérations théoriques, par exemple des connaissances sur les processus écologiques, faute de quoi les résultats sont arbitraires.⁴⁰

La question de la prise en compte de l'hétérogénéité a été posée complètement différemment en économie. L'axiomatique⁴¹ (les économistes parlent de « bonnes propriétés ») des mesures de structuration spatiale requiert de contrôler « la tendance générale de l'activité économique à s'agglomérer ». Cette idée a été formalisée clairement par Brülhart et Traeger⁴² dans le cadre un peu différent des mesures de concentration en espace discret (comme l'indice de Gini⁴³). Elle peut être formulée de la façon suivante.

Les fonctions K et g comparent la densité de voisins, c'est-à-dire leur nombre *par unité de surface* à la valeur attendue si les points étaient distribués indépendamment les uns des autres. Cette approche est appelée « topographique » parce que le nombre de voisins est normalisé par la surface.

Le nombre de voisins peut être comparé à une autre grandeur que la surface. Le nombre de plantes voisines d'une espèce donnée

³⁹A. J. BADDELEY et al. (2000). « Non- and semi-parametric estimation of interaction in inhomogeneous point patterns ». In : *Statistica Neerlandica* 54.3, p. 329–350.

⁴⁰ILLIAN et al. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*, cf. note 22, p. 25, Section 3.3.

⁴¹P.-P. COMBES et H. G. OVERMAN (2004). « The spatial distribution of economic activities in the European Union ». In : *Handbook of Urban and Regional Economics*. Sous la dir. de J. V. HENDERSON et J.-F. THISSE. T. 4. Amsterdam : Elsevier. North Holland. Chap. 64, p. 2845–2909 ; DURANTON et OVERMAN (2005). « Testing for Localisation Using Micro-Geographic Data », cf. note 27, p. 27.

⁴²M. BRÜLHART et R. TRAEGER (2005). « An Account of Geographic Concentration Patterns in Europe ». In : *Regional Science and Urban Economics* 35.6, p. 597–624.

⁴³C. GINI (1912). *Variabilità e mutabilità*. Bologna : C. Cuppini ; 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.

peut être divisé par le nombre total de plantes voisines, de la même façon que l'indice de Gini prend en compte la taille de l'industrie d'un secteur économique donné rapportée à la taille totale de l'industrie pour fournir un poids relatif de ce secteur dans chaque région et intègre l'écart de ce poids relatif local au poids relatif global (sur l'ensemble des régions confondues) pour fournir une statistique globale de la concentration du secteur. Cette approche est appelée « relative ». Appliquée aux statistiques spatiales, les points doivent disposer d'une marque discrète et la proportion moyenne de voisins d'un type particulier autour des points de référence remplace la densité de voisins. Cette proportion locale est normalisée par la proportion globale pour donner une valeur attendue de 1 si les points étaient distribués indépendamment.

Une complication supplémentaire est apportée par la possible pondération des points, portée par une marque continue. Si les objets traités sont des établissements industriels, la distribution de leurs tailles, appelée concentration industrielle, est considérée comme exogène.⁴⁴ La concentration spatiale est évaluée en prenant en compte le poids de chaque établissement (mesuré généralement par son nombre d'employés), mais la mesure utilisée doit « contrôler la concentration industrielle ».⁴⁵

Formellement, la fonction M^{46} permet de traiter la question jusqu'à la distance r choisie :

$$\hat{M}(r) = \frac{\sum_{x_i \in R_A} \frac{\sum_{x_j \neq x_i \in N_A} \mathbf{1}(\|x_i - x_j\| \leq r) w(x_j)}{\sum_{x_j \neq x_i \in X_A} \mathbf{1}(\|x_i - x_j\| \leq r) w(x_j)}}{\sum_{x_i \in R_A} \frac{W_N - w(x_i)}{W - w(x_i)}} \quad (4.19)$$

R_A est l'ensemble des points de référence, N_A celui des voisins, dans la fenêtre A . $w(x)$ est le poids du point x . W_N est le poids total des voisins, W le poids total de tous les points. La proportions des voisins appartenant au secteur d'intérêt N est calculée autour de chaque point du secteur de référence R (numérateur de l'équation) et moyenné (la division par le nombre de points de R_A n'apparaît pas parce qu'elle se simplifie avec le dénominateur). Cette proportion est comparée au poids relatif du secteur N sur l'ensemble du jeu de données, qui n'est pas exactement W_N/W parce que chaque point de référence doit être retiré du poids des voisins.

À la distance r , la fonction m^{47} a le même objectif :

$$\hat{m}(r) = \frac{\sum_{x_i \in R_A} \frac{\sum_{x_j \neq x_i \in N_A} k(\|x_i - x_j\|, r) w(x_j)}{\sum_{x_j \neq x_i \in X_A} k(\|x_i - x_j\|, r) w(x_j)}}{\sum_{x_i \in R_A} \frac{W_N - w(x_i)}{W - w(x_i)}} \quad (4.20)$$

Les types N et R peuvent être identiques ou différents, définissant les fonctions univariées ou bivariées. L'intervalle de confiance de M ou m sous l'hypothèse nulle d'indépendance entre les points

⁴⁴ELLISON et GLAESER (1997). « Geographic Concentration in U.S. Manufacturing Industries : A Dartboard Approach », cf. note 17, p. 23.

⁴⁵DURANTON et OVERMAN (2005). « Testing for Localisation Using Micro-Geographic Data », cf. note 27, p. 27.

⁴⁶MARCON et PUECH (2010). « Measures of the Geographic Concentration of Industries : Improving Distance-Based Methods », cf. note 6, p. 14.

⁴⁷G. LANG et al. (2015). « Distance-Based Measures of Spatial Concentration : Introducing a Relative Density Function ». In : HAL 01082178.version 3, p. 1–14.

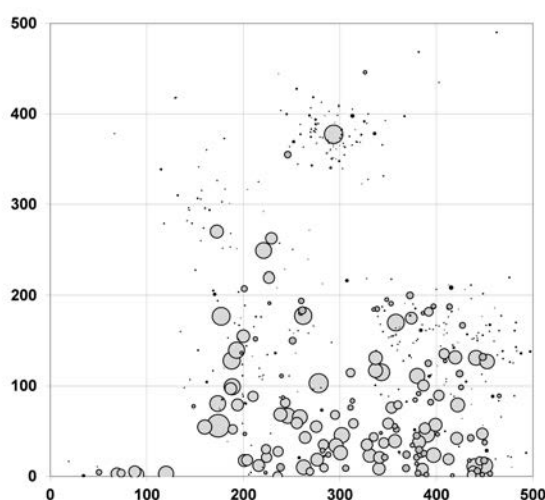
⁴⁸DURANTON et OVERMAN (2005). « Testing for Localisation Using Micro-Geographic Data », cf. note 27, p. 27; MARCON et PUECH (2010). « Measures of the Geographic Concentration of Industries : Improving Distance-Based Methods », cf. note 6, p. 14.

est construite en redistribuant les points marqués (avec leur secteur et leur poids, ce qui permet de conserver la concentration industrielle) sur les emplacements existants.⁴⁸ Cette hypothèse nulle est appelée « localisation aléatoire ». Elle peut être remplacée par un processus de Poisson inhomogène de même intensité que les données, mais la localisation aléatoire est plus simple à mettre en œuvre.

En absence de toute structure, la valeur attendue de M est 1.

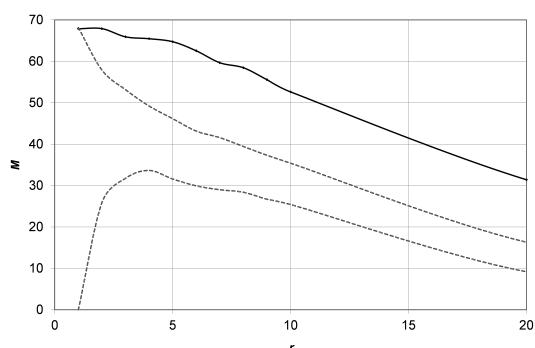
4.3.2 Application

FIGURE 4.6 – Carte de la régénération des Wacapous dans la parcelle 16 de Paracou. La carte représente un carré de 500 m de côté. Tous les Wacapous à partir d'un centimètre de diamètre sont cartographiés. La taille des cercles est proportionnelle à celle des arbres.

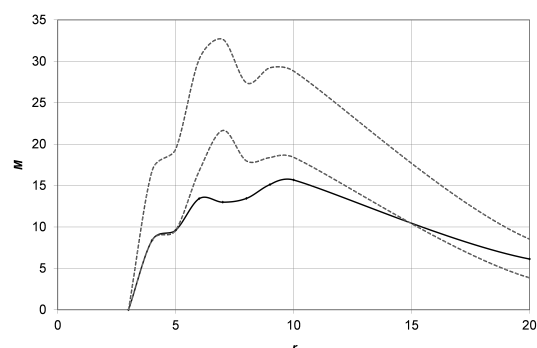


⁴⁹MARCON et al. (2012a). « Characterizing the relative spatial structure of point patterns », cf. note 6, p. 14.

L'exemple de la figure 4.6⁴⁹ teste la structure spatiale de la régénération des Wacapous (*Vouacapoua americana*) dans une parcelle de 25 ha de Paracou. La question posée concerne la structure



(a) Courbe de $M(r)$ de 0 à 20 m pour les juvéniles. La distribution est très agrégée.



(b) Distribution des juvéniles autour des arbres adultes. La répulsion est significative de 5 à 15 m environ.

FIGURE 4.7 – Structure spatiale des juvéniles de Wacapou, et interaction entre arbres adultes et juvéniles. Les courbes pleines représentent la fonction $M(r)$; les pointillés délimitent l'intervalle de confiance de M sous l'hypothèse nulle de localisation aléatoire, au seuil de 99%, obtenue à partir de 10000 simulations.

spatiale des juvéniles, et leur éventuelle attraction ou répulsion par rapport aux arbres adultes. Le jeu de points est très hétérogène, et la taille des arbres adultes est un critère important. L'approche relative se justifie pour ces deux raisons.

Les juvéniles ont une structure agrégative (Figure 4.7a). De 1 à 20 m, le poids des juvéniles (leur surface terrière) est 70 à 30 fois supérieur (valeur de M) à ce qu'il serait si la distribution spatiale des juvéniles était celle des Wacapous en général. Une partie de cet excès est dû à la distribution diamétrique (l'équivalent de la concentration industrielle en économie) : l'enveloppe de M quand la position des arbres est permutée n'est pas centrée sur 1 parce que la distribution diamétrique est très inégalitaire ; mais la courbe de M est largement au dessus de l'intervalle de confiance.

La relation entre les adultes et les juvéniles est évaluée par la fonction M bivariée (Figure 4.7b) : les points de référence sont les arbres potentiellement reproducteurs (diamètre supérieur à 30 cm), les voisins sont les juvéniles (diamètre inférieur à 10 cm). Un déficit significatif de juvéniles est détecté entre 5 et 15 mètres des reproducteurs. La puissance de test n'est pas suffisante pour l'affirmer à distance inférieure à 5 m.

4.3.3 Test de significativité

Le problème de la répétition du test local à chaque distance affecte la fonction M comme la fonction K . En absence de test analytique, un test de qualité d'ajustement⁵⁰ (*Goodness of Fit*) est possible. Le test consiste à calculer l'écart entre chaque valeur de $M(r)$ et la moyenne des valeurs simulées sous l'hypothèse nulle. L'écart peut être la valeur absolue de la différence (test de Kolmogorov-Smirnov) ou son carré (test du χ^2 de Pearson), implémenté dans le package *dbmss*. La somme (sur toutes les valeurs de r) de l'écart est la statistique de test, comparée pour les données réelles aux quantiles de ses valeurs simulées.

L'agrégation des Wacapous juvéniles est significative, avec une p-value inférieure à 0,1%. La répulsion entre juvéniles et reproducteurs est significative avec une p-value de 6,7% seulement.

Le test d'ajustement fournit une p-value mais ne donne pas d'information sur les distances auxquelles les interactions se produisent ; c'est pourquoi le tracé des intervalles de confiance reste utile.

4.4 Typologie des mesures de structure spatiale

Une typologie des fonctions non-paramétriques d'analyse de la structure de second ordre des processus ponctuels (Tableau 4.1)⁵¹ émerge selon deux critères :

⁵⁰HEINRICH (1991). « Goodness-of-fit tests for the second moment function of a stationary multidimensional poisson process », cf. note 34, p. 27.

⁵¹MARCON et PUECH (2015a). « A Typology of Distance-Based Measures of Spatial Concentration », cf. note 9, p. 14.

TABLE 4.1 – Choix de l'outil approprié pour la description d'une structure spatiale.

Choix de la fonction	Topographique, homogène	Topographique, inhomogène	Absolue	Relative
Fonctions de densité	g	g_{inhom}	K_d K^{emp}	m
Fonctions cumulatives	K	K_{inhom} D	Cumulative de K_d Cumulative de K^{emp}	M

- La référence à la surface définit les mesures *topographiques*. Les fonctions utilisées sont différentes selon que l'espace est homogène (rigoureusement, si le processus ponctuel responsable du jeu de points observé est homogène) ou non. Si la référence est différente, typiquement le poids total des voisins, les fonctions sont dites *relatives*. En absence de référence (si les voisins sont simplement comptés), les fonctions sont dites *absolues*.
- La non-indépendance entre les points peut être évaluée à une distance donnée, ou *jusqu'à* cette distance. On parle dans le premier cas de fonctions de *densité* et dans le second de fonctions *cumulatives*, par analogie avec les fonctions de probabilité et parce que la fonction K_d , la première du genre, était réellement une densité de probabilité.

Les fonctions topographiques ont déjà été présentées. La fonction D ⁵² est la différence entre deux fonctions K : celle des points d'intérêt (les cas), et celle des points de référence (les contrôles). Elle permet éventuellement de montrer que les cas sont plus ou moins concentrés que les contrôles, même si les processus concernés ne sont pas homogènes. Elle a été beaucoup utilisée avant d'être supplantée par la fonction K_{inhom} .

La fonction K_d ⁵³ a été développée en économie pour répondre aux axiomes cités plus haut. C'est simplement la densité de probabilité de trouver un point voisin (un établissement) à la distance r d'un point de référence, sans correction des effets de bord. La fonction K^{emp} ⁵⁴ pondère les établissements par leur nombre d'employés (plus généralement, les points par un poids approprié). Les cumulatives de K_d et K^{emp} ⁵⁵ sont simplement la probabilité de trouver un voisin à une distance inférieure ou égale à r . Les valeurs de ces fonctions sont comparées à leur distribution sous l'hypothèse nulle de localisation aléatoire.

Seule la fonction M mesure réellement la concentration relative.⁵⁶

⁵²P. J. DIGGLE et A. G. CHETWYND (1991). « Second-Order Analysis of Spatial Clustering for Inhomogeneous Populations ». In : *Biometrics* 47.3, p. 1155–1163.

⁵³DURANTON et OVERMAN (2005). « Testing for Localisation Using Micro-Geographic Data », cf. note 27, p. 27.

⁵⁴Ibid.

⁵⁵K. BEHRENS et T. BOUGNA (2015). « An Anatomy of the Geographical Concentration of Canadian Manufacturing Industries ». In : *Regional Science and Urban Economics* 51, p. 47–69.

⁵⁶MARCON et PUECH (2015b). « Mesures de la concentration spatiale en espace continu : théorie et applications », cf. note 9, p. 14.

CHAPITRE 5

Mesure de la diversité

L'objectif de mes recherches à ce stade est de traiter la mesure de la biodiversité, pas son intérêt ou ses implications. On se référera par exemple à Chapin *et al.*¹ pour une revue sur cette question ou Cardinale *et al.*² 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,³ 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. J'ai développé un cadre théorique permettant de définir sans ambiguïté la diversité sous tous ses aspects. J'ai également développé des estimateurs permettant d'appliquer ce cadre à des données réelles.

Comme dans le chapitre précédent, la première section présente le cadre et fait une revue de la littérature nécessaire pour la suite. Les sections suivantes introduisent mes travaux.

5.1 La diversité définie comme quantité d'information

5.1.1 Entropie et théorie de l'information

Les textes fondateurs sont Davis⁴ et surtout Theil⁵ en économétrie, et Shannon⁶ pour la mesure de la diversité. Une revue est fournie par Maasoumi.⁷

Considérons une expérience dont les résultats possibles sont $\{r_1, r_2, \dots, r_S\}$. La probabilité d'obtenir r_s est p_s , et $\mathcal{P} = \{p_1, p_2, \dots, p_S\}$. 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

¹F. S. I. CHAPIN et al. (2000). « Consequences of changing biodiversity ». In : *Nature* 405.6783, p. 234–242.

²B. J. CARDINALE et al. (2012). « Biodiversity loss and its impact on humanity ». In : *Nature* 486.7401, p. 59–67.

³A. PURVIS et A. HECTOR (2000). « Getting the measure of biodiversity. » In : *Nature* 405.6783, p. 212–9.

⁴H. T. DAVIS (1941). *The theory of econometrics*. Bloomington, Indiana : The Principia Press.

⁵THEIL (1967). *Economics and Information Theory*, cf. note 16, p. 15.

⁶SHANNON (1948). « A Mathematical Theory of Communication », cf. note 17, p. 15 ; C. E. SHANNON et W. WEAVER (1963). *The Mathematical Theory of Communication*. University of Illinois Press.

⁷E. MAASOUMI (1993). « A compendium to information theory in economics and econometrics ». In : *Econometric Reviews* 12.2, p. 137–181.

⁸PATIL et TAILLIE (1982). « Diversity as a concept and its measurement », cf. note 21, p. 15.

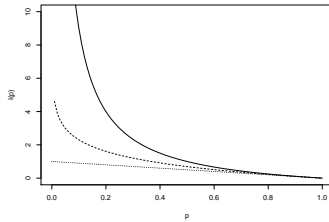


FIGURE 5.1 – Fonctions d'information utilisées dans le nombre d'espèces (trait plein), l'indice de Shannon (pointillés longs) et l'indice de Simpson (pointillés). L'information apportée par l'observation d'espèces rares décroît du nombre d'espèces à l'indice de Simpson.

⁹Ibid.

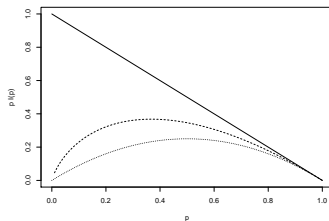


FIGURE 5.2 – Valeur de $p_s I(p_s)$ dans le nombre d'espèces (trait plein), l'indice de Shannon (pointillés longs) et l'indice de Simpson (pointillés). Les espèces rares contribuent peu, sauf pour le nombre d'espèces

¹⁰R. H. MACARTHUR (1955). « Fluctuations of Animal Populations and a Measure of Community Stability ». In : *Ecology* 36.3, p. 533–536.

¹¹R. E. ULANOWICZ (2001). « Information theory in ecology ». In : *Computers & Chemistry* 25.4, p. 393–399.

¹²A. 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.

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 Taillie⁸ appellent l'information « rareté ».

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 probabilités des espèces dans une communauté, Patil et Taillie⁹ 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.

Ces trois fonctions d'information sont représentées en Figure 5.1.

La contribution de chaque espèce à la valeur totale de l'entropie est représentée Figure 5.2.

5.1.2 Application à la biodiversité

MacArthur¹⁰ est le premier à avoir introduit la théorie de l'information en écologie.¹¹ MacArthur s'intéressait aux réseaux trophiques et cherchait à mesurer leur stabilité : l'indice de Shannon qui comptabilise le nombre de relations possibles lui paraissait une bonne façon de l'évaluer. Mais l'efficacité implique la spécialisation, ignorée dans H qui est une mesure neutre (toutes les espèces y jouent le même rôle). MacArthur a abandonné cette voie.

Les premiers travaux consistant à généraliser l'indice de Shannon sont dus à Rényi.¹² L'entropie d'ordre q de Rényi est :

$${}^qR = \frac{1}{1 - q \ln \sum_{q=1}^S p_s^q} \quad (5.1)$$

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

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

Il montre que qR respecte les 3 axiomes.

Patil et Taillie¹³ 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.¹⁴

Hill¹⁵ transforme l'entropie de Rényi en *nombre de Hill*, qui en sont simplement l'exponentielle :

$${}^qD = \left(\sum_{s=1}^S p_s^q \right)^{\frac{1}{1-q}} \quad (5.2)$$

Le souci de Hill était de rendre les indices de diversité intelligibles après l'article remarqué de Hurlbert¹⁶ intitulé « le non-concept de diversité spécifique ». Hurlbert reprochait à la littérature sur la diversité sa trop grande abstraction et son éloignement des réalités biologiques, notamment en fournissant des exemples dans lesquels l'ordre des communautés n'est pas le même selon l'indice de diversité choisi. Les nombres de Hill sont le nombre d'espèces équiprobables donnant la même valeur de diversité que la distribution observée. Ils sont des transformations simples des indices classiques :

- 0D est le nombre d'espèces ;
- ${}^1D = e^H$, l'exponentielle de l'indice de Shannon ;
- ${}^2D = 1/(1 - E)$, l'inverse de l'indice de concentration de Simpson ($1 - E = \sum_s p_s^2$), connu sous le nom d'indice de Stoddart.¹⁷

Ces résultats avaient déjà été obtenus avec une autre approche par MacArthur¹⁸ et repris par Adelman¹⁹ dans la littérature économique.

Les nombres de Hill sont des « nombres effectifs » ou « nombres équivalents ». Le concept a été défini rigoureusement par Gregorius,²⁰ d'après Wright²¹ (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.

Gregorius²² 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). Ces indices doivent vérifier deux

¹³PATIL et TAILLIE (1982). « Diversity as a concept and its measurement », cf. note 21, p. 15.

¹⁴H. DALTON (1920). « The measurement of the inequality of incomes ». In : *The Economic Journal* 30.119, p. 348-361.

¹⁵M. O. HILL (1973). « Diversity and Evenness : A Unifying Notation and Its Consequences ». In : *Ecology* 54.2, p. 427-432.

¹⁶S. H. HURLBERT (1971). « The Nonconcept of Species Diversity : A Critique and Alternative Parameters ». In : *Ecology* 52.4, p. 577-586.

¹⁷J. A. STODDART (1983). « A genotypic diversity measure ». In : *Journal of Heredity* 74, p. 489-490.

¹⁸R. H. MACARTHUR (1965). « Patterns of species diversity ». In : *Biological Reviews* 40.4, p. 510-533.

¹⁹M. 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.

²⁰H.-R. GREGORIUS (1991). « On the concept of effective number. » In : *Theoretical population biology* 40.2, p. 269-83.

²¹S. WRIGHT (1931). « Evolution in Mendelian Populations ». In : *Genetics* 16.2, p. 97-159.

²²H.-R. GREGORIUS (2014). « Partitioning of diversity : the "within communities" component ». In : *Web Ecology* 14, p. 51-60.

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 bien qu'il soit très difficile de lui trouver le moindre sens intuitif : 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 :

$$D = \frac{1}{I^{-1}\left(\sum_{s=1}^S p_s I(p_s)\right)} \quad (5.3)$$

5.1.3 Biais d'estimation

L'entropie est définie comme la somme pondérée sur toutes les espèces de l'information. Dans des systèmes très divers comme la forêt tropicale, inventorier la totalité des espèces est en général impossible. Estimer le nombre d'espèces total par le nombre d'espèces échantillonnées est évidemment incorrect : l'estimation du nombre d'espèces non observées a généré une abondante littérature.²³

Le problème ne se limite pas au nombre d'espèces. Il sera traité en détail dans la section 5.4 mais il est important pour la suite de le définir. Sa conséquence est une sous-estimation de la diversité, appelée « biais d'estimation » par Dauby et Hardy.²⁴ Ce terme est préférable à « biais d'échantillonnage », souvent utilisé, parce qu'il n'est pas lié à un échantillonnage défaillant mais simplement à une variabilité inévitable et à l'impossibilité d'augmenter indéfiniment l'effort d'échantillonnage.

Les espèces rares ont un rôle central dans le biais d'estimation parce qu'elles sont plus difficiles à observer. Les mesures de diversité qui leur donnent une grande importance (l'exemple le plus simple est la richesse spécifique) sont plus biaisées que les mesures qui ne prennent en compte que les espèces dominantes (comme l'indice de Simpson).

5.1.4 Entropie HCDT

Tsallis²⁵ propose une classe de mesures appelée entropie généralisée, définie par Havrda et Charvát²⁶ pour la première fois et redécouverte plusieurs fois, notamment par Daróczy,²⁷ d'où son nom *entropie HCDT* (voir Mendes *et al.*,²⁸ page 451, pour un historique complet) :

$${}^qH = \frac{1}{q-1} \left(1 - \sum_{s=1}^S p_s^q \right) \quad (5.4)$$

Tsallis a montré que les indices de Simpson et de Shannon étaient des cas particuliers d'entropie généralisée. Ces résultats

²³A. CHAO (2004). « Species richness estimation. » In : *Encyclopedia of Statistical Sciences*. Sous la dir. de N BALAKRISHNAN et al. 2nd ed. New York : Wiley.

²⁴G. 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.

²⁵C. TSALLIS (1988). « Possible generalization of Boltzmann-Gibbs statistics ». In : *Journal of Statistical Physics* 52.1, p. 479–487.

²⁶J. HAVRDA et F. CHARVÁT (1967). « Quantification method of classification processes. Concept of structural a-entropy ». In : *Kybernetika* 3.1, p. 30–35.

²⁷Z. DARÓCZY (1970). « Generalized information functions ». In : *Information and Control* 16.1, p. 36–51.

²⁸R. S. MENDES et al. (2008). « A unified index to measure ecological diversity and species rarity ». In : *Ecography* 31.4, p. 450–456.

ont été complétés par d'autres et repris en écologie par Keylock²⁹ et Jost.³⁰ Là encore :

- Le nombre d'espèces moins 1 est 0H ;
- L'indice de Shannon est 1H ;
- L'indice de Gini-Simpson est 2H .

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).

5.1.5 Logarithmes déformés

L'écriture de l'entropie HCDT est largement simplifiée en introduisant le formalisme des logarithmes déformés.³¹ Le logarithme d'ordre q est défini par :

$$\ln_q x = \frac{x^{1-q} - 1}{1 - q} \quad (5.5)$$

Le logarithme déformé converge vers le logarithme naturel quand $q \rightarrow 1$.

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

$$e_q^x = [1 + (1 - q)x]^{\frac{1}{1-q}} \quad (5.6)$$

Enfin, le logarithme déformé est subadditif :

$$\ln_q(xy) = \ln_q x + \ln_q y - (q - 1)(\ln_q x)(\ln_q y) \quad (5.7)$$

Ses propriétés sont les suivantes :

$$\ln_q \frac{1}{x} = -x^{q-1} \ln_q x \quad (5.8)$$

$$\ln_q(xy) = \ln_q x + x^{1-q} \ln_q y \quad (5.9)$$

$$\ln_q \left(\frac{x}{y} \right) = \ln_q x - \left(\frac{x}{y} \right)^{1-q} \ln_q y \quad (5.10)$$

Et :

$$e_q^{x+y} = e_q^x e_q^{\frac{y}{1+(1-q)x}} \quad (5.11)$$

Les calculs utilisant les logarithmes déformés sont donc plus laborieux que ceux limités aux logarithmes classiques mais ils sont efficaces notamment pour décomposer la diversité.³²

²⁹C. J. KEYLOCK (2005). « Simpson diversity and the Shannon-Wiener index as special cases of a generalized entropy ». In : *Oikos* 109.1, p. 203–207.

³⁰JOST (2006). « Entropy and diversity », cf. note 18, p. 15; JOST (2007). « Partitioning diversity into independent alpha and beta components », cf. note 19, p. 15.

³¹C. TSALLIS (1994). « What are the numbers that experiments provide? » In : *Química Nova* 17.6, p. 468–471.

³²MARCON et al. (2014a). « Generalization of the partitioning of Shannon diversity », cf. note 23, p. 15.

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

L'entropie d'ordre q s'écrit :

$${}^qH = \frac{1}{q-1} \left(1 - \sum_{s=1}^S p_s^q \right) = - \sum_s p_s^q \ln_q p_s = \sum_s p_s \ln_q \frac{1}{p_s} \quad (5.12)$$

Ces trois formes sont équivalentes mais les deux dernières s'interprètent comme une généralisation de l'entropie de Shannon.³³

³³MARCON et al. (2014a). « Generalization of the partitioning of Shannon diversity », cf. note 23, p. 15.

5.1.6 Entropie et diversité

On voit immédiatement que l'entropie de Tsallis est le logarithme d'ordre q du nombre de Hill correspondant, comme l'entropie de Rényi en est le logarithme naturel.

$${}^qH = \ln_q {}^qD \quad (5.13)$$

$${}^qD = e_q^{{}^qH} \quad (5.14)$$

L'entropie est utile pour les calculs : la correction des biais d'estimation notamment. Les nombres de Hill, ou *nombres équivalents d'espèces* ou *nombres d'espèces effectives* permettent une appréhension plus intuitive de la notion de biodiversité.³⁴ En raison de leurs propriétés, notamment de décomposition, Jost³⁵ les appelle « vraie diversité ». Hoffmann et Hoffmann³⁶ critiquent cette définition totalitaire³⁷ et fournissent une revue historique plus lointaine sur les origines de ces mesures.

Dauby et Hardy³⁸ écrivent « diversité au sens strict » ; Gregorius³⁹ « diversité explicite ». Quoi qu'il en soit, les nombres de Hill respectent le principe de réplification (voir Chao *et al.*,⁴⁰ section 3 pour une discussion et un historique) : si I communautés de même taille, de même niveau de diversité D , mais sans espèces en commun sont regroupées dans une méta-communauté, la diversité de la méta-communauté doit être $I \times D$.

L'intérêt de ces approches est de fournir une définition paramétrique de la diversité, qui donne plus ou moins d'importance aux espèces rares :

- $^{-\infty}D = 1/\min(p_S)$ est l'inverse de la proportion de la communauté représentée par l'espèce la plus rare (toutes les autres espèces sont ignorées). Le biais d'estimation est incontrôlable : l'espèce la plus rare n'est pas dans l'échantillon tant que l'inventaire n'est pas exhaustif ;
- 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

³⁴JOST (2006). « Entropy and diversity », cf. note 18, p. 15.

³⁵JOST (2007). « Partitioning diversity into independent alpha and beta components », cf. note 19, p. 15.

³⁶S. HOFFMANN et A. HOFFMANN (2008). « Is there a "true" diversity? » In : *Ecological Economics* 65.2, p. 213–215.

³⁷Jost (L. JOST [2009]. « 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).

³⁸DAUBY et HARDY (2012). « Sampled-based estimation of diversity sensu stricto by transforming Hurlbert diversities into effective number of species », cf. note 24, p. 38.

³⁹H.-R. GREGORIUS (2010). « Linking Diversity and Differentiation ». In : *Diversity* 2.3, p. 370–394.

⁴⁰A. CHAO et al. (2010). « Phylogenetic diversity measures based on Hill numbers ». In : *Philosophical Transactions of the Royal Society B* 365.1558, p. 3599–3609.

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 propre ;

- 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. Hill⁴¹ 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-Parker⁴² qui est la proportion de la communauté représentée par l'espèce la plus abondante : $d = \max(p_S)$. Toutes les autres espèces sont ignorées.

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,⁴³ 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 comme la moyenne des entropies de plusieurs communautés, ou l'entropie phylogénétique 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.

5.1.7 Profils de diversité

Hill,⁴⁴ Patil et Taillie,⁴⁵ Tothmeresz⁴⁶ et Kindt *et al.*,⁴⁷ recommandent de tracer des profils de diversité, c'est-à-dire la valeur de la diversité qD en fonction de l'ordre q (Figure 5.3) 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.⁴⁸

Liu *et al.*⁴⁹ nomment *séparables* des communautés dont les profils ne se croisent pas. Ils montrent que les communautés séparables selon le profil de la queue de distribution,⁵⁰ le sont

⁴¹HILL (1973). « Diversity and Evenness : A Unifying Notation and Its Consequences », cf. note 15, p. 37.

⁴²W. H. BERGER et F. L. PARKER (1970). « Diversity of planktonic foraminifera in deep-sea sediments ». In : *Science* 168.3937, p. 1345–1347.

⁴³R. GADAGKAR (1989). « An undesirable property of Hill's diversity index N_2 ». In : *Oecologia* 80, p. 140–141.

⁴⁴HILL (1973). « Diversity and Evenness : A Unifying Notation and Its Consequences », cf. note 15, p. 37.

⁴⁵PATIL et TAILLIE (1982). « Diversity as a concept and its measurement », cf. note 21, p. 15.

⁴⁶B. TOTHMERESZ (1995). « Comparison of different methods for diversity ordering ». In : *Journal of Vegetation Science* 6.2, p. 283–290.

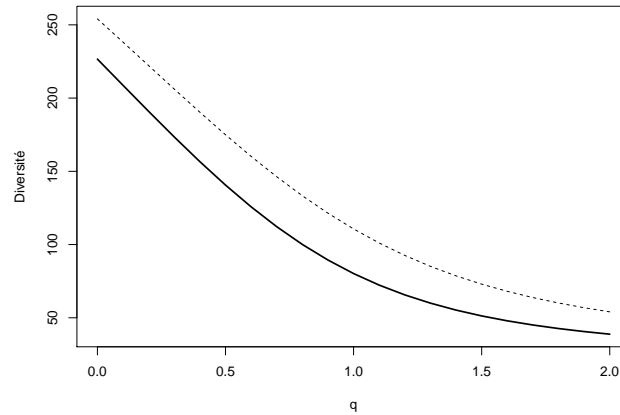
⁴⁷R. KINDT *et al.* (2006). « Tree diversity in western Kenya : Using profiles to characterise richness and evenness ». In : *Biodiversity and Conservation* 15.4, p. 1253–1270.

⁴⁸TOTHMERESZ (1995). Cf. note 46.

⁴⁹C. LIU *et al.* (2006). « Unifying and distinguishing diversity ordering methods for comparing communities ». In : *Population Ecology* 49.2, p. 89–100.

⁵⁰PATIL et TAILLIE (1982). « Diversity as a concept and its measurement », cf. note 21, p. 15.

FIGURE 5.3 – Profil de diversité calculé pour deux parcelles de Paracou (Parcelle 6 : trait plein et Parcelle 18 : trait pointillé).

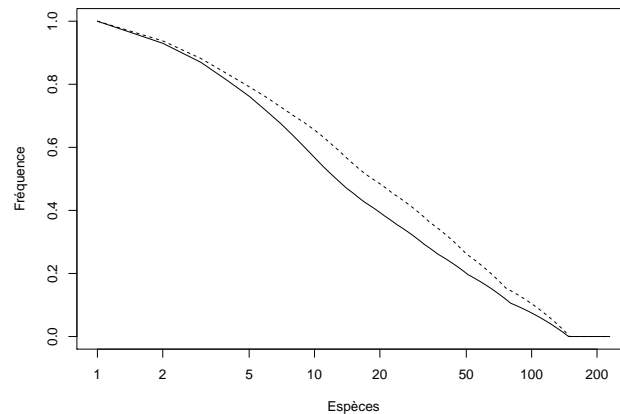


dans tous les cas. Le profil de la queue de distribution est construit en classant les espèces de la plus fréquente à la plus rare et en traçant la probabilité qu'un individu appartienne à une espèce plus rare que l'espèce en abscisse (Figure 5.4). Les coordonnées des points du profil sont définies par :

$$y(x) = \sum_{s=x+1}^S p_{[s]}, \quad x \in \{0, 1, \dots, S\} \quad (5.15)$$

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

FIGURE 5.4 – 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.

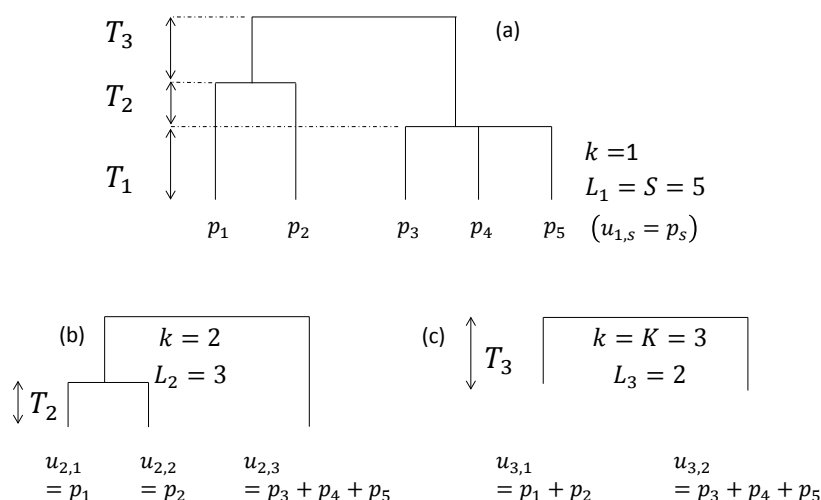


FIGURE 5.5 – Arbres phylogénétiques ou fonctionnels hypothétiques. (a) Arbre complet. 5 espèces sont présentes ($S = 5$). Une période de l'arbre est définie entre deux nœuds successifs : l'arbre contient $K = 3$ périodes. Les hauteurs des périodes sont notées T_k . À chaque période correspond un arbre 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 $u_{k,l}$.

Les données utilisées dans les figures et par la suite sont deux hectares du dispositif forestier permanent de Paracou, parcelles 6 et 18, inventoriés dans le cadre du projet ANR Bridge ⁵¹, et présentées en détail par Marcon *et al.* ⁵²

5.2 Diversité neutre, phylogénétique et fonctionnelle

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.

5.2.1 Phylodiversité

Il s'agit donc de caractériser la différence entre deux classes d'objets, puis de construire des mesures de diversité en rapport. ⁵³ En écologie, ces différences sont fonctionnelles ou phylogénétiques, définissant la diversité fonctionnelle ⁵⁴ ou la diversité phylogénétique (*phylodiversity*). ⁵⁵ Les premières propositions de ce type d'indices sont dues à Rao ⁵⁶ puis, avec nettement moins de succès, Vane-Wright *et al.* ⁵⁷ et Warwick et Clarke. ⁵⁸ Chave *et al.* ⁵⁹ 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. ⁶⁰ Les mesures présentées ici sont les plus utilisées, et notamment celles qui peuvent être

⁵¹Bridging Information on tree Diversity in French Guiana and a test of Ecological theories, ANR-06-BDIV-0004

⁵²MARCON *et al.* (2012b). « The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity », cf. note 20, p. 15.

⁵³E. C. PIELOU (1975). *Ecological Diversity*. New York : Wiley ; 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.

⁵⁴D. TILMAN *et al.* (1997). « The Influence of Functional Diversity and Composition on Ecosystem Processes ». In : *Science* 277.5330, p. 1300–1302.

⁵⁵C. O. WEBB *et al.* (2006). « Integrating Phylogenies into Community Ecology ». In : *Ecology* 87.sp7, S1–S2.

⁵⁶C. R. RAO (1982). « Diversity and dissimilarity coefficients : a unified approach ». In : *Theoretical Population Biology* 21, p. 24–43.

⁵⁷R. VANE-WRIGHT *et al.* (1991). « What to protect ?—Systematics and the agony of choice ». In : *Biological Conservation* 55.3, p. 235–254.

⁵⁸R. M. WARWICK *et al.* (1995). « New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress ». In : *Marine Ecology Progress Series* 129, p. 301–305.

⁵⁹J. 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.

⁶⁰C. RICOTTA (2007). « A semantic taxonomy for diversity measures ». In : *Acta Biotheoretica* 55.1, p. 23–33; 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; S. PAVOINE et M. B. BONSALL (2011). « Measuring biodiversity to explain community assembly : a unified approach ». In : *Biological Reviews* 86.4, p. 792–812.

⁶¹S. 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.

⁶²CHAO et al. (2010). « Phylogenetic diversity measures based on Hill numbers », cf. note 40, p. 40.

ramenées aux mesures classiques en fixant une distance égale entre toutes les espèces.

Pavoine *et al.*⁶¹ 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.*⁶² en numérotant les périodes à partir du présent et en notant T_k leur durée. Figure 5.5, 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 (5.4)) est calculée à chaque période. Figure 5.5, 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.

Définition de l'entropie phylogénétique

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 ${}^q\bar{H}(T)$:

$${}^q\bar{H}(T) = \sum_{k=1}^K \frac{T_k}{T} {}^q_kH \quad (5.16)$$

⁶³K. SHIMATANI (2001). « Multivariate point processes and spatial variation of species diversity ». In : *Forest Ecology and Management* 142.1-3, p. 215–229.

⁶⁴C. RICOTTA (2005b). « On parametric diversity indices in ecology : A historical note ». In : *Community Ecology* 6.2, p. 241–244.

⁶⁵D. P. FAITH (1992). « Conservation evaluation and phylogenetic diversity ». In : *Biological Conservation* 61.1, p. 1–10.

⁶⁶O. L. PETCHEY et K. J. GASTON (2002). « Functional diversity (FD), species richness and community composition ». In : *Ecology Letters* 5, p. 402–411.

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,⁶³ puis Ricotta⁶⁴ 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 (5.16) 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 PD⁶⁵ ou FD⁶⁶ (les mesures de diversité phylogénétique ou diversité fonctionnelle égales à la somme de la longueur des branches de l'arbre), ${}^1\bar{H}(T)$ 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 5.5, 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.

Entropie et diversité

L'entropie ${}^q\bar{H}(T)$ peut être transformée en diversité⁶⁷ de la même façon que ${}^qD = e^{{}^q\bar{H}}$:

$${}^q\bar{D}(T) = e_q^{{}^q\bar{H}(T)} \quad (5.17)$$

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.⁶⁸

Chao *et al.*⁶⁹ obtiennent ce résultat sans recourir explicitement à l'entropie, mais en faisant le même calcul :

$${}^q\bar{D}(T) = \left(\sum_{i \in B_T} \frac{T_i}{T} p_i^q \right)^{\frac{1}{1-q}} \quad (5.18)$$

B_T est l'ensemble des branches de l'arbre. Chaque branche, indexée par i , se situe dans une période : toutes les branches de la période k sont de même longueur T_k . Les probabilités associées aux branches, p_i , sont la somme des probabilités des espèces situées sur les feuilles liées à la branche.

Les entropies ont un comportement linéaire : elles s'additionnent tout au long de l'arbre pour donner ${}^q\bar{H}(T)$. Les diversités qD calculées à chaque période ne peuvent pas être sommées sur le modèle de l'équation (5.16) : ${}^q\bar{D}(T)$ 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.

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.

Arbres non ultramétriques

Chao *et al.*⁷⁰ définissent la diversité phylogénétique selon l'équation (5.18) 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.

⁶⁷MARCON *et al.* (2014a). « Generalization of the partitioning of Shannon diversity », cf. note 23, p. 15.

⁶⁸C. RICOTTA *et* L. SZEIDL (2009). « Diversity partitioning of Rao's quadratic entropy ». In : *Theoretical Population Biology* 76.4, p. 299–302.

⁶⁹CHAO *et al.* (2010). « Phylogenetic diversity measures based on Hill numbers », cf. note 40, p. 40.

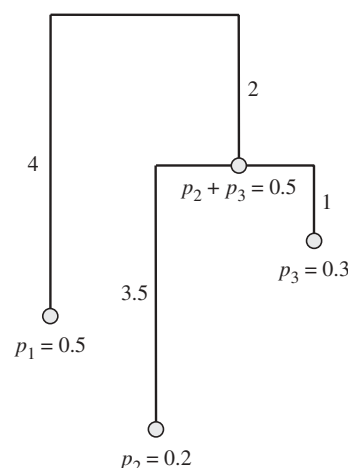


FIGURE 5.6 – Arbre phylogénétique hypothétique non ultramétrique.^a

^aIbid., figure 1b.

⁷⁰CHAO *et al.* (2010). « Phylogenetic diversity measures based on Hill numbers », cf. note 40, p. 40.

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 5.6 peut être découpé en quatre périodes selon la même technique que précédemment 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 Bonsall⁷¹ traitent en détail les résultats aberrants que cause un arbre non ultramétrique dans le cas particulier de l'entropie de Rao. Leinster et Cobbold⁷² 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.

Diversité fonctionnelle

La diversité fonctionnelle peut être calculée par ${}^q\bar{D}(T)$ à condition de définir un arbre ultramétrique pour décrire la divergence fonctionnelle entre les espèces, ce qui est possible mais pas souhaitable.

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.

La première étape consiste donc à choisir un ensemble de traits pertinents et à les mesurer de façon standardisée.⁷³ Toute la stratégie relative à la photosynthèse peut être par exemple assez bien résumée par la masse surfacique des feuilles,⁷⁴ mais, en forêt tropicale, ce trait est décorrélé de la densité du bois.⁷⁵ Les valeurs manquantes peuvent être complétées en utilisant toute l'information disponible par MICE (*multiple imputation by chained equations*),⁷⁶ disponible sous R dans le package *mice*.⁷⁷

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.⁷⁸ La formule de Gower, étendue par Podani⁷⁹ puis Pavoine *et al.*⁸⁰ 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

⁷¹PAVOINE et BONSTALL (2009). « Biological diversity : Distinct distributions can lead to the maximization of Rao's quadratic entropy », cf. note 61, p. 44.

⁷²LEINSTER et COBBOLD (2012). « Measuring diversity : the importance of species similarity », cf. note 25, p. 15.

⁷³J. 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.

⁷⁴I. J. WRIGHT *et al.* (2004). « The worldwide leaf economics spectrum ». In : *Nature* 428, p. 821–827.

⁷⁵C. BARALOTO *et al.* (2010a). « Functional trait variation and sampling strategies in species rich plant communities ». In : *Functional Ecology* 24, p. 208–216.

⁷⁶S. van BUUREN *et al.* (2006). « Fully conditional specification in multivariate imputation ». In : *Journal of Statistical Computation and Simulation* 76.12, p. 1049–1064.

⁷⁷S. van BUUREN et K. GROOTHUIS-OUUDSHOORN (2011). « mice : Multivariate Imputation by Chained Equations in R ». In : *Journal of Statistical Software* 45.3, p. 1–67.

⁷⁸J. C. GOWER (1971). « A General Coefficient of Similarity and Some of Its Properties ». In : *Biometrics* 27.4, p. 857–871.

⁷⁹J. PODANI (1999). « Extending Gower's General Coefficient of Similarity to Ordinal Characters ». In : *Taxon* 48.2, p. 331–340.

⁸⁰S. PAVOINE *et al.* (2011). « Linking patterns in phylogeny, traits, abiotic variables and space : a novel approach to linking environmental filtering and plant community assembly ». English. In : *Journal of Ecology* 99.1, p. 165–175.

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.

Un arbre peut ensuite être construit par classification automatique hiérarchique. Podani et Schmera⁸¹ suggèrent d'utiliser ensuite une classification hiérarchique par UPGMA⁸² qu'ils montrent être la plus robuste (pour le calcul de FD, c'est-à-dire ${}^0\bar{H}(T)$) à l'ajout ou au retrait d'un trait ou d'une espèce.

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 espèces⁸³ : une mesure de diversité qui utilise directement la matrice est préférable, c'est un intérêt de la diversité de Leinster et Cobbold.

5.2.2 Diversité de Leinster et Cobbold

Définitions

Leinster et Cobbold⁸⁴ 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, \mathbf{Z} , décrit par ses valeurs $z_{s,t}$ la similarité entre l'espèce s et l'espèce t comprises entre 0 et 1 ($z_{s,s} = 1$).

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, \mathbf{Z} est la matrice identité \mathbf{I}), au maximum 1 si toutes les espèces sont totalement similaires (\mathbf{Z} ne contient que des 1). Une espèce rare totalement différente des autres est peu banale.

La moyenne généralisée d'ordre r ,⁸⁵ *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 :

$$\bar{x} = \left(\sum_s p_s x_s^r \right)^{\frac{1}{r}} \quad (5.19)$$

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

⁸¹J. PODANI et D. SCHMERA (2006). « On dendrogram-based measures of functional diversity ». In : *Oikos* 115.1, p. 179–185.

⁸²R. 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.

⁸³S. PAVOINE et al. (2005). « 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.

⁸⁴LEINSTER et COBBOLD (2012). « Measuring diversity : the importance of species similarity », cf. note 25, p. 15.

⁸⁵G. H. HARDY et al. (1952). *Inequalities*. Cambridge University Press.

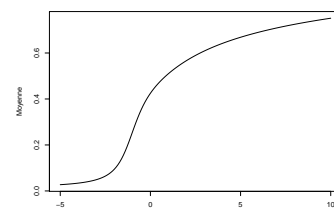


FIGURE 5.7 – Moyenne généralisée d'ordre r (en abscisse) d'une distribution uniforme de 100 valeurs tirées entre 0 et 1. La moyenne arithmétique 0,5 est obtenue pour $r = 1$. La moyenne généralisée tend vers la plus petite valeur (près de 0) quand $r \rightarrow -\infty$, et vers la plus grande valeur (près de 1) quand $r \rightarrow +\infty$.

être négatives), et un fort poids aux grandes valeurs quand r est grand (Figure 5.7).

La banalité peut s'écrire sous forme matricielle en notant \mathbf{p} le vecteur des probabilités p_s , on note alors $(\mathbf{Zp})_s = \sum_t p_t z_{s,t}$.

La banalité moyenne des espèces de la communauté est calculée en prenant $r = q - 1$:

$$(\overline{\mathbf{Zp}}) = \left(\sum_s p_s (\mathbf{Zp})_s^{q-1} \right)^{\frac{1}{q-1}} \quad (5.20)$$

La diversité de la communauté est simplement l'inverse de la banalité moyenne des espèces :

$${}^q D^{\mathbf{Z}} = \left(\sum_s p_s (\mathbf{Zp})_s^{q-1} \right)^{\frac{1}{1-q}} \quad (5.21)$$

${}^q D^{\mathbf{Z}}$ converge vers ${}^1 D^{\mathbf{Z}}$ quand q tend vers 1 :

$${}^1 D^{\mathbf{Z}} = \frac{1}{\prod_s (\mathbf{Zp})_s^{p_s}} \quad (5.22)$$

Diversité neutre

Leinster et Cobbold montrent que la diversité HCDT est un cas particulier de ${}^q D^{\mathbf{Z}}$, pour la matrice identité :

$${}^q D = {}^q D^{\mathbf{I}} \quad (5.23)$$

Diversité non neutre

Une mesure de diversité non neutre est obtenue en utilisant une matrice \mathbf{Z} qui contient l'information sur la similarité entre les espèces. La dissimilarité entre espèces peut être obtenue à partir des arbres.

Dans le cas particulier $q = 2$, ${}^2 D^{\mathbf{Z}}$ est égal à ${}^2 \bar{D}(T)$:

Mais ce n'est pas le cas en général.

⁸⁶C. 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.

Entropie de Ricotta et Szeidl

Ricotta et Szeidl⁸⁶ 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 :

$${}^1 H = - \sum_s p_s \ln \left(1 - \sum_{t \neq s} p_t \right) \quad (5.24)$$

$1 - \sum_{t \neq s} p_t$ est la banalité de l'espèce s si on mesure la diversité neutre. De façon plus générale, $(\mathbf{Zp})_s$ peut exprimer cette banalité. Enfin, l'entropie HCDT peut généraliser l'entropie de Shannon pour définir une mesure Q_α que nous noterons ${}^qH^Z$ pour la cohérence avec les autres mesures ($q = \alpha$) :

$${}^qH^Z = \frac{1 - \sum_s p_s (\mathbf{Zp})_s^{q-1}}{q-1} \quad (5.25)$$

$${}^1H^Z = - \sum_s p_s \ln (\mathbf{Zp})_s \quad (5.26)$$

De façon plus rigoureuse, on peut remarquer que $(\mathbf{Zp})_s$ 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 $(\mathbf{Zp})_s$: la fonction reste décroissante et vaut 0 pour $p_s = 1$ puisque $(\mathbf{Zp})_s = 1$ dans ce cas. ${}^qH^Z$ est donc bien une entropie. La fonction d'information $\ln_q(1/p_s)$ de l'entropie HCDT (5.12) est simplement remplacée par $\ln_q(1/(\mathbf{Zp})_s)$.

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 ${}^qD^Z$ est ${}^qH^Z$:

$$\ln_q {}^qD^Z = {}^qH^Z \quad (5.27)$$

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.⁸⁷

Le problème majeur de la diversité ${}^qD^Z$ est qu'elle fournit souvent des valeurs presque indépendantes de q quand on l'applique à une matrice de similarité fonctionnelle.⁸⁸ 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 :

⁸⁷M. NEI (1972). « Genetic Distance between Populations ». In : *The American Naturalist* 106.949, p. 283–292.

⁸⁸C.-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.

$(\mathbf{Zp})_s^{q-1} \approx 1 + (q-1)[(\mathbf{Zp})_s - 1]$. Cette approximation linéaire implique que $\sum_s p_s (\mathbf{Zp})_s^{q-1} \approx [\sum_s p_s (\mathbf{Zp})_s]^{q-1}$. La puissance $q-1$ disparaît donc dans le calcul de la diversité et l'équation 5.21 devient ${}^q D^{\mathbf{Z}} \approx \bar{z}$ où $\bar{z} = \sum_s p_s (\mathbf{Zp})_s$. Dans les faits, cette approximation vaut pour des valeurs de banalité assez éloignées de 1.

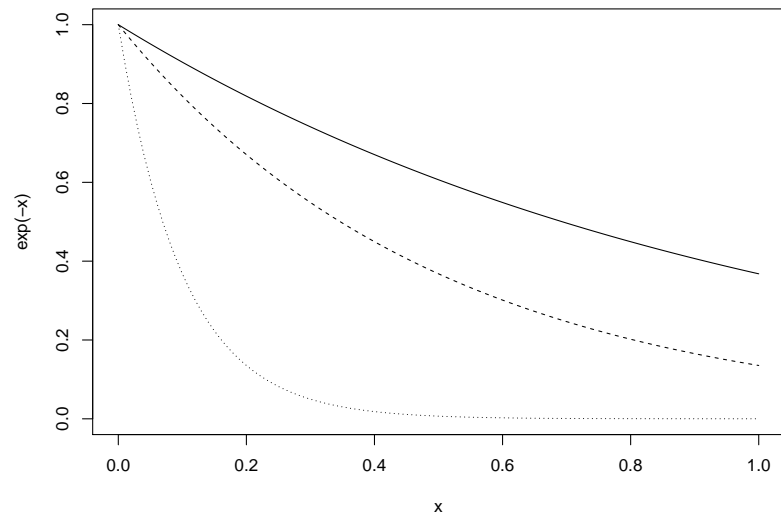
Si une majorité des espèces (au sens de la somme de leurs probabilités respectives) a une banalité supérieure à 0,5, l'approximation linéaire est assez bonne et la moyenne généralisée de la banalité est proche de sa moyenne arithmétique. Ce n'est pas un problème pour la comparaison de profils de diversité 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.

⁸⁹T. LEINSTER (2013). « The Magnitude of Metric Spaces ». In : *Documenta Mathematica* 18, p. 857–905.

La transformation exponentielle négative $z_{s,t} = e^{-u \frac{d_{s,t}}{\max(d_{s,t})}}$ est justifiée par Leinster.⁸⁹ u est une constante positive. Plus u est grand, plus la transformation est convexe : les similarités sont tassées vers 0 (Figure 5.8). 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 .

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.

FIGURE 5.8 – 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.



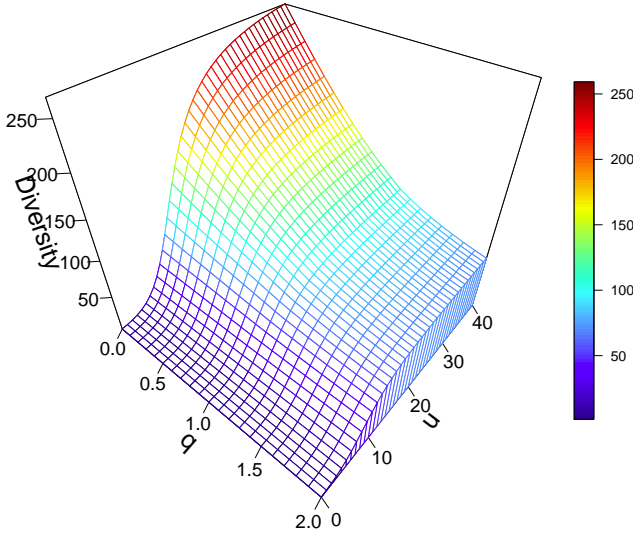


FIGURE 5.9 – Profil bivié de la diversité des deux hectares de forêt de Paracou (parcelles 6 et 18). Le paramètre q contrôle l'importance des espèces originales, la paramètre u la similarité entre les espèces. Les valeurs élevées de u font converger la diversité fonctionnelle vers la diversité neutre.

Leinster relie la valeur de ${}^qD^Z$ 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 Δ , 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 $u \rightarrow +\infty$, $Z \rightarrow I$: 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 (Figure 5.9).

Le paramètre u peut être interprété comme un proxy de la variabilité intraspécifique, ou de façon équivalente, du recouvrement de la niche des espèces voisines.⁹⁰ La variabilité intraspécifique des traits entraîne une superposition des niches des espèces voisines⁹¹ représentée en figure 5.10. Mathématiquement, en supposant que les espèces s et t ont une distribution gaussienne dans l'espace des traits, avec la même variance σ^2 , le recouvrement de leurs niches est :

$$O_{s,t} = 2 \int_{d_{s,t}/2}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{d_{s,t}}{\sigma}\right)^2} \quad (5.28)$$

Le recouvrement ne change pas tant que $d_{s,t}/\sigma$ est constant. Comme les distances ont été normalisées, la variable d'intérêt est l'écart-type σ qui représente la variabilité intraspécifique des traits. Considérons les deux espèces les plus différentes, pour lesquelles $d_{s,t} = 1$. Si $\sigma = 1/6$, leur similarité est très petite (environ 1%). Des valeurs supérieures de σ impliquent qu'aucune paire d'espèces ne peut être considérée comme totalement dissimilaire : $z_{s,t} > 0$,

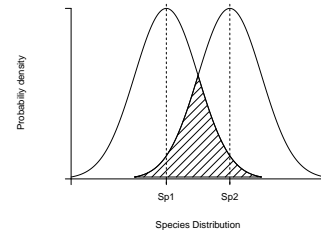


FIGURE 5.10 – Représentation de la superposition de la niche de deux espèces. Les espèces sont placées dans l'espace multidimensionnel des traits. La variabilité intraspécifique entraîne une distribution gaussienne des individus autour de la position moyenne. La variance est supposée identique pour toutes les espèces et toutes les dimensions (tous les traits). Deux espèces sont représentées sur la figure, avec leur densité de probabilité, projetées sur l'axe qui les relie. L'écart-type de la distribution sur la figure est la moitié de la distance entre les deux espèces.

⁹⁰MARCON et al. (2014b). « The Decomposition of Similarity-Based Diversity and its Bias Correction », cf. note 26, p. 15.

⁹¹J. LEPŠ et al. (2006). « Quantifying and interpreting functional diversity of natural communities : practical considerations matter ». In : *Preslia* 78, p. 481–501.

⁹²S. 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.

quelles que soient s et t . À l’opposé, une variabilité plus faible permet $z_{s,t} \approx 0$ pour des paires d’espèces plus proches.

Une assez bonne approximation de $O_{s,t}$ est $e^{-\frac{d}{\sigma\sqrt{2}}}$. En choisissant $u = 1/\sigma\sqrt{2}$, la similarité $z_{s,t} = e^{-ud_{s,t}}$ représente le recouvrement des espèces. La valeur de u correspondant à $\sigma = 1/6$ est de l’ordre de 4. De plus grandes échelles correspondent à moins de variabilité intraspécifique.

Le lien entre les propriétés mathématiques du paramètre d’échelle u et sa signification biologique est donc établi, même si les hypothèses du modèle, à savoir la variabilité identique de tous les traits pour toutes les espèces, n’est pas réaliste. Elles sont plus satisfaisantes que l’hypothèse habituelle d’absence complète de variabilité. La diversité des valeurs propres,⁹² présentée plus bas, permet de mieux la prendre en compte mais a d’autres limites plus critiques.

Diversité phylogénétique

La diversité phylogénétique ${}^q\bar{D}(T)$ est un cas particulier de ${}^qD^Z$ pour une matrice 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 5.5, page 43, 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 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 correspondant 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.

Synthèse

${}^qD(T)$ est l’exponentielle d’ordre q de la moyenne pondérée de l’entropie HCdT calculée à chaque période de l’arbre phylogénétique.

${}^qD^Z$ 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.

${}^qD^Z$ et qD sont des nombres effectifs d’espèces, respectent toutes les propriétés demandées à une mesure de diversité dont le principe de réplique. Leurs valeurs sont identiques pour $q = 2$

(diversité de Rao). ${}^qD^Z$ est moins sensible au paramètre q que ${}^qD(T)$: la Figure 5.11 compare les deux mesures.

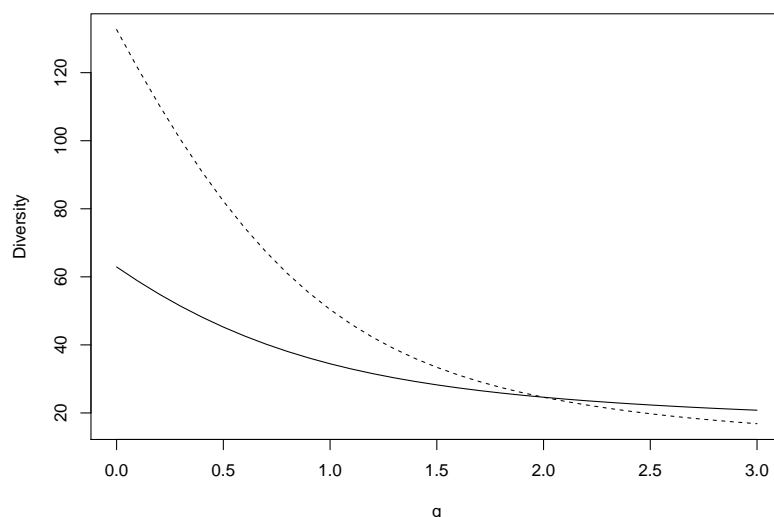


FIGURE 5.11 – Diversité phylogénétique (l'arbre est une taxonomie) d'ordre q des deux hectares de forêt de Paracou (parcelles 6 et 18) : ${}^qD^Z$ (trait plein) et ${}^qD(T)$ (pointillé).

${}^qD^Z$ 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.⁹³ 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 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.

Diversité individuelle

La diversité ${}^qD^Z$ 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 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 probabilités vérifiant $p_s = p_{s'} + p_{s''}$. L'opération peut être répétée jusqu'à la désagrégation complète de la matrice où chaque ligne

⁹³PAVOINE et al. (2005). « Is the originality of a species measurable ? », cf. note 83, p. 47 ; PODANI et SCHMERA (2006). « On dendrogram-based measures of functional diversity », cf. note 81, p. 47 ; PODANI et SCHMERA (2007). « How should a dendrogram-based measure of functional diversity function ? A rejoinder to Petchey and Gaston », cf. note 83, p. 47.

correspondrait à un individu et les probabilités seraient égales à $1/N$. Il n'y a 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 entropies 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.⁹⁴ 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.

Formellement, si l'espèce s est composée de deux groupes s' et s'' , la banalité du groupe s' est :

$$(\mathbf{Zp})_{s'} = \sum_{t \neq s', s''} p_t z_{s', t} + p_{s'} + p_{s''} z_{s', s''}$$

Celle du groupe s'' est :

$$(\mathbf{Zp})_{s''} = \sum_{t \neq s', s''} p_t z_{s'', t} + p_{s'} z_{s', s''} + p_{s''}$$

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

$$\sum_{s'} p_{s'} (\mathbf{Zp})_{s'}^{q-1} + \sum_{s''} p_{s''} (\mathbf{Zp})_{s''}^{q-1}$$

est remplacée après regroupement par

$$\sum_s p_s (\mathbf{Zp})_s^{q-1}$$

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

$$\sum_{s'} p_{s'} (\mathbf{Zp})_{s'} + \sum_{s''} p_{s''} (\mathbf{Zp})_{s''} = \sum_s p_s (\mathbf{Zp})_s$$

⁹⁴S. 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 88, p. 49.

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 .

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.

5.2.3 Diversité des valeurs propres

L'approche de Pavoine et Izsák⁹⁵ 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 Isaza⁹⁶ qui ont relié la diversité de Simpson au volume de la sphère sur laquelle se trouve le point définissant la communauté (Figure 5.12).

⁹⁵PAVOINE et IZSÁK (2014). « New biodiversity measure that includes consistent interspecific and intraspecific components », cf. note 92, p. 52.

⁹⁶D. CAMPOS et J. F. ISAZA (2009). « A geometrical index for measuring species diversity ». In : *Ecological Indicators* 9.4, p. 651–658.

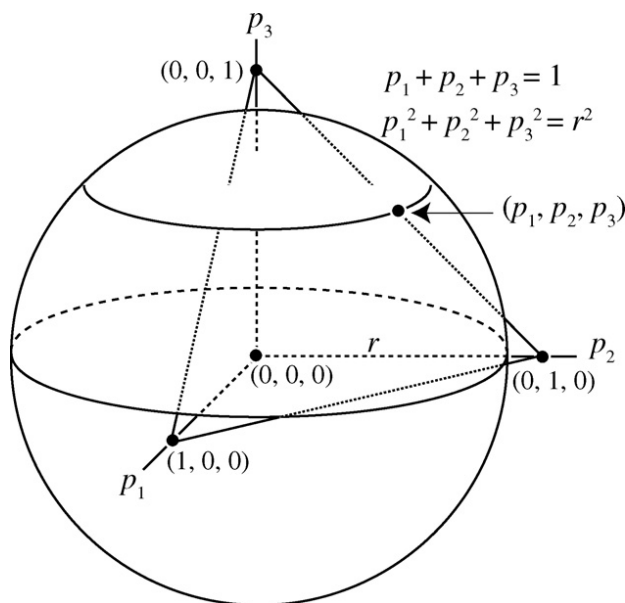


FIGURE 5.12 – Représentation d'une communauté dans un espace multidimensionnel dont chaque axe correspond à une espèce. La communauté est composée de trois espèces. Le point représentant la communauté se trouve sur la sphère de rayon $r = \sqrt{\sum_s p_s^2}$. Le plan (qui est en réalité un hyperplan de dimension $S - 1$) d'équation $\sum_s p_s = 1$ est représenté par un triangle. (in Campos et Isaza, 2009)

La matrice de similarité entre les espèces \mathbf{Z} permet d'affiner la représentation. Soit la matrice $\mathbf{C} = \sqrt{\mathbf{Z}}$. 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 en compte la similarité. Dans le cas extrême de la diversité neutre,

$\mathbf{C} = \mathbf{I}$, chaque espèce est située uniquement sur son axe. Dans l'autre cas extrême de totale similarité, où \mathbf{C} ne contient que des 1, toutes les espèces sont colinéaires.

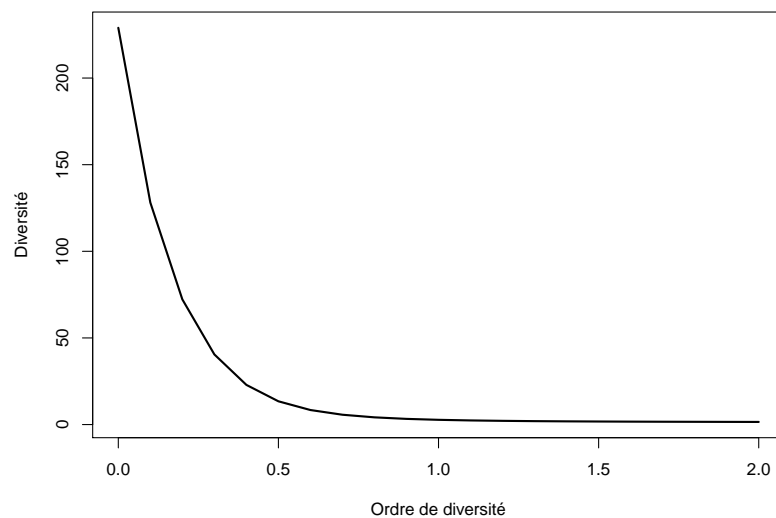
Il est possible de calculer les S valeurs propres notées λ_s de la matrice des coordonnées des espèces et de les normaliser : $\mu_s = \lambda_s / \sum_s \lambda_s$. La communauté peut maintenant être représentée par ses valeurs propres μ_s correspondant à la proportion d'une « espèce composite » pure 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 ${}^qD^{\mathbf{Z}}(\Lambda)$.

La définition de cette diversité autorise toute matrice \mathbf{C} symétrique, dont les valeurs de similarité sont positives ou nulles, strictement positives sur la diagonale. Si \mathbf{C} est la racine carrée d'une matrice de similarité au sens strict \mathbf{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 ${}^2D^{\mathbf{Z}}(\Lambda)$ est égale à la diversité de Rao appliquée à une matrice de distances $\mathbf{D} = 1 - \mathbf{Z}$. La définition de \mathbf{C} comme racine carrée de \mathbf{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 $\mathbf{Z} = \mathbf{I}$, l'ACP ne modifie pas le nuage de points original et ${}^qD^{\mathbf{Z}}(\Lambda) = {}^qD$.

Le profil de diversité se trouve en Figure 5.13.

FIGURE 5.13 – Profil de diversité des valeurs propres des deux hectares de forêt de Paracou (parcelles 6 et 18).



La diversité des valeurs propres permet de ramener le problème de la diversité d'espèces partiellement similaires au calcul

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

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.

5.3 Diversité β et décomposition

La notion de diversité β a été introduite par Whittaker⁹⁷ 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

⁹⁷R. H. WHITTAKER (1960). « Vegetation of the Siskiyou Mountains, Oregon and California ». In : *Ecological Monographs* 30.3, p. 279–338, page 320.

⁹⁸M. 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.

⁹⁹A. M. ELLISON (2010). « Partitioning diversity ». In : *Ecology* 91.7, p. 1962–1963.

¹⁰⁰A. BASELGA (2010). « Multiplicative partition of true diversity yields independent alpha and beta components; additive partition does not ». In : *Ecology* 91.7, p. 1974–1981; L. JOST (2010). « 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.

¹⁰¹A. CHAO et al. (2012). « Proposing a resolution to debates on diversity partitioning ». In : *Ecology* 93.9, p. 2037–2051.

¹⁰²H. 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.

¹⁰³TUOMISTO (2010a). Cf. note 102.

¹⁰⁴JOST (2006). « Entropy and diversity », cf. note 18, p. 15; JOST (2007). « Partitioning diversity into independent alpha and beta components », cf. note 19, p. 15.

¹⁰⁵H. 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.

en une définition sans ambiguïté est encore une question de recherche et de débats. Anderson *et al.*⁹⁸ fournissent une revue des analyses utiles de la diversité β en forme de guide à destination des écologues.

Dans la littérature, la diversité β est généralement une mesure *dérivée*,⁹⁹ 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 .¹⁰⁰ Un forum a été consacré à la question dans la revue *Ecology*, conclu par Chao *et al.*,¹⁰¹ et une revue a été faite par Tuomisto.¹⁰²

Pour simplifier l'exposé, les individus seront échantillonnés dans des communautés, appartenant à une méta-communauté. Le Tableau 5.1 résume les notations. Toute ambiguïté sera évitée en appelant « entropie » les indices de diversité qui sont des mesures d'entropie qH (indices de Shannon et de Simpson par exemple) et « diversité » leur nombre équivalent qD .

5.3.1 Définitions de la diversité β , mesure dérivée

Tuomisto¹⁰³ passe en revue l'ensemble des définitions de la diversité β dérivée des diversités γ et α . Toutes ont en commun :

- Une définition de la mesure de diversité, appliquée à la diversité γ , qui est généralement une des mesures vues plus haut ;
- 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 :
 - $\beta = \gamma/\alpha$
 - $\beta = \gamma - \alpha$

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 »¹⁰⁴ β qui est un nombre de communautés équivalentes (*compositionnal units*) similaire au nombre d'espèces équivalentes des diversités α et γ . Tuomisto¹⁰⁵ 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.

5.3.2 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 .

Whittaker¹⁰⁶ est l'auteur de ce concept. Il a posé le principe que la diversité γ devait être le produit des diversités α et β .

Lande¹⁰⁷ 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 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 exprimées 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.*¹⁰⁸

Un débat assez stérile a découlé de l'opposition entre les deux approches, principalement dû à la transformation logarithmique,¹⁰⁹ à des définitions imprécises et des démonstrations empiriques remises en question.¹¹⁰ Il reste que la décomposition multiplicative permet seule la définition de la diversité β en tant que diversité au sens strict.¹¹¹

Jurasinski *et al.*¹¹² 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 .

J'ai montré¹¹³ 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 . 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, j'ai généralisé¹¹⁴ ce résultat à toutes les mesures de diversités dérivées de l'entropie généralisée de Tsallis.

5.3.3 Décomposition multiplicative de la diversité

Jost¹¹⁵ et Chao *et al.*¹¹⁶ ont montré que la décomposition des

¹⁰⁶WHITTAKER (1960). « Vegetation of the Siskiyou Mountains, Oregon and California », cf. note 97, p. 57 ; R. H. WHITTAKER (1972). « Evolution and Measurement of Species Diversity ». In : *Taxon* 21.2/3, p. 213–251.

¹⁰⁷R. LANDE (1996). « Statistics and partitioning of species diversity, and similarity among multiple communities ». In : *Oikos* 76, p. 5–13.

¹⁰⁸J. A. VEECH *et al.* (2002). « The additive partitioning of species diversity : recent revival of an old idea ». In : *Oikos* 99.1, p. 3–9.

¹⁰⁹C. RICOTTA (2005a). « On hierarchical diversity decomposition ». In : *Journal of Vegetation Science* 16.2, p. 223–226.

¹¹⁰BASELGA (2010). Cf. note 100 ; VEECH *et CRIST* (2010). Cf. note 100.

¹¹¹CHAO *et al.* (2012). Cf. note 101.

¹¹²G. JURASINSKI *et al.* (2009). « Inventory, differentiation, and proportional diversity : a consistent terminology for quantifying species diversity ». English. In : *Oecologia* 159.1, p. 15–26.

¹¹³MARCON *et al.* (2012b). « The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity », cf. note 20, p. 15.

¹¹⁴MARCON *et al.* (2014a). « Generalization of the partitioning of Shannon diversity », cf. note 23, p. 15.

¹¹⁵JOST (2007). « Partitioning diversity into independent alpha and beta components », cf. note 19, p. 15.

¹¹⁶CHAO *et al.* (2012). Cf. note 101.

nombre de Hill en éléments indépendants est multiplicative :

$${}^qD_\gamma = {}^qD_\alpha {}^qD_\beta \quad (5.29)$$

${}^qD_\alpha$ 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 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, 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 ${}^qD_\beta$ n'est pas contrainte par celle de ${}^qD_\alpha$. Cette propriété est souvent considérée comme importante,¹¹⁷ et sera discutée ici.

¹¹⁷M. 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 39, p. 40.

5.3.4 Définitions de la diversité α

La diversité α est calculée pour chaque communauté, et notée ${}^qD_\alpha$ (l'entropie correspondante est notée ${}^qH_\alpha$). 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.¹¹⁸ La définition classique est selon Routledge¹¹⁹ :

$${}^qH_\alpha = \sum_i w_i {}^qH_{\alpha,i} \quad (5.30)$$

$${}^qD_\alpha = \left(\sum_s \sum_i w_i p_{s,i}^q \right)^{1/(1-q)} \quad (5.31)$$

La pondération proposée par Jost¹²⁰ est :

$${}^qH_\alpha = \sum_i \frac{w_i^q}{\sum_i w_i^q} {}^qH_{\alpha,i} \quad (5.32)$$

$${}^qD_\alpha = \left(\sum_s \sum_i \frac{w_i^q}{\sum_i w_i^q} p_{s,i}^q \right)^{1/(1-q)} \quad (5.33)$$

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, Jost¹²¹ a montré que l'indépendance n'est possible que si la diversité α est définie selon sa pondération.

¹¹⁸CHAO et al. (2012). « Proposing a resolution to debates on diversity partitioning », cf. note 101, p. 58.

¹¹⁹R. D. ROUTLEDGE (1979). « Diversity indices : Which ones are admissible? » In : *Journal of Theoretical Biology* 76.4, p. 503-515.

¹²⁰JOST (2007). « Partitioning diversity into independent alpha and beta components », cf. note 19, p. 15.

¹²¹Ibid.

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 pour des communautés de même poids ou pour $q = 0$ ou $q = 1$.

Chiu *et al.*¹²² 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 :

$${}^qD_\alpha = \frac{1}{I} \left(\sum_s \sum_i (w_i p_{s,i})^q \right)^{1/(1-q)} \quad (5.34)$$

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.

Marcon *et al.*¹²³ préfèrent la définition de Routledge qui est plus intuitive, garantit que la diversité γ est supérieure à la diversité α . La dépendance entre ${}^qD_\beta$ et ${}^qD_\alpha$ est le prix à payer.

5.3.5 Décomposition de l'entropie HCdT

Décomposition de l'indice de Shannon

La décomposition explicite est due à Marcon *et al.*¹²⁴ mais avait été établie plus ou moins clairement plusieurs fois dans la littérature, notamment par Rao et Nayak¹²⁵ : 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.¹²⁶

Lewontin,¹²⁷ dans son célèbre article sur la diversité génétique humaine, définit la diversité de Shannon inter-groupe comme la 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,¹²⁸ 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.¹²⁹

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.

Test de significativité L'objectif est de tester si deux communautés ne sont pas simplement deux échantillons d'une même

¹²²C.-H. CHIU *et al.* (2014). « Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers ». In : *Ecological Monographs* 84.1, p. 21–44.

¹²³MARCON *et al.* (2014a). « Generalization of the partitioning of Shannon diversity », cf. note 23, p. 15.

¹²⁴MARCON *et al.* (2012b). « The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity », cf. note 20, p. 15.

¹²⁵C. 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.

¹²⁶J. LIN (1991). « Divergence Measures Based on the Shannon Entropy ». In : *IEEE Transactions on Information Theory* 37.1, p. 145–151.

¹²⁷R. LEWONTIN (1972). « The apportionment of human diversity ». In : *Evolutionary biology* 6, p. 381–398.

¹²⁸C. RICOTTA *et* G. AVENA (2003). « An information-theoretical measure of beta-diversity ». In : *Plant Biosystems* 137.1, p. 57–61.

¹²⁹A. LUDOVISI *et* M. I. TATICCHI (2006). « Investigating beta diversity by Kullback-Leibler information measures ». In : *Ecological Modelling* 192.1-2, p. 299–313.

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 .

Le test est réalisé de la façon suivante :

- Les effectifs de chaque communauté i sont tirés dans une loi multinomiale $\mathcal{M}(n_{+i}, n_{s,i}/n_{+i})$ et H_β 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 H_β 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 question¹³⁰ : les communautés réelles ne pouvant pas être exactement identiques, il suffit d'augmenter la taille de l'échantillonnage pour prouver leur différence.

¹³⁰D. JONES et N. MATLOFF (1986). « Statistical Hypothesis Testing in Biology : A Contradiction in Terms ». In : *Journal of Economic Entomology* 79.5, p. 1156–1160.

Intervalle de confiance de H_β L'intervalle de confiance de l'estimateur de H_β peut être calculé de la même manière en simulant les communautés par des tirages dans des lois multinomiales suivant leurs fréquences : $\mathcal{M}(n_{+i}, n_{s,i}/n_{+i})$.

Si l'intervalle de confiance ne contient pas 0, l'égalité des distributions est rejetée.

La figure 5.14 présente la diversité (plutôt que l'entropie) de Shannon des deux hectares (parcelles 6 et 18) de Paracou avec ses intervalles de confiance. La valeur 1 (1 communauté effective, correspondant à une entropie H_β nulle) n'est pas dans l'intervalle de confiance de la diversité β .

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.*¹³¹ puis Chao *et al.*¹³² effectuent la deuxième correction par un recentrage de la distribution de H_β

¹³¹MARCON *et al.* (2012b). « The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity », cf. note 20, p. 15.

¹³²A. 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.

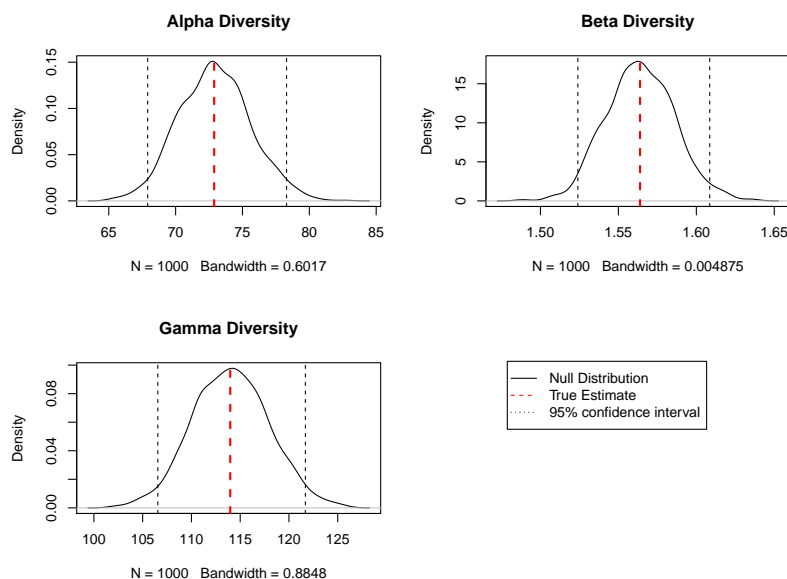


FIGURE 5.14 – Estimation de la diversité des deux hectares de forêt de Paracou (parcelles 6 et 18).

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 sur-estimée.

Décomposition de l'entropie généralisée

La décomposition additive de l'entropie selon Lande¹³³ doit être la suivante :

$${}^qH_\gamma = {}^qH_\alpha + {}^qH_\beta \quad (5.35)$$

Bourguignon¹³⁴ comme Lande¹³⁵ 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 $H_\alpha = \sum_i w_i H_{\alpha_i}$.
- 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, Jost¹³⁶ 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 postulat¹³⁷ permet de décomposer la diversité d'ordre

¹³³LANDE (1996). « Statistics and partitioning of species diversity, and similarity among multiple communities », cf. note 107, p. 59.

¹³⁴F. BOURGUIGNON (1979). « Decomposable Income Inequality Measures ». In : *Econometrica* 47.4, p. 901-920.

¹³⁵LANDE (1996). « Statistics and partitioning of species diversity, and similarity among multiple communities », cf. note 107, p. 59.

¹³⁶JOST (2007). « Partitioning diversity into independent alpha and beta components », cf. note 19, p. 15.

¹³⁷MARCON et al. (2014a). « Generalization of the partitioning of Shannon diversity », cf. note 23, p. 15.

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 (5.29) :

$${}^qH_\gamma = {}^qH_\alpha + \ln_q {}^qD_\beta - (q-1) {}^qH_\alpha \ln_q {}^qD_\beta \quad (5.36)$$

¹³⁸JOST (2007). « Partitioning diversity into independent alpha and beta components », cf. note 19, p. 15.

Ceci est la décomposition de Jost.¹³⁸ 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 .

Les deux derniers termes peuvent être regroupés et réarrangés pour obtenir ${}^qH_\beta$ conformément à la décomposition additive de l'équation (5.35) :

$${}^qH_\beta = \sum_i w_i \sum_s p_{s,i}^q \ln_q \frac{p_{s,i}}{p_s} \quad (5.37)$$

L'entropie β est la somme pondérée par w_i (et pas autrement) des contributions de chaque communauté :

$${}_i^qH_\beta = \sum_s p_{s,i}^q \ln_q \frac{p_{s,i}}{p_s} \quad (5.38)$$

¹³⁹L. BORLAND et al. (1998). « Information gain within nonextensive thermostatics ». 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.

${}_i^qH_\beta$ est une divergence de Kullback-Leibler généralisée.¹³⁹ De même, ${}^qH_\gamma$ peut s'écrire sous la forme ${}^qH_\gamma = -\sum_s p_s^q \ln_q p_s$: la formalisation de la décomposition est une généralisation de la décomposition de l'entropie de Shannon, résumée dans le Tableau 5.2.

Il est intéressant d'écrire ${}_i^qH_\beta$ et ${}^qH_\gamma$ sous la forme d'entropies, pour faire apparaître leur fonction d'information. Pour q différent de 1, l'entropie γ est :

$${}^qH_\gamma = \sum_s p_s \frac{1 - p_s^{q-1}}{q-1} = \sum_s p_s \ln_q \frac{1}{p_s} \quad (5.39)$$

L'entropie β est :

$${}_i^qH_\beta = \sum_s p_{s,i} \frac{p_{s,i}^{q-1} - p_s^{q-1}}{q-1} = \sum_s p_{s,i} \left(\ln_q \frac{1}{p_s} - \ln_q \frac{1}{p_{s,i}} \right) \quad (5.40)$$

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 :

$${}^0D_\beta = \frac{S}{\bar{S}} \quad (5.41)$$

${}^0D_\beta$ peut être supérieur au nombre de communautés¹⁴⁰ 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.*¹⁴¹ 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.

¹⁴⁰CHAO et al. (2012). « Proposing a resolution to debates on diversity partitioning », cf. note 101, p. 58.

¹⁴¹MARCON et al. (2014a). « Generalization of the partitioning of Shannon diversity », cf. note 23, p. 15.

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 $p_s(1 - p_s)$. Cette variance peut être décomposée entre les communautés, où la probabilité est $p_{s,i}$. La décomposition entre variance intra et inter-communautés est :

$$p_s(1 - p_s) = \sum_i w_i \left[p_{s,i}(1 - p_{s,i}) + (p_{s,i} - p_s)^2 \right] \quad (5.42)$$

Cette égalité peut être sommée sur toutes les espèces pour donner :

$$E_\gamma = \sum_i w_i E_{\alpha,i} + \sum_i w_i \sum_s (p_{s,i} - p_s)^2 = E_\alpha + E_\beta \quad (5.43)$$

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.

En génétique, l'égalité s'écrit¹⁴² $H_t = H_s + D_{s,t}$. 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 (5.40) pour le cas général, mais la contribution

¹⁴²M. 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.

TABLE 5.2 – 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)	${}^qD_\beta = e_q^{\frac{{}^qH_\beta}{1-(q-1){}^qH_\alpha}}$	${}^1D_\beta = e^{{}^1H_\beta}$

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.

¹⁴³JOST (2007). « Partitioning diversity into independent alpha and beta components », cf. note 19, p. 15; L. JOST (2008). « GST and its relatives do not measure differentiation ». In : *Molecular Ecology* 17.18, p. 4015–4026.

¹⁴⁴GREGORIUS (2014). « Partitioning of diversity : the "within communities" component », cf. note 22, p. 37.

Cette décomposition est remise en cause par Jost.¹⁴³ 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 69). Gregorius¹⁴⁴ 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 :

$${}^qD_\gamma = {}^qD_\alpha {}^qD_\beta \Leftrightarrow \frac{1}{1-E_\gamma} = \frac{1}{1-E_\alpha} \frac{1}{1-E_\beta} \quad (5.44)$$

D'où une décomposition additive différente, qui correspond à sa décomposition générale (5.36) :

$$E_\gamma = E_\alpha + E_\beta - E_\alpha E_\beta \quad (5.45)$$

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 (5.35) est la décomposition multiplicative (5.29).

Le nombre équivalent de communautés a une forme légèrement différente d'un nombre de Hill (Tableau 5.2).

Une interprétation géométrique de la décomposition est la suivante.¹⁴⁵ Les espèces peuvent être placées dans un espace multidimensionnel construit par une Analyse en Coordonnées Principales¹⁴⁶ 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

¹⁴⁵S. 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.

¹⁴⁶J. C. GOWER (1966). « Some distance properties of latent root and vector methods used in multivariate analysis ». In : *Biometrika* 53.3, p. 325–338.

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 ,¹⁴⁷ dans la tradition de l'utilisation de la dissimilarité entre paires de communautés.¹⁴⁸

Cette partition suppose que les poids des communautés sont proportionnels à leur nombre d'individus.¹⁴⁹ Hardy et Senterre¹⁵⁰ décomposent (par différence entre γ et α) l'indice de Rao de communautés de poids égaux. Hardy et Jost¹⁵¹ montrent que les deux pondérations sont valides mais l'absence de cadre général assurant que la diversité γ de Rao est supérieure à la diversité α ¹⁵² motive Guiasu et Guiasu¹⁵³ 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 $\gamma \geq \alpha$). 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.¹⁵⁴ Ce résultat est valide pour les arbres ultramétriques dans le cadre plus général de la décomposition de l'entropie phylogénétique (voir section 5.3.5 page 61).

La décomposition de la diversité (et non seulement de l'entropie) de Rao a été établie par Ricotta et Szeidl.¹⁵⁵

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 5.2.

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 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.¹⁵⁶

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,¹⁵⁷ 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

¹⁴⁷C. 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.

¹⁴⁸J. C. GOWER et P. LEGENDRE (1986). « Metric and Euclidean properties of dissimilarity coefficients ». In : *Journal of classification* 48, p. 5–48.

¹⁴⁹RAO (1982). « Diversity and dissimilarity coefficients : a unified approach », cf. note 56, p. 43 ; S. VILÉGER 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.

¹⁵⁰O. 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.

¹⁵¹O. J. HARDY et L. JOST (2008). « Interpreting and estimating measures of community phylogenetic structuring ». In : *Journal of Ecology* 96.5, p. 849–852.

¹⁵²F. 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.

¹⁵³R. 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.

¹⁵⁴S. CHAMPELY et D. CHESSEL (2002). « Measuring biological diversity using Euclidean metrics ». In : *Environmental and Ecological Statistics* 9.2, p. 167–177.

¹⁵⁵RICOTTA et SZEIDL (2009). « Diversity partitioning of Rao's quadratic entropy », cf. note 68, p. 45.

¹⁵⁶T. 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. (2012b). « The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity », cf. note 20, p. 15.

¹⁵⁷MARCON et al. (2012b). « The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity », cf. note 20, p. 15 ; J. BECK et al. (2013). « Undersampling and the measurement of beta diversity ». In : *Methods in Ecology and Evolution* 4.4, p. 370–382.

¹⁵⁸MARCON et al. (2014a). « Generalization of the partitioning of Shannon diversity », cf. note 23, p. 15; MARCON et al. (2014b). « The Decomposition of Similarity-Based Diversity and its Bias Correction », cf. note 26, p. 15.

¹⁵⁹CRIST et al. (2003). « Partitioning species diversity across landscapes and regions : A hierarchical analysis of alpha, beta, and gamma diversity », cf. note 156, p. 67.

¹⁶⁰JONES et MATLOFF (1986). « Statistical Hypothesis Testing in Biology : A Contradiction in Terms », cf. note 130, p. 62.

¹⁶¹R. CONDIT (1998). *Tropical Forest Census Plots*. Berlin, Germany, et Georgetown, Texas : Springer-Verlag et R. G. Landes Company, p. 1-224; S. P. HUBBELL (1999). « Light-Gap Disturbances, Recruitment Limitation, and Tree Diversity in a Neotropical Forest ». In : *Science* 283.5401, p. 554-557; S. P. HUBBELL et al. (2005). *Barro Colorado Forest Census Plot Data*.

¹⁶²JOST (2007). « Partitioning diversity into independent alpha and beta components », cf. note 19, p. 15.

générale consiste donc à corriger les entropies α et γ , calculer l'entropie β par différence et transformer les résultats en nombres équivalents.¹⁵⁸

Enfin, il est possible¹⁵⁹ mais controversé¹⁶⁰ 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.

5.3.6 Normalisation

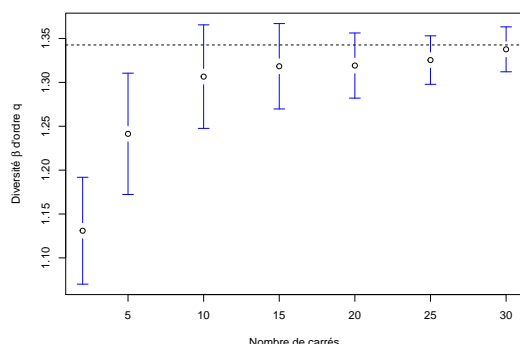
La mesure de diversité ${}^qD_\beta$ est le nombre équivalent de communautés totalement distinctes qui fourniraient ce niveau de diversité.

Nécessité de normaliser la diversité β

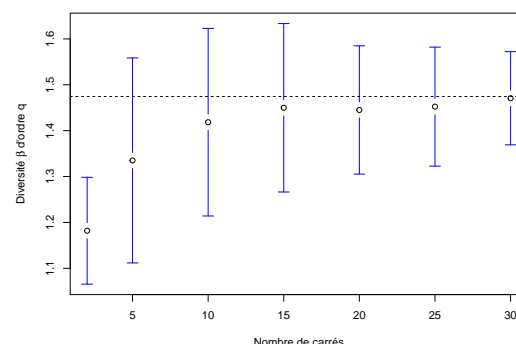
Selon une première approche, cette mesure n'a de sens que comparée à nombre de communautés. La diversité entre un hectare de forêt tropicale et un hectare de forêt tempérée (de même poids, sans espèce commune) est égale à 2 quel que soit q . La diversité entre un nombre suffisant d'échantillons de forêt tropicale relativement homogène (avec de nombreuses espèces communes) peut facilement dépasser 2 : c'est le cas des 50 carrés d'un hectare de la forêt de Barro-Colorado Island (BCI) à Panama¹⁶¹ 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, Jost¹⁶² 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.

Un autre cas est envisageable : les communautés peuvent être des échantillons d'une communauté clairement définie. La diversité β mesure alors la variabilité de l'échantillonnage à l'intérieur de la 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 5.37, page 64) de la divergence de Jensen-Shannon entre la distribution des espèces d'un échantillon et celle de la 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 5.15 présente les valeurs estimées de diversité neutre β entre carrés de BCI tirés aléatoirement, en fonction du nombre de carrés.



(a) Pour $q = 1$, la diversité est sous-estimée parce que la correction du biais d'estimation n'est pas suffisamment efficace.



(b) 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.

FIGURE 5.15 – 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.

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. Pour $q = 2$, l'estimateur est sans biais, la diversité est très bien estimée même avec un échantillonnage réduit. Dans les deux cas, on peut accepter que l'estimation de la diversité β est constante dès 10 carrés.

La diversité β ne doit donc pas être normalisée systématiquement, mais seulement dans le premier cas.

L'indice de chevauchement C_{qN}

Plutôt qu'une simple normalisation, Chao *et al.*¹⁶³ 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 équiréquentes, dont A seraient communes à toutes les communautés et les autres représentées dans une seule, A/S serait cet indice de chevauchement.

Pour $q \neq 1$:

¹⁶³A. CHAO *et al.* (2008). « A Two-Stage Probabilistic Approach to Multiple-Community Similarity Indices ». In : *Biometrics* 64.4, p. 1178–1186.

$$C_{qN} = \left[\left(\frac{1}{qD_\beta} \right)^{q-1} - \left(\frac{1}{N} \right)^{q-1} \right] / \left[1 - \left(\frac{1}{N} \right)^{q-1} \right] \quad (5.46)$$

Et :

$$C_{1N} = \frac{1}{\ln N} \sum_{s=1}^S \sum_{i=1}^N \frac{p_{s,i}}{N} \ln \left(1 + \frac{\sum_{j \neq i} p_{sj}}{p_{s,i}} \right) = 1 - \frac{H_\beta}{\ln N} \quad (5.47)$$

¹⁶⁴CHAO et al. (2012). « Proposing a resolution to debates on diversity partitioning », cf. note 101, p. 58.

C_{qN} a été défini à l'origine pour $q \geq 2$. Chao *et al.*¹⁶⁴ l'étendent sans précaution à q quelconque. Sa valeur peut être négative pour $q < 1$, on se limitera donc à $q \geq 1$.

5.3.7 Décomposition de la diversité phylogénétique

¹⁶⁵MARCON et HÉRAULT (2015a). « Decomposing Phylodiversity », cf. note 24, p. 15.

L'entropie phylogénétique est une combinaison linéaire de l'entropie HCDT donc sa décomposition est identique.¹⁶⁵ En combinant l'équation (5.16) et le Tableau 5.2, on obtient :

$${}^q\bar{H}_\gamma(T) = {}^q\bar{H}_\alpha(T) + {}^q\bar{H}_\beta(T) \quad (5.48)$$

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 $p_s = \sum_i w_i p_{s,i}$:

$${}^q\bar{H}_\alpha(T) = \sum_i w_i {}^q\bar{H}_i(T) \quad (5.49)$$

L'entropie β est :

$${}^q\bar{H}_\beta(T) = \sum_i w_i {}^q\bar{H}_{i\beta}(T) \quad (5.50)$$

$$= \sum_i w_i \sum_k \frac{T_k}{T} {}^q\bar{H}_{ik\beta} \quad (5.51)$$

$$= \sum_i w_i \sum_k \frac{T_k}{T} \sum_u p_{k,u,i}^q \ln_q \frac{p_{k,u,i}}{p_{k,u}} \quad (5.52)$$

où $p_{k,u}$ est la probabilité d'observer une des espèces de la branche u de l'arbre à la période k (Figure 5.5) 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.*¹⁶⁶), c'est-à-dire dans la cas particulier $q = 1$, a été établie par Mouchet et Mouillot.¹⁶⁷

La Figure 5.16 donne une représentation graphique de la décomposition de la diversité. La différence de taille entre les

¹⁶⁶B. 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.

¹⁶⁷M. A. MOUCHET et D. MOUILLOT (2011). « Decomposing phylogenetic entropy into α , β and γ components ». In : *Biology Letters* 7.2, p. 205–209.

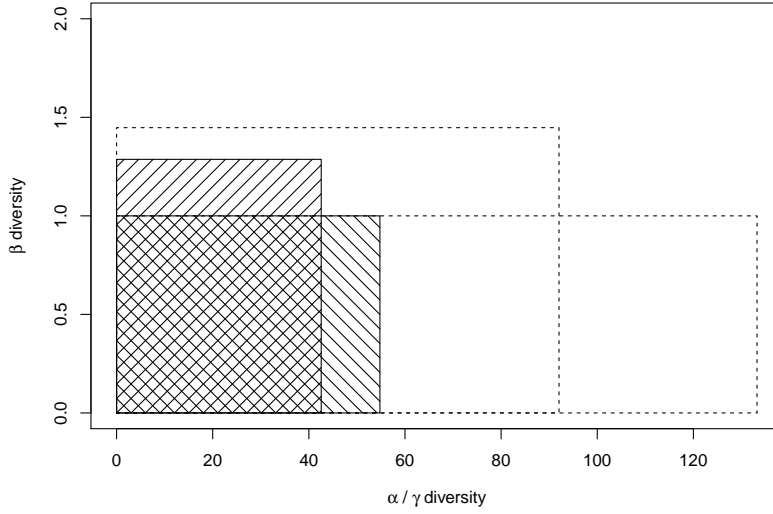


FIGURE 5.16 – Représentation graphique de la diversité des deux hectares de forêt de Paracou (parcelles 6 et 18). Les rectangles transparents représentent la diversité neutre, les rectangles hachurés la diversité phylogénétique (l'arbre est la taxonomie des espèces, $q = 1$). Dans chaque cas, le rectangle 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.

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 ici).

5.3.8 Partitionnement de la diversité de Leinster et Cobbold

Selon les mêmes principes que pour la décomposition de l'entropie HCDT, Marcon *et al.*¹⁶⁸ ont décomposé l'entropie de Ricotta et Szeidl.¹⁶⁹ L'entropie α de la communauté i est :

$${}^qH_{\alpha}^Z = \frac{1 - \sum_s p_{s,i}(\mathbf{Zp})_{s,i}^{q-1}}{q - 1} \quad (5.53)$$

L'entropie α est la somme pondérée des entropies des communautés :

$${}^qH_{\alpha}^Z = \sum_i w_i {}^qH_{\alpha}^Z = \frac{1 - \sum_i w_i \sum_s p_{s,i}(\mathbf{Zp})_{s,i}^{q-1}}{q - 1} \quad (5.54)$$

L'entropie β est similaire à l'entropie β HCDT : c'est la divergence généralisée de Jensen-Shannon entre les distributions de \mathbf{Zp} des communautés. Formellement :

$${}^qH_{\beta}^Z = \sum_i w_i \sum_s p_{s,i} \left(\ln_q \frac{1}{(\mathbf{Zp})_s} - \ln_q \frac{1}{(\mathbf{Zp})_{s,i}} \right) \quad (5.55)$$

¹⁶⁸MARCON *et al.* (2014b). « The Decomposition of Similarity-Based Diversity and its Bias Correction », cf. note 26, p. 15.

¹⁶⁹RICOTTA *et* SZEIDL (2006). « Towards a unifying approach to diversity measures : Bridging the gap between the Shannon entropy and Rao's quadratic index », cf. note 86, p. 48.

La diversité β est obtenue en calculant l'exponentielle déformée de la décomposition additive de l'entropie :

$${}^qD_\beta^Z = e_q^{\frac{{}^qH^Z}{1+(1-q){}^qH_\alpha^Z}} \quad (5.56)$$

La décomposition de la diversité peut être faite directement, sans passer par l'entropie.

L'inverse de ${}^qD_\gamma^Z$ est la moyenne généralisée d'ordre $q - 1$ de $(Zp)_s$:

$$\frac{1}{{}^qD_\gamma^Z} = \left[\sum_i p_s(Zp)_s^{q-1} \right]^{\frac{1}{q-1}} \quad (5.57)$$

Partant de ${}^qD_\gamma = {}^qD_\alpha {}^qD_\beta$ et de la définition de Routledge de la diversité α , on obtient :

$${}^qD_\beta^Z = \left[\sum_i w_i \left(\frac{1/{}_i^qD_\alpha^Z}{1/{}_i^qD_\gamma^Z} \right)^{q-1} \right]^{\frac{1}{q-1}} \quad (5.58)$$

De la même façon que $1/{}_i^qD_\gamma^Z$ est la banalité moyenne des espèces, ${}_i^qD_\beta^Z$ est la banalité moyenne normalisée des communautés, où la banalité normalisée de la communauté i est définie comme $(1/{}_i^qD_\alpha^Z)/(1/{}_i^qD_\gamma^Z)$, c'est-à-dire la banalité moyenne de ses espèces divisée par celle de la méta-communauté.

5.3.9 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.

L'indice de Simpson spatialement explicite

¹⁷⁰J. 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.

¹⁷¹R. CONDIT et al. (2002). « Beta-diversity in tropical forest trees ». In : *Science* 295.5555, p. 666–669.

¹⁷²G. SHEN et al. (2013b). « Quantifying spatial phylogenetic structures of fully stem-mapped plant communities ». In : *Methods in Ecology and Evolution* 4.12, p. 1132–1141.

Chave et Leigh¹⁷⁰ écrivent un modèle neutre de dispersion des espèces forestières, appliqué à l'échelle de l'Amazonie occidentale.¹⁷¹ 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-à-dire l'indice de concentration de Simpson appliqué à des paires d'individus distants. $\beta_s(r) = 1 - F(r)$ est appelé « indice de Simpson spatialement explicite » par Shen *et al.*¹⁷² En pratique il est estimé à 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,¹⁷³ voir page 58, 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 .

Une décomposition due à Nei¹⁷⁴ reprise par Chave *et al.*¹⁷⁵ 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 :

$$E_{i,j} = 1 - \sum_s p_{s,i} p_{s,j} \quad (5.59)$$

E_{ii} est l'entropie de Simpson calculée dans la parcelle i . L'entropie β de Simpson (équation 5.43, dans le cadre de la décomposition de la variance) pour des parcelles de poids égal vaut :

$$E_\beta = \frac{1}{I^2} \sum_i \sum_{j \neq i} \left(E_{i,j} - \frac{E_{ii} + E_{jj}}{2} \right) \quad (5.60)$$

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 $\beta_{phy}(r)$.

Le variogramme de complémentarité

Wagner¹⁷⁶ 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'indicateur $\mathbf{1}_{s,i}$ 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 :

$$\hat{\gamma}_s(r) = \frac{1}{2n_r} \sum_{i,j; d(i,j) \approx r} (\mathbf{1}_{s,i} - \mathbf{1}_{s,j})^2 \quad (5.61)$$

¹⁷³JURASINSKI et al. (2009). « Inventory, differentiation, and proportional diversity : a consistent terminology for quantifying species diversity », cf. note 112, p. 59.

¹⁷⁴NEI (1973). « Analysis of Gene Diversity in Subdivided Populations », cf. note 142, p. 65.

¹⁷⁵CHAVE et al. (2007). « The importance of phylogenetic structure in biodiversity studies », cf. note 59, p. 43.

¹⁷⁶H. H. WAGNER (2003). « Spatial covariance in plant communities : Integrating ordination, geostatistics, and variance testing ». In : *Ecology* 84.4, p. 1045-1057.

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 \mathbf{S}_i dont chaque terme est $\mathbf{1}_{s,i}$. La semi-variance de ce vecteur de composition spécifique est :

$$\hat{\gamma}(r) = \frac{1}{2n_r} \sum_{i,j;d(i,j) \approx r} \|\mathbf{S}_i - \mathbf{S}_j\|^2 = \sum_s \hat{\gamma}_s(r) \quad (5.62)$$

$\|\mathbf{S}_i - \mathbf{S}_j\|$ 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.

¹⁷⁷ G. BACARO et C. RICOTTA (2007). « A spatially explicit measure of beta diversity ». In : *Community Ecology* 8.1, p. 41–46.

Bacaro et Ricotta¹⁷⁷ l'utilisent en tant que mesure de diversité β .

$\mathbb{E}(\mathbf{1}_s)$ 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 :

$$\hat{\gamma}_{s,t}(r) = \frac{1}{2n_r} \sum_{i,j;d(i,j) \approx r} (\mathbf{1}_{s,i} - \mathbf{1}_{s,j})(\mathbf{1}_{t,i} - \mathbf{1}_{t,j}) \quad (5.63)$$

Wagner montre que l'estimateur de la variance de la richesse spécifique par parcelle est :

$$\hat{\text{Var}}(S) = \sum_r \frac{n_r}{\sum_r n_r} \sum_{s,t} \hat{\gamma}_{s,t}(r) \quad (5.64)$$

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 $\text{Var}(S)$ 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.

5.4 Estimation

5.4.1 Le biais d'estimation

Les mesures d'entropie sont sujettes à des biais d'estimation,¹⁷⁸ qui entraînent presque toujours une sous-estimation de la diversité. Plus précisément, ces biais sont dus à deux causes. La première est le non échantillonnage des espèces rares. Historiquement, Basharin¹⁷⁹ a évalué son effet sur l'estimation de l'entropie de Shannon. Elle a été très largement traitée par la littérature d'écologie statistique,¹⁸⁰ principalement dans le cas particulier de l'estimation du nombre d'espèces. La deuxième cause a été traitée par la littérature de physique statistique : il s'agit de la non-linéarité de l'entropie. Même si toutes les espèces sont observées dans un échantillon, estimer la probabilité d'une espèce p_s à la puissance $q > 0$ par son estimateur à la puissance q entraîne une sous-estimation systématique.¹⁸¹

5.4.2 Techniques d'estimation

Il existe quatre méthodes principales. La plus simple consiste à injecter dans la formule de la diversité l'estimateur de p_s , c'est-à-dire $\hat{p}_s = n_s/n$, pour obtenir l'estimateur dit « plug-in » ou « naïf ».

$${}^q\hat{H} = - \sum_s \hat{p}_s \ln_q \hat{p}_s \quad (5.65)$$

L'estimateur plug-in est inutilisable dans des communautés très diverses parce qu'il sous-estime sévèrement la diversité à cause des espèces non échantillonnées et de la non-linéarité des estimateurs.

La deuxième méthode repose sur l'estimateur développé par Horvitz et Thompson¹⁸² pour la somme pondérée d'une fonction de ses éléments, de la forme $\sum_s p_s f(s)$ quand une partie des termes ne sont pas observés.

Un estimateur non-biaisé de la somme est obtenue en divisant chaque terme par la probabilité qu'il soit observé : $1 - (1 - p_s)^n$. Chao et Shen¹⁸³ ont proposé de le combiner à l'estimateur du taux de couverture : conditionnellement aux espèces observées, un estimateur sans biais¹⁸⁴ de p_s est $\tilde{p}_s = \hat{C}\hat{p}_s$. Chao et Shen ont estimé l'entropie de Shannon ; la méthode a été étendue ensuite à l'entropie HCDT¹⁸⁵ et à la diversité de Leinster et Cobbold.¹⁸⁶

¹⁷⁸D. MOUILLOT et A. LEPRÊTRE (1999). « A comparison of species diversity estimators ». In : *Researches on Population Ecology* 41.2, p. 203–215.

¹⁷⁹G. 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.

¹⁸⁰J. 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.

¹⁸¹J. A. BONACHELA et al. (2008). « Entropy estimates of small data sets ». In : *Journal of Physics A : Mathematical and Theoretical* 41.202001, p. 1–9.

¹⁸²D. 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.

¹⁸³A. 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.

¹⁸⁴J. 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.

¹⁸⁵MARCON et al. (2014a). « Generalization of the partitioning of Shannon diversity », cf. note 23, p. 15.

¹⁸⁶MARCON et al. (2014b). « The Decomposition of Similarity-Based Diversity and its Bias Correction », cf. note 26, p. 15.

$${}^q\tilde{H} = - \sum_{s=1}^{s_{\neq 0}^n} \frac{\hat{C}\hat{p}_s \ln_q(\hat{C}\hat{p}_s)}{1 - (1 - \hat{C}\hat{p}_s)^n} \quad (5.66)$$

$s_{\neq 0}^n$ est le nombre d'espèces représentées par ν individus dans l'échantillon de taille n ; $s_{\neq 0}^n$ est le nombre d'espèces observées au moins une fois.

¹⁸⁷P. GRASSBERGER (1988). « Finite sample corrections to entropy and dimension estimates ». In : *Physics Letters A* 128.6-7, p. 369–373.

¹⁸⁸MARCON et al. (2014a). « Generalization of the partitioning of Shannon diversity », cf. note 23, p. 15.

La troisième méthode a été développée par Grassberger¹⁸⁷ qui a fourni un estimateur à biais réduit de la valeur d'un entier à la puissance q . La somme $\sum_s p_s^q$ est écrite comme $1/n^q \sum_s n_s^q$ et chaque terme n_s^q est estimé séparément.¹⁸⁸

$${}^q\tilde{H} = \frac{1 - \sum_{s=1}^{s_{\neq 0}^n} \widetilde{p}_s^q}{q - 1} \quad (5.67)$$

$$\widetilde{p}_s^q = n_s^{-q} \left(\frac{\Gamma(n_s + 1)}{\Gamma(n_s - q + 1)} + \frac{(-1)^n \Gamma(1 + q) \sin \pi q}{\pi(n + 1)} \right) \quad (5.68)$$

$\Gamma(\cdot)$ est la fonction gamma (pour $n \in \mathbb{N}$, $\Gamma(n + 1) = n!$).

¹⁸⁹Voir la revue dans 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, Appendix A.

La dernière méthode a généré une importante littérature dans les dix dernières années.¹⁸⁹ Elle repose sur l'estimation de la somme $h_q = \sum_{s=1}^S p_s^q$, qui peut être écrite de la façon suivante :

$$h_q = \sum_{r=0}^{\infty} \binom{q-1}{r} (-1)^r \zeta_q \quad (5.69)$$

¹⁹⁰Z. 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.

¹⁹¹Z. ZHANG (2013). « Asymptotic normality of an entropy estimator with exponentially decaying bias ». In : *IEEE Transactions on Information Theory* 59.1, p. 504–508.

¹⁹²Z. ZHANG et M. GRABCHAK (2014). « Entropic Representation and Estimation of Diversity Indices ». In : *arXiv* 1403.3031.v. 2, p. 1–12.

¹⁹³Z. ZHANG et M. GRABCHAK (2013). « Bias adjustment for a non-parametric entropy estimator ». In : *Entropy* 15.6, p. 1999–2011.

¹⁹⁴CHAO et JOST (2015). « Estimating diversity and entropy profiles via discovery rates of new species », cf. note 132, p. 62.

¹⁹⁵CHAO et al. (2013). Cf. note 189.

¹⁹⁶I. J. GOOD (1953). « The Population Frequency of Species and the Estimation of Population Parameters ». In : *Biometrika* 40.3/4, p. 237–264.

ζ_q est l'entropie de Simpson généralisée $\sum_s p_s(1 - p_s)^q$ définie par Zhang et Zhou.¹⁹⁰ Les n premiers éléments de la somme peuvent être estimés sans biais :

$$\tilde{h}_q = \sum_{s=1}^S \hat{p}_s \sum_{v=1}^{n-n_s} \left[\prod_{i=1}^v \frac{i-q}{i} \prod_{j=1}^v \left(1 - \frac{n_s-1}{n-j} \right) \right] \quad (5.70)$$

Zhang¹⁹¹ montre que le biais dû à l'ignorance des autres termes est asymptotiquement normal et décroît exponentiellement vite. J'appelle estimateur de Zhang et Grabchak¹⁹² l'estimateur limité à n termes.

Des tentatives ont été faites pour estimer le biais résiduel.¹⁹³ La plus aboutie est celle de Chao et Jost,¹⁹⁴ en complément de Chao, Wang et Jost.¹⁹⁵ Elle s'appuie sur deux hypothèses : le nombre total d'espèces est évalué par l'estimateur Chao1 et les probabilités des espèces observées le même nombre de fois ν peuvent être estimées toutes égales par $\hat{\alpha}_{\nu}$. Une conséquence est que \hat{s}_1^n et $n s_0$ sont égaux. L'estimateur Chao1 pour le nombre d'espèces manquantes combiné à la relation de Good-Turing¹⁹⁶ pour établir le rapport entre le nombre d'espèces observées ν fois

et $\alpha_{\nu+1}$ fait que les estimateurs de la probabilité des singletons et des espèces non observées sont égaux : $\hat{\alpha}_0 = \hat{\alpha}_1$. Leur valeur est notée A . Elle vaut

$$\frac{2s_2^n}{(n-1)s_1^n + 2s_2^n}$$

en présence de singletons et de doubletons,

$$\frac{2}{(n-1)(s_1^n - 1) + 2}$$

si les doubletons sont absents. En absence de singletons et doubletons, la valeur de A est fixée à 1. L'estimateur de Chao-Wang-Jost pour l'entropie HCDT est :

$${}^q\tilde{H} = \frac{1}{q-1} \left[1 - \tilde{h}_q - \frac{s_1^n}{n} (1-A)^{1-n} \left(A^{q-1} - \sum_{r=0}^{n-1} \binom{q-1}{r} (A-1)^r \right) \right] \quad (5.71)$$

J'ai proposé¹⁹⁷ deux nouveaux estimateurs à partir des progrès récents dans l'estimation de la distribution réelle des probabilités des espèces.¹⁹⁸ Les estimateurs conditionnels des probabilités des espèces observées \tilde{p}_s peuvent être améliorées en ajustant un modèle de distribution aux données. Ces nouveaux estimateurs de probabilités peuvent être injectés dans l'estimateur de Chao-Shen pour améliorer ses performances de façon significative. L'ajustement de la distribution des probabilités repose sur l'estimation du taux de couverture généralisé, d'où le nom d'estimateur « du taux de couverture généralisé ».

La distribution des espèces non échantillonnées peut être ajoutée en estimant leur nombre et en choisissant une loi. Chao *et al.*¹⁹⁹ ont utilisé l'estimateur Chao1 et une distribution géométrique. J'ai utilisé l'estimateur jackknife²⁰⁰ parce que son ordre peut être adapté aux données quand l'effort d'échantillonnage est trop faible pour que l'estimateur Chao1 soit performant.²⁰¹ Un simple estimateur plug-in, appelé « estimateur révélé », peut ensuite être appliqué à cette distribution révélée.

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.

5.4.3 Pratique de l'estimation

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.²⁰² Un résultat représentatif est donné ici.

¹⁹⁷É. MARCON (2015). « Practical Estimation of Diversity from Abundance Data ». In : *HAL* 01212435, version 1, p. 1–27.

¹⁹⁸A. CHAO et al. (2015). « Unveiling the Species-Rank Abundance Distribution by Generalizing Good-Turing Sample Coverage Theory ». In : *Ecology* 96.5, p. 1189–1201.

¹⁹⁹Ibid.

²⁰⁰K. 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.

²⁰¹U. BROSE et al. (2003). « Estimating species richness : Sensitivity to sample coverage and insensitivity to spatial patterns ». In : *Ecology* 84.9, p. 2364–2377.

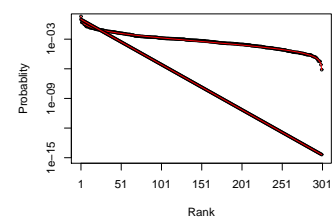


FIGURE 5.17 – Courbes rang-abondance de 300 espèces suivant une distribution log-normale (courbe du haut, logarithme de l'écart-type égal à 2) ou géométrique (ligne droite, paramètre 0,1 : la fréquence de chaque espèce est 0,9 fois celle de l'espèce immédiatement plus abondante). Les courbes rouges sont les distributions théoriques ajustées.

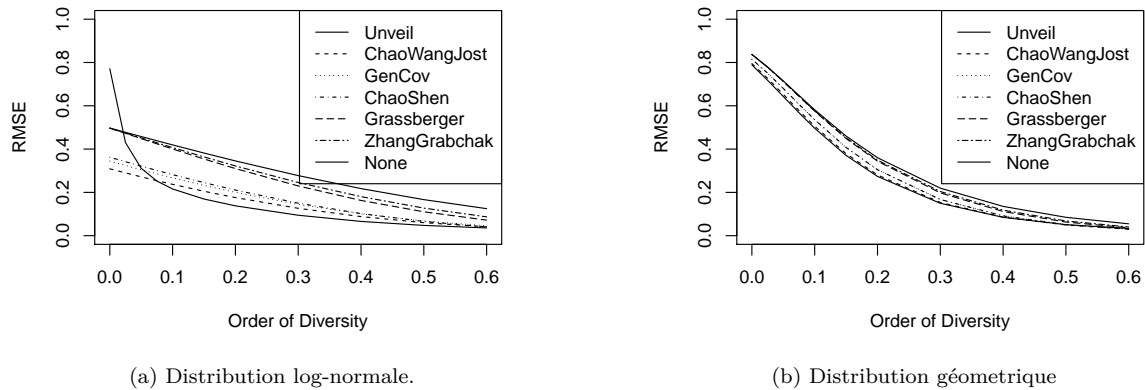


FIGURE 5.18 – Erreur relative des estimateurs de diversité estimée à partir de 1000 échantillons de 1000 individus de chaque distribution (log-normale et géométrique) de 300 espèces. L'erreur est très importante pour les ordres de diversité q faibles, particulièrement pour la distribution géométrique. Les valeurs de q supérieures à 0,6 ne sont pas montrées parce que les estimateurs y ont des performances de plus en plus similaires. La légende liste les estimateurs par erreur croissante pour $q > 0,1$, où le meilleur est l'estimateur révélé, puis celui de Chao-Wang-Jost. Près de $q = 0$, l'estimateur révélé a la plus forte variance, ce qui en fait l'estimateur le moins fiable.

La figure 5.18 compare les performances des estimateurs. Les données de test sont deux distributions de 300 espèces (Figure 5.17), une log-normale (correspondant aux arbres d'une forêt tropicale par exemple) et une géométrique (correspondant plutôt à une communauté microbienne), dont la queue de distribution est plus longue. Le critère d'évaluation est l'erreur relative moyenne. L'erreur moyenne est la somme du biais d'estimation au carré et de la variance de l'estimateur, évalués à partir de simulations. Sa racine carrée (RMSE : *root-mean-squared error*) est normalisée par la vraie valeur pour obtenir l'erreur relative moyenne. Elle est de l'ordre de 80% pour l'estimation du nombre d'espèces de la distribution géométrique (Figure 5.18b), ce qui invalide toute tentative d'application à des données réelles. Elle tombe autour de 10% dès $q = 0,3$ pour la distribution log-normale avec les meilleurs estimateurs, qui sont celui de Chao-Wang-Jost et l'estimateur révélé.

La figure 5.19 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 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.

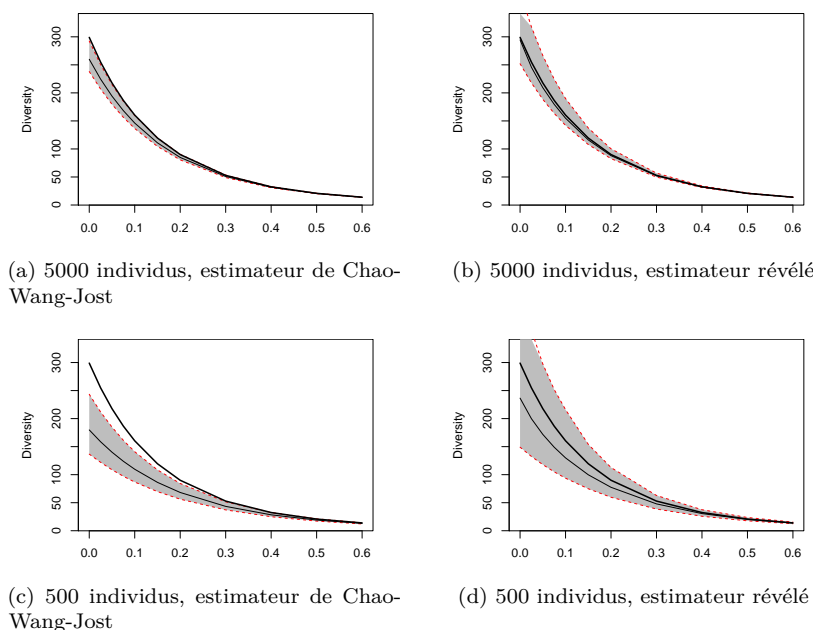


FIGURE 5.19 – Estimation de la diversité d’une communauté log-normale de 300 espèces. Un inventaire de 500 (en bas) ou 5000 (en haut) individus est répété 1000 fois et la diversité estimée à chaque fois par l’estimateur de Chao-Wang-Jost (à gauche) ou l’estimateur révélé (à droite). La courbe noire en 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.

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 jackknife 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). Dans des conditions beaucoup plus sévères (des échantillons microbiens dans des communautés de plusieurs millions d’espèces avec une distribution géométrique de paramètre $1/2$, c’est-à-dire dans laquelle chaque espèce est deux fois moins fréquente que celle qui lui est immédiatement plus abondante), Haegeman *et al.*²⁰³ ont montré que la diversité ne pouvait être estimée correctement qu’à

²⁰³B. HAEGEMAN et al. (2013). « Robust estimation of microbial diversity in theory and in practice ». In : *The ISME journal* 7.6, p. 1092–101.

partir de $q = 1$.

Tous ces estimateurs sont applicables à la phylodiversité ${}^q\bar{D}(T)$ (ils sont appliqués à l'entropie à chaque période k , puis sommés et l'entropie est transformée en diversité) comme à la diversité neutre. Les estimateurs de la diversité de Leinster et Cobbold sont différents.²⁰⁴ Tous sont implémentés dans le package *entropart*.²⁰⁵

²⁰⁴MARCON et al. (2014b). « The Decomposition of Similarity-Based Diversity and its Bias Correction », cf. note 26, p. 15.

²⁰⁵MARCON et HÉRAULT (2015b). « entropart, an R Package to Measure and Partition Diversity », cf. note 27, p. 16.

²⁰⁶PATIL et TAILLIE (1982). « Diversity as a concept and its measurement », cf. note 21, p. 15.

²⁰⁷MARCON et HÉRAULT (2015a). « Decomposing Phylodiversity », cf. note 24, p. 15.

²⁰⁸MARCON et al. (2012b). « The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity », cf. note 20, p. 15; MARCON et al. (2014a). « Generalization of the partitioning of Shannon diversity », cf. note 23, p. 15; MARCON et HÉRAULT (2015a). « Decomposing Phylodiversity », cf. note 24, p. 15.

²⁰⁹MARCON et HÉRAULT (2015a). « Decomposing Phylodiversity », cf. note 24, p. 15; MARCON et al. (2014b). « The Decomposition of Similarity-Based Diversity and its Bias Correction », cf. note 26, p. 15.

²¹⁰MARCON et HÉRAULT (2015b). « entropart, an R Package to Measure and Partition Diversity », cf. note 27, p. 16.

5.5 Conclusion

Mes travaux dans le domaine de la mesure de la biodiversité sont plus intriqués dans la littérature que ceux concernant les statistiques spatiales. Il est donc utile ici de faire un bilan de ma contribution réelle.

J'ai généralisé le cadre théorique établi par Patil et Taillie²⁰⁶ pour la diversité neutre, qui permet de définir d'une manière unique la diversité et la dualité entropie-diversité :

- l'entropie est une mesure de surprise (ou d'information, de rareté) qui permet de comparer des communautés ;
- la diversité est une transformation monotone de l'entropie (son exponentielle) qui permet de traduire l'entropie en nombre effectif d'espèces, y compris pour la diversité phylogénétique et la diversité fonctionnelle ;²⁰⁷
- l'entropie β est une divergence moyenne entre chaque communauté et la méta-communauté ;²⁰⁸
- la diversité β est un nombre effectif de communautés, y compris pour la diversité phylogénétique et la diversité fonctionnelle ;²⁰⁹

J'ai systématiquement développé des estimateurs pour les mesures de diversité et un package²¹⁰ pour R permettant leur application.

CHAPITRE 6

Perspectives

LES travaux présentés dans ce mémoire de synthèse concernent la méthodologie de l'écologie et de l'économie. Je me suis efforcé de proposer des outils et des méthodes pour caractériser la concentration spatiale ou la diversité qui dépassent les simples indices mais permettent la mesure, c'est-à-dire une quantification physique des phénomènes concernés. Les mesures des statistiques spatiales que j'ai proposées (y compris dans le cadre de la simple normalisation de fonctions existantes) sont des quotients de localisation¹ : leur valeur est le rapport entre le nombre ou des proportions de voisins observés et le nombre ou la proportion attendue. Les mesures de diversité sont des nombres effectifs d'espèces ou de communautés. Leur estimation est permise par des méthodes mathématiques et les outils informatiques appropriés.

De nombreuses questions méthodologiques restent à traiter. La spatialisation de la diversité est par exemple problématique. La simple cartographie de la diversité α n'est pas triviale² parce que la diversité augmente avec le grain de la représentation. La cartographie d'une variable est simple quand le passage à une échelle plus grossière revient simplement à moyenner sa valeur (par exemple, la densité d'une population). Ce n'est pas le cas pour la diversité.

Un autre sujet méthodologique est le traitement de l'incertitude de la mesure. Je travaille par exemple à la mise au point d'un test statistique, déjà ébauché,³ analogue à l'ANOVA pour l'analyse de la diversité β au-delà de deux communautés : il s'agit de tester si la diversité observée peut être le seul effet de la stochasticité, ce qui signifie que les communautés testées peuvent être considérées comme des échantillons d'une même communauté plus vaste.

Un dernier exemple est le transfert vers d'autres objets des méthodes que j'ai développées. La caractérisation des réseaux trophiques en terme de nombre effectif de flux et de rôles⁴ est très similaire à l'approche des nombres de Hill, mais limitée à l'entropie de Shannon. La généralisation de l'approche d'Ulanowicz

¹P. S. FLORENCE (1972). *The Logic of British and American Industry : A Realistic Analysis of Economic Structure and Government*. 3rd ed. London : Routledge & Kegan Paul.

²V. GRANGER et al. (2015). « Mapping diversity indices : not a trivial issue ». In : *Methods in Ecology and Evolution* 6.6, p. 688–696.

³C. 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.

⁴R. E. ULANOWICZ et al. (2014). « Limits on ecosystem trophic complexity : Insights from ecological network analysis ». In : *Ecology Letters* 17.2, p. 127–136.

est relativement simple à mettre en œuvre.

Je présente en détail ici deux sujets de recherche méthodologique : la mesure spatialement explicite de la diversité, qui est l'étape théorique préalable à la cartographie, et le transfert des méthodes de mesure de la diversité à l'économie géographique.

J'engage progressivement des recherches nouvelles, au-delà de la méthodologie, pour répondre à des questions d'écologie ou d'économie. Je présente ici la première, traitée en profondeur dans le cadre d'une thèse, qui consiste à comprendre l'impact de l'exploitation forestière sur les divers aspects de la diversité des arbres.

6.1 Mesure de la diversité spatialement explicite

⁵J. B. PLOTKIN et al. (2000). « Species-area curves, spatial aggregation, and habitat specialization in tropical forests ». In : *Journal of Theoretical Biology* 207.1, p. 81–99.

⁶SHIMATANI (2001). « Multivariate point processes and spatial variation of species diversity », cf. note 63, p. 44.

⁷G. SHEN et al. (2013a). « Quantifying effects of habitat heterogeneity and other clustering processes on spatial distributions of tree species ». In : *Ecology* 94.11, p. 2436–2443.

⁸MARCON et al. (2014a). « Generalization of the partitioning of Shannon diversity », cf. note 23, p. 15; MARCON et HÉRAULT (2015a). « Decomposing Phylodiversity », cf. note 24, p. 15.

⁹R. 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.

¹⁰L. GÖTZENBERGER et al. (2012). « Ecological assembly rules in plant communities - approaches, patterns and prospects ». In : *Biological Reviews* 87.1, p. 111–127.

¹¹A. MARSHALL (1890). *Principle of Economics*. London : Macmillan ; A. WEBER (1909). *Über den Standort der Industrien*. Tübingen. English translation edited in 1971, "Theory of the location of industries", Russell & Russell ; P. KRUGMAN (1991). *Geography and Trade*. London : MIT Press.

¹²M. HOUEBINE (1999). « Concentration Géographique des Activités et Spécialisation des Départements Français ». In : *Economie et Statistique* 326-327.6-7, p. 189–204 ; CUTRINI (2010). « Specialization and Concentration from a Twofold Geographical Perspective : Evidence from Europe », cf. note 15, p. 15.

La définition classique de la diversité α s'applique à un ensemble d'individus sans structure spatiale, qui est plutôt considérée comme un problème éventuel pour l'échantillonnage⁵ que comme une caractéristique écologique.

Une approche différente a été initiée par Shimatani⁶ qui a défini une entropie de Simpson dépendant de la distance, et l'a reliée à la fonction K de Ripley. Cette idée a été reprise récemment⁷ et étendue à l'entropie de Rao pour définir une diversité phylogénétique spatialement explicite. Une application immédiate en écologie consiste à comparer la relation entre diversité et distance à sa valeur sous diverses hypothèses nulles : absences de structuration spatiale, phylogénétique ou fonctionnelle.

Un important travail de conceptualisation reste à faire pour :

- généraliser cette approche à la diversité paramétrique, au delà de l'entropie de Rao qui ne prend en compte que les espèces dominantes ;⁸
- définir une relation théorique entre la diversité et la distance, similaire à la courbe d'accumulation des espèces ;⁹
- relier les écarts à la distribution théorique aux processus écologiques.¹⁰

6.2 Transfert à l'économie géographique des méthodes de la biodiversité

Les recherches sur la structure spatiale de l'activité économique se sont principalement intéressées à la concentration spatiale, source d'externalités positives.¹¹ La concentration spatiale va de pair avec la spécialisation.¹² Le cadre conceptuel est le suivant : des employés peuvent être localisés dans une région quelconque d'un pays donné, et travailler dans un secteur économique quelconque. Les données sont le nombre d'employés de chaque secteur dans

chaque région. Sous l'hypothèse nulle d'une distribution non structurée, la connaissance de la taille relative de chaque secteur et de chaque région donne l'espérance de ce nombre. La concentration spatiale d'un secteur économique mesurée par l'indice d'Ellison et Glaeser¹³ est l'écart entre la part de chaque région dans ce secteur et la taille relative des régions. De façon symétrique, la spécialisation d'une région peut être définie comme l'écart de la distribution des poids relatifs de ses secteurs économiques à leurs poids dans l'ensemble du pays. Les deux peuvent être combinés pour définir une mesure de diversité jointe,¹⁴ écart entre la distribution des couples secteur \times région et leur valeur attendue en absence de structuration. Cutrini a défini cette diversité jointe, mesurée par l'entropie de Shannon, comme un « indice de localisation globale ».

Les développements méthodologiques du domaine de la diversité peuvent être appliqués à ce cadre pour généraliser cette mesure de localisation globale à l'entropie HCDT. Les problèmes classiques de sensibilité des mesures de concentration spatiale et de spécialisation en espace discret¹⁵ peuvent être largement réduits en considérant l'emboîtement des échelles spatiales ou des secteurs économiques plus ou moins agrégés de la même façon qu'une phylogénie. À titre d'exemple, le problème d'échelle est l'incohérence des mesures de concentration géographique considérées à des échelles différentes. Sa résolution théorique est immédiate dans le cadre de la décomposition de la diversité présentée au chapitre précédent : la concentration (ou la spécialisation) à un niveau agrégé (par exemple un pays) est égale à sa valeur moyenne au niveau désagrégé (par exemple les régions de ce pays) à laquelle s'ajoute la divergence entre régions (analogue à la diversité β) dont l'ignorance est le fondement du problème.

6.3 Trajectoires de la diversité

La possibilité de gestion durable des forêts tropicales implique l'existence d'un régime périodique de fonctionnement de la forêt exploitée, forcément différent de celui de la forêt primaire, mais soutenable à long terme. Cette question a été traitée largement du point de vue des gestionnaires forestiers, sous l'angle du maintien de la ressource disponible en essences commerciales.¹⁶ Elle a été complétée plus récemment par l'étude de la reconstitution du stock de carbone entre deux exploitations,¹⁷ en lien avec les préoccupations liées au cycle du carbone.

Les connaissances sur l'impact de l'exploitation sur la biodiversité sont plus fragmentaires. Les approches classiquement utilisées évaluent ponctuellement, à court ou moyen terme, l'impact de l'exploitation sur la richesse spécifique ou le cortège floristique.¹⁸ Le rôle des forêts exploitées dans les politiques de conservation de la biodiversité nécessite de mieux prendre en compte (1) la dynamique temporelle post-exploitation ainsi que (2) les différentes

¹³ELLISON et GLAESER (1997). « Geographic Concentration in U.S. Manufacturing Industries : A Dartboard Approach », cf. note 17, p. 23.

¹⁴GREGORIUS (2010). « Linking Diversity and Differentiation », cf. note 39, p. 40.

¹⁵G. ARBIA (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Dordrecht : Kluwer ; S. OPENSHAW et P. J. TAYLOR (1979). « A million or so correlation coefficients : three experiments on the modifiable areal unit problem ». In : *Statistical Applications in the Spatial Sciences*. Sous la dir. de N WRIGLEY. London : Pion, p. 127–144.

¹⁶GOURLET-FLEURY et al. (2004). *Ecology & management of a neotropical rainforest. Lessons drawn from Paracou, a long-term experimental research site in French Guiana*, cf. note 23, p. 25 ; S. GOURLET-FLEURY et al. (2005). « Using models to predict recovery and assess tree species vulnerability in logged tropical forests : A case study from French Guiana ». In : *Forest Ecology and Management* 209.1-2, p. 69–86.

¹⁷BLANC et al. (2009). « Dynamics of aboveground carbon stocks in a selectively logged tropical forest », cf. note 4, p. 14.

¹⁸C. B. HALPERN et T. A. SPIES (1995). « Plant species diversity in natural and managed forests of the Pacific Northwest ». In : *Ecological Applications* 5.4, p. 913–934.

¹⁹S. G. LETCHER et al. (2012). « Phylogenetic community structure during succession : Evidence from three Neotropical forest sites ». In : *Perspectives in Plant Ecology, Evolution and Systematics* 14.2, p. 79–87.

²⁰C. H. CANNON et al. (1998). « Tree Species Diversity in Commercially Logged Bornean Rainforest ». In : *Science* 281.5381, p. 1366–1368.

²¹S. A. SCHNITZER et W. P. CARSON (2001). « Treefall Gaps and the Maintenance of Species Diversity in a Tropical Forest ». In : *Ecology* 82.4, p. 913–919.

²²O. L. OSAZUWA-PETERS et al. (2015). « Selective logging : does the imprint remain on tree structure and composition after 45 years? » In : *Conservation Physiology* 3.1, cov012.

²³S. DUFOUR-KOWALSKI et al. (2012). « Capsis : An open software framework and community for forest growth modelling ». In : *Annals of Forest Science* 69.2, p. 221–233.

²⁴S. GUITET et al. (2014). « Estimating tropical tree diversity indices from forestry surveys : A method to integrate taxonomic uncertainty ». In : *Forest Ecology and Management* 328, p. 270–281.

dimensions (spécifique, phylogénétique et fonctionnelle)¹⁹ de cette diversité. Les conséquences de l'exploitation forestière ne sont pas facilement prévisibles,²⁰ de l'augmentation de la diversité apportée par les perturbations régulières²¹ à l'appauvrissement de l'écosystème dû à la disparition des espèces de succession tardive,²² exploitées pour leur intérêt technologique.

Je dirige une thèse sur ce sujet, dont l'objectif principal sera de caractériser des trajectoires de biodiversité (spécifique, phylogénétique, fonctionnelle) au cours du cycle d'exploitation. La stabilité et la robustesse de ces trajectoires au cours des cycles d'exploitation est un critère essentiel de la durabilité. Des développements méthodologiques seront nécessaires : la prise en compte de l'incertitude sur la détermination botanique et de la variabilité intraspécifique des traits fonctionnels notamment ; l'intégration des mesures aux simulateurs existants (Selva, sous Capsis²³).

Les données de Paracou, fournissant un recul de trente années après l'exploitation et différents niveaux de prélèvement, permettent ces analyses. La détermination botanique des parcelles exploitées n'est pas complète mais devrait progresser pendant la thèse. Les méthodes développées récemment²⁴ seront utilisées et étendues pour évaluer la biodiversité à partir de relevés des essences forestières (et non des espèces botaniques), avec l'incertitude associée : ce type de données est le seul généralement disponible.

6.4 Conclusion

Au-delà des pistes de recherche immédiates détaillées ici, ma ligne directrice consiste toujours à caractériser aussi précisément que possible les patrons de diversité ou de distribution spatiale pour pouvoir les relier ensuite aux processus.

En économie géographique, la théorie des externalités positives prévoit que la croissance soit liée positivement à la concentration spatiale. Pour le vérifier empiriquement, il est nécessaire de définir une mesure de concentration spatiale dont la valeur soit comparable entre les jeux de données, et dont la quantification soit sans ambiguïté.

En écologie, la diversité peut être par exemple utilisée comme variable explicative de la stabilité d'un écosystème. Le paramétrage de l'ordre de la diversité, et son expression en tant que nombre effectif d'espèces, permettent de choisir l'importance donnée aux espèces rares tout en disposant d'une variable explicative comparable d'un scénario à l'autre.

Mes travaux sur la caractérisation sont assez avancés pour que je commence à m'intéresser à ces questions. A court terme, la méthodologie restera mon activité de recherche principale, avec de plus en plus d'application à des données réelles dans le cadre de

projets de recherches plus vastes : estimation de la biodiversité par metabarcoding (nécessitant des développements méthodologiques pour prendre en compte l'incertitude sur les données²⁵) ou cartographie de la diversité à partir d'imagerie hyperspectrale²⁶ par exemple. La modélisation des processus générant les distributions observées (distributions de probabilités et distributions spatiales) sera l'étape suivante mais nécessite d'acquérir plus d'expertise sur les objets pour dépasser la seule approche méthodologique.

²⁵G. F. FICETOLA et al. (2015). « Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data ». In : *Molecular Ecology Resources* 15.3, p. 543–556.

²⁶J.-B. FÉRET et G. P. ASNER (2014). « Mapping tropical forest canopy diversity using high-fidelity imaging spectroscopy ». In : *Ecological Applications* 24.6, p. 1289–1296.

Bibliographie

- 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. 37).
- 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. 70).
- ANDERSON, M. J., T. O. CRIST, J. M. CHASE, M. VELLEND, B. D. INOUE, 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. 58).
- ARBA, G. (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Dordrecht : Kluwer (cf. p. 83).
- 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. 75).
- BACARO, G. et C. RICOTTA (2007). « A spatially explicit measure of beta diversity ». In : *Community Ecology* 8.1, p. 41–46 (cf. p. 74).
- BADDELEY, A. J., J. MØLLER et R. P. WAAGEPETERSEN (2000). « Non- and semi-parametric estimation of interaction in inhomogeneous point patterns ». In : *Statistica Neerlandica* 54.3, p. 329–350 (cf. p. 30).
- BARALOTO, C., C. E. T. PAINE, S. PATIÑO, D. BONAL, B. HÉRAULT et J. CHAVE (2010a). « Functional trait variation and sampling strategies in species rich plant communities ». In : *Functional Ecology* 24, p. 208–216 (cf. p. 46).
- BARALOTO, C., E. MARCON, F. MORNEAU, S. PAVOINE et J.-C. ROGGY (2010b). « Integrating functional diversity into tropical forest plantation designs to study ecosystem processes ». In : *Annals of Forest Science* 67.3, p. 303 (cf. p. 14).
- BASELGA, A. (2010). « Multiplicative partition of true diversity yields independent alpha and beta components ; additive partition does not ». In : *Ecology* 91.7, p. 1974–1981 (cf. p. 58, 59).
- 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. 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. 75).
- BECK, J., J. D. HOLLOWAY et W. SCHWANGHART (2013). « Undersampling and the measurement of beta diversity ». In : *Methods in Ecology and Evolution* 4.4, p. 370–382 (cf. p. 67).
- BEHRENS, K. et T. BOUGNA (2015). « An Anatomy of the Geographical Concentration of Canadian Manufacturing Industries ». In : *Regional Science and Urban Economics* 51, p. 47–69 (cf. p. 34).
- 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. 41).
- BESAG, J. E. (1977). « Comments on Ripley's paper ». In : *Journal of the Royal Statistical Society B* 39.2, p. 193–195 (cf. p. 14, 26).
- BESAG, J. E. et P. J. DIGGLE (1977). « Simple Monte Carlo Tests for Spatial Pattern ». In : *Applied Statistics* 26.3, p. 327–333 (cf. p. 24, 26).
- BLANC, L., M. ECHARD, B. HÉRAULT, D. BONAL, É. MARCON, J. CHAVE et C. BARALOTO (2009). « Dynamics of aboveground carbon stocks in a selectively logged tropical forest ». In : *Ecological Applications* 19.6, p. 1397–1404 (cf. p. 14, 83).
- 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. 75).
- BONAL, D., C. BORN, C. BRECHET, S. COSTE, É. MARCON, J. C. ROGGY et J. M. GUEHL (2007). « The successional status of tropical rainforest tree species is associated with differences in leaf carbon isotope discrimination and functional traits ». In : *Annals of Forest Science* 64.2, p. 169–176 (cf. p. 14).
- BORLAND, L., A. R. PLASTINO et C. TSALLIS (1998). « Information gain within nonextensive thermostatics ». In : *Journal of Mathematical Physics* 39.12, p. 6490–6501 (cf. p. 64).
- BOURGUIGNON, F. (1979). « Decomposable Income Inequality Measures ». In : *Econometrica* 47.4, p. 901–920 (cf. p. 63).
- 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. 77).
- 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. 30).
- 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. 77).
- BUUREN, S. van et K. GROOTHUIS-ODSHOORN (2011). « mice : Multivariate Imputation by Chained Equations in R ». In : *Journal of Statistical Software* 45.3, p. 1–67 (cf. p. 46).
- BUUREN, S. van, J. P. L. BRAND, C. G. M. GROOTHUIS-ODSHOORN et D. B. RUBIN (2006). « Fully conditional specification in multivariate imputation ». In : *Jour-*

- nal of Statistical Computation and Simulation* 76.12, p. 1049–1064 (cf. p. 46).
- CAMPOS, D. et J. F. ISAZA (2009). « A geometrical index for measuring species diversity ». In : *Ecological Indicators* 9.4, p. 651–658 (cf. p. 55).
- CANNON, C. H., D. R. PEART et M. LEIGHTON (1998). « Tree Species Diversity in Commercially Logged Bornean Rainforest ». In : *Science* 281.5381, p. 1366–1368 (cf. p. 84).
- 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. 35).
- 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. 30).
- CHAMPELY, S. et D. CHESSEL (2002). « Measuring biological diversity using Euclidean metrics ». In : *Environmental and Ecological Statistics* 9.2, p. 167–177 (cf. p. 67).
- CHAO, A. (2004). « Species richness estimation. » In : *Encyclopedia of Statistical Sciences*. Sous la dir. de N BALAKRISHNAN, C. B. READ et B. VIDAKOVIC. 2nd ed. New York : Wiley (cf. p. 38).
- CHAO, A. 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 (cf. p. 62, 76).
- 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. 75).
- 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. 69).
- 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. 40, 44, 45).
- 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. 58–60, 65, 70).
- 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. 76).
- CHAO, A., T. C. HSIEH, R. L. CHAZDON, R. K. COLWELL et N. J. GOTELLI (2015). « Unveiling the Species-Rank Abundance Distribution by Generalizing Good-Turing Sample Coverage Theory ». In : *Ecology* 96.5, p. 1189–1201 (cf. p. 77).
- CHAPIN, F. S. I., E. S. ZAVALA, V. T. EVINER, R. L. NAYLOR, P. M. VITOUSEK, H. L. REYNOLDS, D. U. HOOPER, S. LAVOREL, O. E. SALA, S. E. HOBIE, M. C. MACK et S. DIAZ (2000). « Consequences of changing biodiversity ». In : *Nature* 405.6783, p. 234–242 (cf. p. 35).
- 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. 72).
- 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. 43, 73).
- 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. 49, 54).
- CHIU, C.-H., L. JOST et A. CHAO (2014). « Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers ». In : *Ecological Monographs* 84.1, p. 21–44 (cf. p. 61).
- CHIU, S. N. (2007). « Correction to Koen's critical values in testing spatial randomness ». In : *Journal of Statistical Computation and Simulation* 77.11-12, p. 1001–1004 (cf. p. 27).
- 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. 82).
- COMBES, P.-P. et H. G. OVERMAN (2004). « The spatial distribution of economic activities in the European Union ». In : *Handbook of Urban and Regional Economics*. Sous la dir. de J. V. HENDERSON et J.-F. THISSE. T. 4. Amsterdam : Elsevier. North Holland. Chap. 64, p. 2845–2909 (cf. p. 30).
- COMBES, P.-P., T. MAYER et J.-F. THISSE (2008). *Economic Geography*. Princeton, New Jersey : Princeton University Press, p. 1–416 (cf. p. 13).
- CONDIT, R. (1998). *Tropical Forest Census Plots*. Berlin, Germany, et Georgetown, Texas : Springer-Verlag et R. G. Landes Company, p. 1–224 (cf. p. 68).
- CONDIT, R., N. PITMAN, E. G. J. LEIGH, J. CHAVE, J. TERBORGH, R. B. FOSTER, P. NÚÑEZ, S. AGUILAR, R. VALENZUELA, 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. 72).
- 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. 46).
- COSTE, S., C. BARALOTO, C. LEROY, É. MARCON, A. REINAUD, A. D. RICHARDSON, J.-C. ROGGY, H. SCHIMANN, J. UDDLING et B. HÉRAULT (2010). « Assessing foliar chlorophyll contents with the SPAD-502 chlorophyll meter : a calibration test with thirteen tree species of tropical rainforest in French Guiana ». In : *Annals of Forest Science* 67.6, p. 607 (cf. p. 14).
- COUSINS, S. H. (1991). « Species diversity measurement : Choosing the right index ». In : *Trends in Ecology and Evolution* 6.6, p. 190–192 (cf. p. 43).
- COX, D. R. (1955). « Some Statistical Methods Connected with Series of Events ». In : *Journal of the Royal Statistical Society B* 17.2, p. 129–164 (cf. p. 22).
- CRESSIE, N. A. (1993). *Statistics for spatial data*. New York : John Wiley & Sons, p. 1–900 (cf. p. 20, 27).
- 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. 67, 68).
- CUTRINI, E. (2010). « Specialization and Concentration from a Twofold Geographical Perspective : Evidence from Europe ». In : *Regional Studies* 44.3, p. 315–336 (cf. p. 15, 82).
- DALTON, H. (1920). « The measurement of the inequality of incomes ». In : *The Economic Journal* 30.119, p. 348–361 (cf. p. 37).

- DARÓCZY, Z. (1970). « Generalized information functions ». In : *Information and Control* 16.1, p. 36–51 (cf. p. 38).
- 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. 38, 40).
- DAVIS, H. T. (1941). *The theory of econometrics*. Bloomington, Indiana : The Principia Press (cf. p. 35).
- DE BELLO, F., S. LAVERGNE, C. N. MEYNARD, J. LEPSŠ 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. 67).
- DIGGLE, P. J. (1983). *Statistical analysis of spatial point patterns*. London : Academic Press, p. 1–148 (cf. p. 19–21, 27).
- DIGGLE, P. J. et A. G. CHETWYND (1991). « Second-Order Analysis of Spatial Clustering for Inhomogeneous Populations ». In : *Biometrics* 47.3, p. 1155–1163 (cf. p. 34).
- DUFOUR-KOWALSKI, S., B. COURBAUD, P. DREYFUS, C. MEREDIEU et F. DE COLIGNY (2012). « Capsis : An open software framework and community for forest growth modelling ». In : *Annals of Forest Science* 69.2, p. 221–233 (cf. p. 84).
- DURANTON, G. et H. G. OVERMAN (2005). « Testing for Localisation Using Micro-Geographic Data ». In : *Review of Economic Studies* 72.4, p. 1077–1106 (cf. p. 27, 30–32, 34).
- ELLISON, A. M. (2010). « Partitioning diversity ». In : *Ecology* 91.7, p. 1962–1963 (cf. p. 58).
- ELLISON, G. et E. L. GLAESER (1997). « Geographic Concentration in U.S. Manufacturing Industries : A Dartboard Approach ». In : *Journal of Political Economy* 105.5, p. 889–927 (cf. p. 23, 31, 83).
- FAITH, D. P. (1992). « Conservation evaluation and phylogenetic diversity ». In : *Biological Conservation* 61.1, p. 1–10 (cf. p. 44).
- FELLER, W. (1943). « On a general class of contagious distributions ». In : *The Annals of Mathematical Statistics* 14, p. 389–400 (cf. p. 23).
- FÉRET, J.-B. et G. P. ASNER (2014). « Mapping tropical forest canopy diversity using high-fidelity imaging spectroscopy ». In : *Ecological Applications* 24.6, p. 1289–1296 (cf. p. 85).
- FICETOLA, G. F., J. PANSU, A. BONIN, E. COISSAC, C. GIGUET-COVEX, M. DE BARBA, L. GIELLY, C. M. LOPES, F. BOYER, F. POMPANON, G. RAYÉ et P. TABERLET (2015). « Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data ». In : *Molecular Ecology Resources* 15.3, p. 543–556 (cf. p. 85).
- FLORENCE, P. S. (1972). *The Logic of British and American Industry : A Realistic Analysis of Economic Structure and Government*. 3rd ed. London : Routledge & Kegan Paul (cf. p. 81).
- GADAGKAR, R. (1989). « An undesirable property of Hill's diversity index N_2 ». In : *Oecologia* 80, p. 140–141 (cf. p. 41).
- GIGNOUX, J., C. DUBY et S. BAROT (1999). « Comparing the performances of Diggle's test of spatial randomness for small samples with or without edge effect correction : application to ecological data ». In : *Biometrics* 55.1, p. 156–164 (cf. p. 28).
- GINI, C. (1912). *Variabilità e mutabilità*. Bologna : C. Cuppini (cf. p. 30).
- 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. 76).
- GOREAUD, F. (2000). « Apports de l'analyse de la structure spatiale en forêt tempérée à l'étude de la modélisation des peuplements complexes ». Thèse de doct. Nancy : ENGREF (cf. p. 14, 20, 22).
- GÖTZENBERGER, L., F. de BELLO, K. A. BRÅTHEN, J. DAVISON, A. DUBUIS, A. GUIAN, J. LEPSŠ, R. LINDBORG, M. MOORA, M. PÄRTEL, L. PELLISSIER, J. POTTIER, P. VITTOZ, K. ZOBEL et M. ZOBEL (2012). « Ecological assembly rules in plant communities - approaches, patterns and prospects ». In : *Biological Reviews* 87.1, p. 111–127 (cf. p. 82).
- GOURLET-FLEURY, S., J. M. GUEHL et O. LAROUSSINIE (2004). *Ecology & management of a neotropical rainforest. Lessons drawn from Paracou, a long-term experimental research site in French Guiana*. Paris (cf. p. 25, 83).
- GOURLET-FLEURY, S., G. COMU, S. JESEL, H. DESSARD, J.-G. JOURGET, L. BLANC et N. PICARD (2005). « Using models to predict recovery and assess tree species vulnerability in logged tropical forests : A case study from French Guiana ». In : *Forest Ecology and Management* 209.1-2, p. 69–86 (cf. p. 83).
- 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. 66).
- (1971). « A General Coefficient of Similarity and Some of Its Properties ». In : *Biometrics* 27.4, p. 857–871 (cf. p. 46).
- GOWER, J. C. et P. LEGENDRE (1986). « Metric and Euclidean properties of dissimilarity coefficients ». In : *Journal of classification* 48, p. 5–48 (cf. p. 67).
- GRANGER, V., N. BEZ, J.-M. FROMENTIN, C. MEYNARD, A. JADAUD et B. MÉRIGOT (2015). « Mapping diversity indices : not a trivial issue ». In : *Methods in Ecology and Evolution* 6.6, p. 688–696 (cf. p. 81).
- GRASSBERGER, P. (1988). « Finite sample corrections to entropy and dimension estimates ». In : *Physics Letters A* 128.6-7, p. 369–373 (cf. p. 76).
- GREGORIUS, H.-R. (1991). « On the concept of effective number ». In : *Theoretical population biology* 40.2, p. 269–83 (cf. p. 37).
- (2010). « Linking Diversity and Differentiation ». In : *Diversity* 2.3, p. 370–394 (cf. p. 40, 60, 83).
- (2014). « Partitioning of diversity : the "within communities" component ». In : *Web Ecology* 14, p. 51–60 (cf. p. 37, 66).
- 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. 67).
- GUITET, S., D. SABATIER, O. BRUNAUX, B. HÉRAULT, M. AUBRY-KIENTZ, J.-F. MOLINO et C. BARALOTO (2014). « Estimating tropical tree diversity indices from forestry surveys : A method to integrate taxonomic uncertainty ». In : *Forest Ecology and Management* 328, p. 270–281 (cf. p. 84).
- 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. 1092–1101 (cf. p. 79).
- HALPERN, C. B. et T. A. SPIES (1995). « Plant species diversity in natural and managed forests of the Pacific Northwest ». In : *Ecological Applications* 5.4, p. 913–934 (cf. p. 83).
- HARDY, G. H., J. E. LITTLEWOOD et G. PÓLYA (1952). *Inequalities*. Cambridge University Press (cf. p. 47).
- 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. 67).
- 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. 67).
- HAVRDA, J. et F. CHARVÁT (1967). « Quantification method of classification processes. Concept of structural α -entropy ». In : *Kybernetika* 3.1, p. 30–35 (cf. p. 38).
- HEINRICH, L. (1991). « Goodness-of-fit tests for the second moment function of a stationary multidimensional poisson process ». In : *Statistics : A Journal of Theoretical and Applied Statistics* 22.2, p. 245–268 (cf. p. 27, 33).
- HILL, M. O. (1973). « Diversity and Evenness : A Unifying Notation and Its Consequences ». In : *Ecology* 54.2, p. 427–432 (cf. p. 37, 41).
- HOFFMANN, S. et A. HOFFMANN (2008). « Is there a "true" diversity ? » In : *Ecological Economics* 65.2, p. 213–215 (cf. p. 40).
- 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. 75).
- HOUBEINE, M. (1999). « Concentration Géographique des Activités et Spécialisation des Départements Français ». In : *Economie et Statistique* 326-327.6-7, p. 189–204 (cf. p. 82).
- HUBBELL, S. P. (1999). « Light-Gap Disturbances, Recruitment Limitation, and Tree Diversity in a Neotropical Forest ». In : *Science* 283.5401, p. 554–557 (cf. p. 68).
- HUBBELL, S. P., R. CONDIT et R. B. FOSTER (2005). *Barro Colorado Forest Census Plot Data* (cf. p. 68).
- HURLBERT, S. H. (1971). « The Nonconcept of Species Diversity : A Critique and Alternative Parameters ». In : *Ecology* 52.4, p. 577–586 (cf. p. 37).
- ILLIAN, J., A. PENTTINEN, H. STOYAN et D. STOYAN (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Chichester : Wiley-Interscience, p. 1–534 (cf. p. 25, 27, 30).
- 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. 62, 68).
- JOST, L. (2006). « Entropy and diversity ». In : *Oikos* 113.2, p. 363–375 (cf. p. 15, 39, 40, 58).
- (2007). « Partitioning diversity into independent alpha and beta components ». In : *Ecology* 88.10, p. 2427–2439 (cf. p. 15, 39, 40, 58–60, 63, 64, 66, 68).
- (2008). « GST and its relatives do not measure differentiation ». In : *Molecular Ecology* 17.18, p. 4015–4026 (cf. p. 66).
- (2009). « Mismeasuring biological diversity : Response to Hoffmann and Hoffmann (2008) ». In : *Ecological Economics* 68, p. 925–928 (cf. p. 40).
- (2010). « Independence of alpha and beta diversities ». In : *Ecology* 91.7, p. 1969–1994 (cf. p. 58).
- 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. 59, 73).
- 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. 39).
- 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. 41).
- KOEN, C. (1991). « Approximate confidence bounds for Ripley's statistic for random points in a square ». In : *Biometrical Journal* 33, p. 173–177 (cf. p. 27).
- KRUGMAN, P. (1991). *Geography and Trade*. London : MIT Press (cf. p. 82).
- LANDE, R. (1996). « Statistics and partitioning of species diversity, and similarity among multiple communities ». In : *Oikos* 76, p. 5–13 (cf. p. 59, 63).
- LANG, G. et É. MARCON (2013). « Testing randomness of spatial point patterns with the Ripley statistic ». In : *ESAIM : Probability and Statistics* 17, p. 767–788. arXiv : 1006.1567 (cf. p. 15, 24, 27).
- LANG, G., É. MARCON et F. PUECH (2015). « Distance-Based Measures of Spatial Concentration : Introducing a Relative Density Function ». In : *HAL* 01082178.version 3, p. 1–14 (cf. p. 31).
- LAW, R., J. ILLIAN, D. F. R. P. BURSLEM, G. GRATZER, C. V. S. GUNATILLEKE et I. A. U. N. GUNATILLEKE (2009). « Ecological information from spatial patterns of plants : insights from point process theory ». English. In : *Journal of Ecology* 97.4, p. 616–628 (cf. p. 14, 20).
- LEINSTER, T. (2013). « The Magnitude of Metric Spaces ». In : *Documenta Mathematica* 18, p. 857–905 (cf. p. 50).
- LEINSTER, T. et C. COBBOLD (2012). « Measuring diversity : the importance of species similarity ». In : *Ecology* 93.3, p. 477–489 (cf. p. 15, 46, 47).
- LEPŠ, J., F. DE BELLO, S. LAVOREL et S. BERMAN (2006). « Quantifying and interpreting functional diversity of natural communities : practical considerations matter ». In : *Preslia* 78, p. 481–501 (cf. p. 52).
- LETCHER, S. G., R. L. CHAZDON, A. C. S. ANDRADE, F. BONGERS, M. van BREUGEL, B. FINEGAN, S. G. LAURANCE, R. C. G. MESQUITA, M. MARTÍNEZ-RAMOS et G. B. WILLIAMSON (2012). « Phylogenetic community structure during succession : Evidence from three Neotropical forest sites ». In : *Perspectives in Plant Ecology, Evolution and Systematics* 14.2, p. 79–87 (cf. p. 84).
- LEWONTIN, R. (1972). « The apportionment of human diversity ». In : *Evolutionary biology* 6, p. 381–398 (cf. p. 61).
- LIN, J. (1991). « Divergence Measures Based on the Shannon Entropy ». In : *IEEE Transactions on Information Theory* 37.1, p. 145–151 (cf. p. 61).
- 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. 41).
- LOOSMORE, N. B. et E. D. FORD (2006). « Statistical inference using the G or K point pattern spatial statistics ». In : *Ecology* 87.8, p. 1925–1931 (cf. p. 27).
- 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. 61).
- MAASOUMI, E. (1993). « A compendium to information theory in economics and econometrics ». In : *Econometric Reviews* 12.2, p. 137–181 (cf. p. 35).
- MACARTHUR, R. H. (1955). « Fluctuations of Animal Populations and a Measure of Community Stability ». In : *Ecology* 36.3, p. 533–536 (cf. p. 36).
- (1965). « Patterns of species diversity ». In : *Biological Reviews* 40.4, p. 510–533 (cf. p. 37).
- MARCON, É. (1999). « Forest surveys on a tree by tree basis : A theoretical and practical approach ». In : *Revue Forestière Française* 51.1, p. 57–69 (cf. p. 13).
- (2010). « Statistiques spatiales avec applications à l'écologie et à l'économie ». Thèse de doct. AgroParisTech (cf. p. 22).
- (2015). « Practical Estimation of Diversity from Abundance Data ». In : *HAL* 01212435.version 1, p. 1–27 (cf. p. 77).
- MARCON, É. et B. HÉRAULT (2015a). « Decomposing Phylo-diversity ». In : *Methods in Ecology and Evolution* 6.3, p. 333–339 (cf. p. 15, 70, 80, 82).

- (2015b). « entropart, an R Package to Measure and Partition Diversity ». In : *Journal of Statistical Software* 67.8, p. 1–26 (cf. p. 16, 80).
- MARCON, É. et F. PUECH (2003). « Evaluating the geographic concentration of industries using distance-based methods ». In : *Journal of Economic Geography* 3.4, p. 409–428 (cf. p. 13).
- (2010). « Measures of the Geographic Concentration of Industries : Improving Distance-Based Methods ». In : *Journal of Economic Geography* 10.5, p. 745–762 (cf. p. 14, 31, 32).
- (2015a). « A Typology of Distance-Based Measures of Spatial Concentration ». In : *HAL SHS* 00679993.version 4, p. 1–16 (cf. p. 14, 26, 33).
- (2015b). « Mesures de la concentration spatiale en espace continu : théorie et applications ». In : *Economie et Statistique* 474, p. 105–131 (cf. p. 14, 34).
- MARCON, É., F. PUECH et S. TRAISSAC (2012a). « Characterizing the relative spatial structure of point patterns ». In : *International Journal of Ecology* 2012.Article ID 619281, p. 11 (cf. p. 14, 32).
- MARCON, É., B. HÉRAULT, C. BARALOTO et G. LANG (2012b). « The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity ». In : *Oikos* 121.4, p. 516–522 (cf. p. 15, 43, 59, 61, 62, 67, 80).
- MARCON, É., S. TRAISSAC et G. LANG (2013). « A Statistical Test for Ripley's Function Rejection of Poisson Null Hypothesis ». In : *ISRN Ecology* 2013.Article ID 753475, p. 9 (cf. p. 15, 27, 28).
- 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. 15, 39, 40, 45, 59, 61, 63, 65, 68, 75, 76, 80, 82).
- MARCON, É., Z. ZHANG et B. HÉRAULT (2014b). « The Decomposition of Similarity-Based Diversity and its Bias Correction ». In : *HAL* 00989454.version 3, p. 1–12 (cf. p. 15, 51, 68, 71, 75, 80).
- MARCON, É., S. TRAISSAC, F. PUECH et G. LANG (2015). « Tools to Characterize Point Patterns : dbmss for R ». In : *Journal of Statistical Software* 67.3, p. 1–15 (cf. p. 15, 28).
- MARSHALL, A. (1890). *Principle of Economics*. London : Macmillan (cf. p. 82).
- MATÉRN, B. (1960). « Spatial variation ». In : *Meddelanden från Statens Skogsforskningsinstitut* 49.5, p. 1–144 (cf. p. 22).
- MAY, R. M. (1990). « Taxonomy as Destiny ». In : *Nature* 347, p. 129–130 (cf. p. 43).
- 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. 38).
- MØLLER, J. et R. P. WAAGEPETERSEN (2004). *Statistical Inference and Simulation for Spatial Point Processes*. T. 100. Chapman et Hall, p. 1–300 (cf. p. 17, 22).
- 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. 70).
- MOUILLOT, D. et A. LEPRÊTRE (1999). « A comparison of species diversity estimators ». In : *Researches on Population Ecology* 41.2, p. 203–215 (cf. p. 75).
- NEI, M. (1972). « Genetic Distance between Populations ». In : *The American Naturalist* 106.949, p. 283–292 (cf. p. 49).
- (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. 65, 73).
- OLLIVIER, M., C. BARALOTO et É. MARCON (2007). « A trait database for Guianan rain forest trees permits intra- and inter-specific contrasts ». In : *Annals of Forest Science* 64.7, p. 781–786 (cf. p. 4, 14).
- OPENSHAW, S. et P. J. TAYLOR (1979). « A million or so correlation coefficients : three experiments on the modifiable areal unit problem ». In : *Statistical Applications in the Spatial Sciences*. Sous la dir. de N. WRIGLEY. London : Pion, p. 127–144 (cf. p. 83).
- OSAZUWA-PETERS, O. L., C. A. CHAPMAN et A. E. ZANNE (2015). « Selective logging : does the imprint remain on tree structure and composition after 45 years? ». In : *Conservation Physiology* 3.1, cov012 (cf. p. 84).
- 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. 15, 36, 37, 41, 80).
- PAVOINE, S. et M. B. BONSALE (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. 44, 46).
- (2011). « Measuring biodiversity to explain community assembly : a unified approach ». In : *Biological Reviews* 86.4, p. 792–812 (cf. p. 44).
- 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. 52, 55).
- 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. 54).
- 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. 66).
- PAVOINE, S., S. OLLIER et A.-B. DUFOUR (2005). « Is the originality of a species measurable? ». In : *Ecology Letters* 8, p. 579–586 (cf. p. 47, 53).
- PAVOINE, S., E. VELA, S. GACHET, G. de BELAIR et M. B. BONSALE (2011). « Linking patterns in phylogeny, traits, abiotic variables and space : a novel approach to linking environmental filtering and plant community assembly ». English. In : *Journal of Ecology* 99.1, p. 165–175 (cf. p. 46).
- PÉLISSIER, R. et F. GOREAUD (2001). « A practical approach to the study of spatial structure in simple cases of heterogeneous vegetation ». In : *Journal of Vegetation Science* 12.1, p. 99–108 (cf. p. 14).
- PETCHY, O. L. et K. J. GASTON (2002). « Functional diversity (FD), species richness and community composition ». In : *Ecology Letters* 5, p. 402–411 (cf. p. 44).
- PIELOU, E. C. (1975). *Ecological Diversity*. New York : Wiley (cf. p. 43).
- PLOTKIN, J. B., M. D. POTTS, N. LESLIE, N. MANOKARAN, J. V. LAFRANKIE et P. S. ASHTON (2000). « Species-area curves, spatial aggregation, and habitat specialization in tropical forests ». In : *Journal of Theoretical Biology* 207.1, p. 81–99 (cf. p. 82).
- PODANI, J. (1999). « Extending Gower's General Coefficient of Similarity to Ordinal Characters ». In : *Taxon* 48.2, p. 331–340 (cf. p. 46).
- PODANI, J. et D. SCHMERA (2006). « On dendrogram-based measures of functional diversity ». In : *Oikos* 115.1, p. 179–185 (cf. p. 47, 53).
- (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. 47, 53).

- PURVIS, A. et A. HECTOR (2000). « Getting the measure of biodiversity. » In : *Nature* 405.6783, p. 212–9 (cf. p. 35).
- R CORE TEAM (2015). *R : A Language and Environment for Statistical Computing*. Vienna, Austria : R Foundation for Statistical Computing (cf. p. 15).
- RAO, C. R. (1982). « Diversity and dissimilarity coefficients : a unified approach ». In : *Theoretical Population Biology* 21, p. 24–43 (cf. p. 43, 67).
- 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. 61).
- 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. 36).
- RICHARD-HANSEN, C., G. JAUEN, 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. 81).
- RICOTTA, C. (2005a). « On hierarchical diversity decomposition ». In : *Journal of Vegetation Science* 16.2, p. 223–226 (cf. p. 59).
- (2005b). « On parametric diversity indices in ecology : A historical note ». In : *Community Ecology* 6.2, p. 241–244 (cf. p. 44).
- (2007). « A semantic taxonomy for diversity measures ». In : *Acta Biotheoretica* 55.1, p. 23–33 (cf. p. 44).
- RICOTTA, C. et G. AVENA (2003). « An information-theoretical measure of beta-diversity ». In : *Plant Biosystems* 137.1, p. 57–61 (cf. p. 61).
- 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. 48, 71).
- (2009). « Diversity partitioning of Rao's quadratic entropy ». In : *Theoretical Population Biology* 76.4, p. 299–302 (cf. p. 45, 67).
- 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. 67).
- RIPLEY, B. D. (1976). « The Foundations of Stochastic Geometry ». In : *Annals of Probability* 4.6, p. 995–998 (cf. p. 14, 24).
- (1977). « Modelling Spatial Patterns ». In : *Journal of the Royal Statistical Society B* 39.2, p. 172–212 (cf. p. 14, 20, 24).
- (1979). « Tests of 'randomness' for spatial point patterns ». In : *Journal of the Royal Statistical Society B* 41.3, p. 368–374 (cf. p. 27).
- (1981). *Spatial statistics*. New York : John Wiley & Sons, p. 1–255 (cf. p. 27).
- ROUTLEDGE, R. D. (1979). « Diversity indices : Which ones are admissible? » In : *Journal of Theoretical Biology* 76.4, p. 503–515 (cf. p. 60).
- SCHNITZER, S. A. et W. P. CARSON (2001). « Treefall Gaps and the Maintenance of Species Diversity in a Tropical Forest ». In : *Ecology* 82.4, p. 913–919 (cf. p. 84).
- SHANNON, C. E. (1948). « A Mathematical Theory of Communication ». In : *The Bell System Technical Journal* 27, p. 379–423, 623–656 (cf. p. 15, 35).
- SHANNON, C. E. et W. WEAVER (1963). *The Mathematical Theory of Communication*. University of Illinois Press (cf. p. 35).
- SHEN, G., F. HE, R. WAAGEPETERSEN, I. F. SUN, Z. HAO, Z.-S. CHEN et M. YU (2013a). « Quantifying effects of habitat heterogeneity and other clustering processes on spatial distributions of tree species ». In : *Ecology* 94.11, p. 2436–2443 (cf. p. 82).
- SHEN, G., T. WIEGAND, X. MI et F. HE (2013b). « Quantifying spatial phylogenetic structures of fully stem-mapped plant communities ». In : *Methods in Ecology and Evolution* 4.12, p. 1132–1141 (cf. p. 72).
- SHIMATANI, K. (2001). « Multivariate point processes and spatial variation of species diversity ». In : *Forest Ecology and Management* 142.1-3, p. 215–229 (cf. p. 44, 82).
- 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. 47).
- STODDART, J. A. (1983). « A genotypic diversity measure ». In : *Journal of Heredity* 74, p. 489–490 (cf. p. 37).
- STOYAN, D., W. S. KENDALL et J. MECKE (1987). *Stochastic Geometry and its Applications*. New York : John Wiley & Sons, 345 p. (Cf. p. 21, 27).
- THEIL, H. (1967). *Economics and Information Theory*. Chicago : Rand McNally et Company (cf. p. 15, 35).
- THOMAS, M. (1949). « A Generalization of Poisson's Binomial Limit for Use in Ecology ». In : *Biometrika* 36.1/2, p. 18–25 (cf. p. 22).
- 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. 43).
- TOMPPA, E. (1986). *Models and methods for analysing spatial patterns of trees*. T. 138. Helsinki, Finland : The Finnish forest research institute, p. 1–65 (cf. p. 22).
- TOTHMERESZ, B. (1995). « Comparison of different methods for diversity ordering ». In : *Journal of Vegetation Science* 6.2, p. 283–290 (cf. p. 41).
- TSALLIS, C. (1988). « Possible generalization of Boltzmann-Gibbs statistics ». In : *Journal of Statistical Physics* 52.1, p. 479–487 (cf. p. 38).
- (1994). « What are the numbers that experiments provide? » In : *Química Nova* 17.6, p. 468–471 (cf. p. 39).
- TSALLIS, C. et E. BRIGATTI (2004). « Nonextensive statistical mechanics : A brief introduction ». English. In : *Continuum Mechanics and Thermodynamics* 16.3, p. 223–235 (cf. p. 15).
- 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. 64).
- 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. 58).
- (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. 58).
- (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. 58).
- ULANOWICZ, R. E. (2001). « Information theory in ecology ». In : *Computers & Chemistry* 25.4, p. 393–399 (cf. p. 36).
- ULANOWICZ, R. E., R. D. HOLT et M. BARFIELD (2014). « Limits on ecosystem trophic complexity : Insights from ecological network analysis ». In : *Ecology Letters* 17.2, p. 127–136 (cf. p. 81).

- 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. 43).
- 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. 58, 59).
- 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. 59).
- 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. 44).
- 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. 67).
- WAGNER, H. H. (2003). « Spatial covariance in plant communities : Integrating ordination, geostatistics, and variance testing ». In : *Ecology* 84.4, p. 1045–1057 (cf. p. 73).
- WARD, J. S. et F. J. FERRANDINO (1999). « New derivation reduces bias and increases power of Ripley's L index ». In : *Ecological Modelling* 116.2-3, p. 225–236 (cf. p. 27).
- 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. 43).
- WEBB, C. O., J. B. LOSOS et A. A. AGRAWAL (2006). « Integrating Phylogenies into Community Ecology ». In : *Ecology* 87.sp7, S1–S2 (cf. p. 43).
- WEBER, A. (1909). *Über den Standort der Industrien*. Tübingen. English translation edited in 1971, "Theory of the location of industries", Russell & Russell. (cf. p. 82).
- WHITTAKER, R. H. (1960). « Vegetation of the Siskiyou Mountains, Oregon and California ». In : *Ecological Monographs* 30.3, p. 279–338 (cf. p. 57, 59).
- (1972). « Evolution and Measurement of Species Diversity ». In : *Taxon* 21.2/3, p. 213–251 (cf. p. 59).
- 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. 60).
- 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. 46).
- WRIGHT, S. (1931). « Evolution in Mendelian Populations ». In : *Genetics* 16.2, p. 97–159 (cf. p. 37).
- ZHANG, Z. (2013). « Asymptotic normality of an entropy estimator with exponentially decaying bias ». In : *IEEE Transactions on Information Theory* 59.1, p. 504–508 (cf. p. 76).
- ZHANG, Z. et M. GRABCHAK (2013). « Bias adjustment for a nonparametric entropy estimator ». In : *Entropy* 15.6, p. 1999–2011 (cf. p. 76).
- (2014). « Entropic Representation and Estimation of Diversity Indices ». In : *arXiv* 1403.3031.v. 2, p. 1–12 (cf. p. 76).
- 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. 76).

Troisième partie

Annexes : Publications

APPENDIX A

entropart, an R Package to Measure and Partition Diversity

Marcon, E. (2015). « entropart, an R Package to Measure and Partition Diversity ». In : Journal of Statistical Software 67.8, p. 1–26



entropart: An R Package to Measure and Partition Diversity

Eric Marcon
AgroParisTech
UMR EcoFoG

Bruno Hérault
Cirad
UMR EcoFoG

Abstract

entropart is a package for R designed to estimate diversity based on HCDDT entropy or similarity-based entropy. It allows calculating species-neutral, phylogenetic and functional entropy and diversity, partitioning them and correcting them for estimation bias.

Keywords: biodiversity, entropy, partitioning.

1. Introduction

Diversity measurement can be done through a quite rigorous framework based on entropy, i.e., the amount of uncertainty calculated from the frequency distribution of a community (Patil and Taillie 1982; Jost 2006; Marcon, Scotti, Hérault, Rossi, and Lang 2014a). Tsallis entropy, also known as HCDDT entropy (Havrdá and Charvát 1967; Daróczy 1970; Tsallis 1988), is of particular interest (Jost 2006; Marcon *et al.* 2014a) namely because it gathers the number of species and Shannon (1948a,b) and Simpson (1949) indices of diversity into a single framework. Interpretation of entropy is not straightforward but one can easily transform it into Hill numbers (Hill 1973) which have many desirable properties (Jost 2007): mainly, they are the number of equally-frequent species that would give the same level of diversity as the data.

Marcon and Hérault (2015a) generalized the duality of entropy and diversity, deriving the relation between phylogenetic or functional diversity (Chao, Chiu, and Jost 2010) and phylogenetic or functional entropy (we will write *phylodiversity* and *phyloentropy* for short), as introduced by Pavoine, Love, and Bonsall (2009). Special cases are the well-known indices PD for phylogenetic diversity (Faith 1992) and FD for functional diversity (Petchey and Gas-

ton 2002) and Rao’s (1982) quadratic entropy. The same relation holds between Ricotta and Szeidl entropy of a community (Ricotta and Szeidl 2006) and similarity-based diversity (Leinster and Cobbold 2012).

The **entropart** package (Marcon and Hérault 2015b) for R (R Core Team 2015) enables calculation of all these measures of diversity and entropy and their partitioning and is available from the Comprehensive R Archive Network (CRAN) at <http://CRAN.R-project.org/package=entropart>.

Diversity partitioning means that, in a given area, the γ diversity D_γ of all individuals found may be split into within (α diversity, D_α) and between (β diversity, D_β) local assemblages. α diversity reflects the diversity of individuals *in* local assemblages whereas β diversity reflects the diversity *of* the local assemblages. Marcon *et al.* (2014a) derived the decomposition of Tsallis γ entropy into its α and β components, generalized to phylodiversity (Marcon and Hérault 2015a) and similarity-based diversity (Marcon, Zhang, and Hérault 2014b).

Estimators of diversity are biased because of unseen species and also because they are not linear functions of probabilities (Marcon *et al.* 2014a). α and γ diversities are underestimated by naive estimators (Chao and Shen 2003; Dauby and Hardy 2012). β diversity is severely biased too when sampling is not sufficient (Beck, Holloway, and Schwanghart 2013). Bias-corrected estimators of phylodiversity have been developed by Marcon and Hérault (2015a). Estimators of similarity-based diversity were derived by Marcon *et al.* (2014b). The package includes them all.

In summary, the framework supported by the package is as follows. First, an information function is chosen to describe the amount of surprise brought by the observation of each individual. In the simplest case of species-neutral diversity, it is just a decreasing function of probability: Observing an individual of a rarer species brings more surprise. Various information functions allow evaluating species-neutral, phylogenetic or functional entropy. Surprise is averaged among all individuals of a community to obtain its entropy. Entropy is systematically transformed into diversity for interpretation. Diversity is an effective number of species, i.e., the number of equally-different and equally-frequent species that would give the same entropy as the data. The average entropy of communities of an assemblage is α entropy, while the entropy of the assemblage is γ entropy. Their difference is β entropy. After transformation, β diversity is the ratio of γ to α diversity. It is an effective number of communities, i.e., the number of equally-weighted communities with no species in common necessary to obtain the same diversity as the data. correction of estimation bias is more easily applied to entropy before transforming it into diversity.

This framework is somehow different from that of Chao, Chiu, and Jost (2014) who define α diversity in another way (see Marcon and Hérault 2015a, for a detailed comparison), such that α entropy is not the average surprise of an assemblage. They also propose a definition of functional diversity (Chiu and Chao 2014) based on the information brought by pairs of individuals that is not supported in the package.

The subsequent sections of this paper present the package features, illustrated by worked examples based on the data included in the package.

2. Package organization

2.1. Data

Most functions of the package calculate entropy or diversity of a community or of an assemblage of communities called a “meta-community”. Community functions accept a vector of probabilities or of abundances for species data. Each element of the vector contains the probability or the number of occurrences of a species in a given community. Meta-community functions require a particular data organization in a ‘MetaCommunity’ object described here.

A ‘MetaCommunity’ object is basically a list. Its main components are `Nsi`, a matrix containing the species abundances whose rows are species, columns are communities and `Wi`, a vector containing community weights. Creating a ‘MetaCommunity’ object is the purpose of the `MetaCommunity()` function. Arguments are a dataframe containing the number of individuals per species (rows) in each community (columns), and a vector containing the community weights. The following example creates a ‘MetaCommunity’ object consisting of three communities of unequal weights with 4 species. The weighted average probabilities of occurrence of species and the total number of individuals define the meta-community as the assemblage of communities.

```
R> library("entropart")
R> (df <- data.frame(C1 = c(10, 10, 10, 10), C2 = c(0, 20, 35, 5),
+   C3 = c(25, 15, 0, 2), row.names = c("sp1", "sp2", "sp3", "sp4")))

      C1 C2 C3
sp1  10  0 25
sp2  10 20 15
sp3  10 35  0
sp4  10  5  2

R> w <- c(1, 2, 1)
R> MC <- MetaCommunity(Abundances = df, Weights = w)
```

A meta-community is partitioned into several local communities (indexed by $i = 1, 2, \dots, I$). n_i individuals are sampled in community i . Let $s = 1, 2, \dots, S$ denote the species that compose the meta-community, $n_{s,i}$ the number of individuals of species s sampled in the local community i , $n_s = \sum_i n_{s,i}$ the total number of individuals of species s , $n = \sum_s \sum_i n_{s,i}$ the total number of sampled individuals. Within each community i , the probability $p_{s,i}$ for an individual to belong to species s is estimated by $\hat{p}_{s,i} = n_{s,i}/n_i$. The same probability for the meta-community is p_s . Communities have a weight w_i , satisfying $p_s = \sum_i w_i p_{s,i}$. The commonly used $w_i = n_i/n$ is a possible weight, but the weighting may be arbitrary (e.g., depending on the sampled areas). The component `Ps` of a ‘MetaCommunity’ object contains the probability of occurrence of each species in the meta-community, calculated in this way:

```
R> MC$Ps

      sp1      sp2      sp3      sp4
0.2113095 0.3184524 0.3541667 0.1160714
```

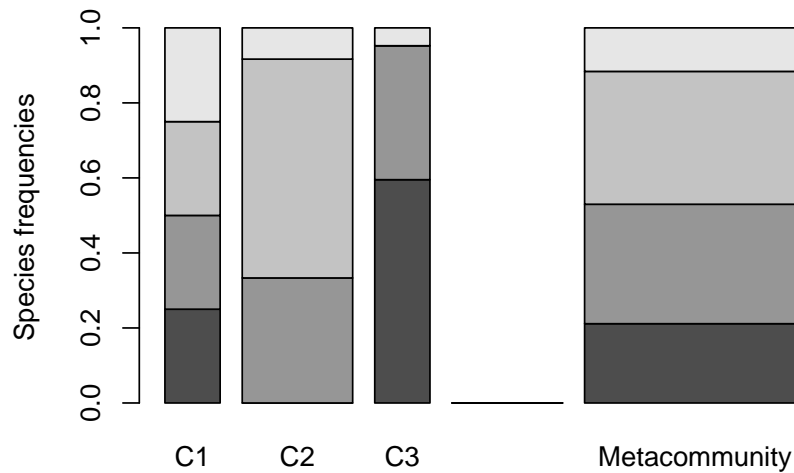


Figure 1: Plot of a ‘MetaCommunity’ object. Communities (named C1, C2 ad C3) are represented in the left part of the figure, the meta-community to the right. Bar widths are proportional to community weights. Species abundances are represented vertically: 4 species are present in the meta-community, only 3 of them in communities C2 and C3.

A ‘MetaCommunity’ object can be summarized and plotted (Figure 1).

The package contains an example dataset containing the inventory of two 1-ha tropical forest plots in Paracou, French Guiana (Marcon, Hérault, Baraloto, and Lang 2012):

```
R> data("Paracou618", package = "entropart")
R> 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
```

Paracou618.MC is a meta-community made of two communities named “P006” and “P018”, containing 425 species (their name is *Family_Genus_Species*, abbreviated to 4 characters). The values of the abundance matrix are the number of individuals of each species in each community. Sample coverage will be explained later.

The dataset also contains a taxonomy and a functional tree. `Paracou618.Taxonomy` is an object of class ‘phylog’, defined in package `ade4` (Dray and Dufour 2007), namely a phylogenetic tree. This example data is only a taxonomy, containing family, genus and species levels

for the sake of simplicity. `Paracou618.Functional` is an object of class ‘`hclust`’ containing a functional tree based on leaf, height, stem and seed functional traits (Hérault and Honnay 2007; Marcon and Hérault 2015a). The package accepts any ultrametric tree of class ‘`phylog`’ or ‘`hclust`’. `Paracou618.dist` is the distance matrix (actually a ‘`dist`’ object) used to build the functional tree.

2.2. Utilities

The deformed logarithm formalism (Tsallis 1994) is very convenient to manipulate entropies. The deformed logarithm of order q is defined as:

$$\ln_q x = \frac{x^{1-q} - 1}{1 - q}. \quad (1)$$

It converges to \ln when $q \rightarrow 1$.

The inverse function of $\ln_q x$ is the deformed exponential:

$$e_q^x = [1 + (1 - q)x]^{\frac{1}{1-q}}. \quad (2)$$

The corresponding functions in the package are `lnq(x, q)` and `expq(x, q)`.

3. Species-neutral diversity

3.1. Community functions

HCDT entropy

Species-neutral HCDT entropy of order q of a community is defined as:

$${}^qH = \frac{1 - \sum_s p_s^q}{q - 1} = - \sum_s p_s^q \ln_q p_s. \quad (3)$$

q is the order of diversity (e.g., 1 for Shannon). Entropy can be calculated by the `Tsallis` function. Paracou meta-community entropy of order 1 is:

```
R> Tsallis(Ps = Paracou618.MC$Ps, q = 1)
```

```
[1] 4.736023
```

For convenience, special cases of entropy of order q have their own functions with clear names: `Richness` for $q = 0$, `Shannon` for $q = 1$, `Simpson` for $q = 2$.

```
R> Shannon(Ps = Paracou618.MC$Ps)
```

```
[1] 4.736023
```


Entropy values have no intuitive interpretation in general, except for the number of species 0H and Simpson entropy 2H which is the probability for two randomly chosen individuals to belong to different species.

Sample coverage

A useful indicator of sampling quality is the sample coverage (Good 1953; Chao, Lee, and Chen 1988; Zhang and Huang 2007), that is the probability for a species of the community to be observed in the actual sample. It equals the sum of the probability of occurrences of all observed species. Its historical estimator is (Good 1953):

$$\hat{C} = 1 - \frac{S^1}{n}. \quad (4)$$

S^1 is the number of singletons (species observed once) of the sample, and n is its size. The estimator has been improved by taking into account the whole distribution of species (Zhang and Huang 2007). The `Coverage` function calculates it, allowing to choose the estimator, using Zhang and Huang's method by default:

```
R> Coverage(Ns = Paracou618.MC$Ns)
```

```
[1] 0.9220438
```

The sample coverage cannot be estimated from probability data; abundances are required.

Its interpretation is straightforward: Some species have not been sampled. Their number is unknown but their total probability of occurrence can be estimated accurately. Here, it is a bit less than 8%. From another point of view, the probability for an individual of the community to belong to a sampled species is C : 8% of them belong to missed species. If the number of missed species are of interest, they can be estimated using other software packages (e.g., the R package **SPECIES**, Wang 2011), but we will not discuss this in detail here. The sample coverage is the foundation of many estimators of entropy.

Bias-corrected estimators

Correction of estimation bias is used to improve the estimation of entropy despite unobserved species and also mathematical issues (Bonachela, Hinrichsen, and Muñoz 2008). Bias-corrected estimators (often relying on sample coverage) are returned by functions whose names are prefixed by `bc`, such as `bcTsallis`. They are similar to the non-corrected ones but they use abundance data and propose several bias correction techniques which can be selected by the `Correction` argument. A “Best” correction is calculated by default as detailed in the help file of each function.

```
R> bcTsallis(Ns = Paracou618.MC$Ns, q = 1)
```

```
[1] 4.898061
```

The best correction for Tsallis entropy follows Marcon *et al.* (2014a). `bcSimpson` returns Lande's correction (Lande 1996) and `bcShannon` returns the very efficient correction by Chao, Wang, and Jost (2013), so that their results are different (and more accurate) than those of the general `bcTsallis` function.

```
R> bcShannon(Ns = Paracou618.MC$Ns)
```

```
[1] 4.892159
```

Bias-corrected entropy is ready to be transformed into explicit diversity.

Effective numbers of species

Entropy should be converted into “true diversity” (Jost 2007), i.e., effective number of species equal to Hill (1973) numbers:

$${}^qD = \left(\sum_s p_s^q \right)^{\frac{1}{1-q}}. \quad (5)$$

This can be done by the deformed exponential function, or using directly the `Diversity` or `bcDiversity` functions (equal to the deformed exponential of order q of Tsallis or `bcTsallis`)

```
R> expq(Simpson(Ps = Paracou618.MC$Ps), q = 2)
```

```
[1] 68.7215
```

```
R> Diversity(Ps = Paracou618.MC$Ps, q = 2)
```

```
[1] 68.7215
```

```
R> expq(bcTsallis(Ns = Paracou618.MC$Ns, q = 2), q = 2)
```

```
[1] 73.19676
```

```
R> bcDiversity(Ns = Paracou618.MC$Ns, q = 2)
```

```
[1] 73.19676
```

The effective number of species of the Paracou dataset is estimated to be 73 after bias correction (rather than 69 without it). It means that a community made of 73 equally-frequent species has the same Simpson entropy as the actual one. This is much less than the actual 425 sampled species but Simpson’s entropy focuses on dominant species.

3.2. Meta-community functions

Meta-community functions allow partitioning diversity according to Patil and Taillie’s concept of diversity of a mixture (Patil and Taillie 1982), i.e., α entropy of a meta-community is defined as the weighted average of community entropy, following Routledge (1979):

$${}^qH_\alpha = \sum_i w_i {}^qH_\alpha. \quad (6)$$

${}^q_i H_\alpha$ is the entropy of community i :

$${}^q_i H_\alpha = \frac{1 - \sum_s p_{s,i}^q}{q - 1} = - \sum_s p_{s,i}^q \ln_q p_{s,i}. \quad (7)$$

Jost's (2007) definition of α entropy is not supported explicitly in the package since it only allows partitioning of equally weighted communities. In this particular case, both definitions are identical.

γ entropy of the meta-community is defined as α entropy of a community. β entropy, the difference between γ and α , is the generalized Jensen-Shannon divergence between the species distribution of the meta-community and those of communities (Marcon *et al.* 2014a):

$${}^q H_\beta = {}^q H_\gamma - {}^q H_\alpha = \sum_s p_{s,i}^q \ln_q \frac{p_{s,i}}{p_s}. \quad (8)$$

β entropy should be transformed into diversity, i.e., an effective number of communities:

$${}^q D_\beta = e_q^{\frac{{}^q H_\beta}{1 - (q-1){}^q H_\alpha}}. \quad (9)$$

Basic meta-community functions

These values can be estimated by the meta-community functions named `AlphaEntropy`, `AlphaDiversity`, `BetaEntropy`, `BetaDiversity`. They accept a 'MetaCommunity' object and an order of diversity q as arguments, and return an 'MCEntropy' or 'MCDiversity' object which can be summarized and plotted. `GammaEntropy` and `GammaDiversity` return a number. Corrections of estimation bias are applied by default:

```
R> e <- AlphaEntropy(Paracou618.MC, q = 1)
R> summary(e)
```

```
Neutral alpha entropy of order 1 of metaCommunity Paracou618.MC
with correction: Best
```

```
Entropy of communities:
      P006      P018
4.403435 4.673620
Average entropy of the communities:
[1] 4.519057
```

The Shannon α entropy of the meta-community is 4.52. It is the weighted average entropy of communities.

Diversity partition of a meta-community

The `DivPart` function calculates everything at once. Its arguments are the same, but bias correction is not applied by default. It can be, using the argument `Biased = FALSE`, and the correction is chosen by the argument `Correction`. It returns a 'DivPart' object which can be summarized (entropy is not printed by `summary`) and plotted:

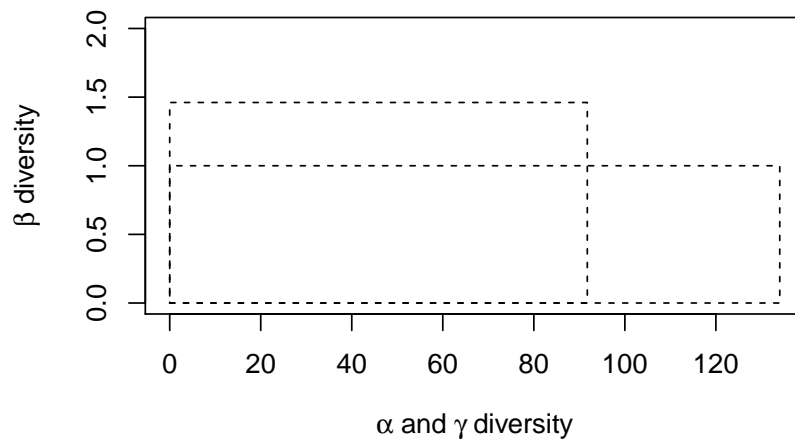


Figure 2: Plot of the diversity partition of the meta-community `Paracou618.MC`. The long rectangle of height 1 represents γ diversity, equal to 134 effective species. The narrower and higher rectangle has the same area: its horizontal size is α diversity (92 effective species) and its height is β diversity (1.46 effective communities).

```
R> p <- DivPart(q = 1, MC = Paracou618.MC, Biased = FALSE)
R> summary(p)
```

```
HCDT diversity partitioning of order 1 of metaCommunity Paracou618.MC
with correction: Best
Alpha diversity of communities:
  P006      P018
81.73115 107.08473
Total alpha diversity of the communities:
[1] 91.74905
Beta diversity of the communities:
[1] 1.460828
Gamma diversity of the metacommunity:
[1] 134.0296
```

```
R> p$CommunityAlphaEntropies
```

```
  P006      P018
4.403435 4.673620
```

The α diversity of communities is 92 effective species, which is the exponential of the entropy calculated previously. This is more than Simpson diversity (73 species, calculated above), because less frequent species are taken into account. γ diversity of the meta-community is 134 effective species. β diversity is 1.46 effective communities, i.e., the two actual communities are as different from each other as 1.46 ones with equal weights and no species in common.

Diversity estimation of a meta-community

The `DivEst` function decomposes diversity and estimates confidence intervals of α , β and γ diversity following [Marcon et al. \(2012\)](#). If the observed species frequencies of a community

are assumed to be a realization of a multinomial distribution, they can be drawn again to obtain a distribution of entropy.

```
R> de <- DivEst(q = 1, Paracou618.MC, Biased = FALSE, Correction = "Best",
+   Simulations = 1000)
```

```
=====
```

```
R> summary(de)
```

```
Diversity partitioning of order 1 of MetaCommunity MC
with correction: Best
```

```
Alpha diversity of communities:
```

```
      P006      P018
```

```
81.73115 107.08473
```

```
Total alpha diversity of the communities:
```

```
[1] 91.74905
```

```
Beta diversity of the communities:
```

```
[1] 1.460828
```

```
Gamma diversity of the metacommunity:
```

```
[1] 134.0296
```

```
Quantiles of simulations (alpha, beta and gamma diversity):
```

0%	1%	2.5%	5%	10%	25%	50%	
80.67265	83.71585	84.62767	85.59966	87.15648	89.44841	91.89227	
75%	90%	95%	97.5%	99%	100%		
94.14842	96.49882	97.55203	98.98351	100.59886	103.39811		
0%	1%	2.5%	5%	10%	25%	50%	75%
1.388795	1.403627	1.416081	1.421326	1.430602	1.444479	1.461347	1.477984
90%	95%	97.5%	99%	100%			
1.492434	1.499196	1.505406	1.512316	1.526236			
0%	1%	2.5%	5%	10%	25%	50%	75%
119.4739	122.1538	124.1964	125.9115	127.6125	130.7931	134.1313	137.4341
90%	95%	97.5%	99%	100%			
140.3785	142.1482	143.7599	145.5349	149.8479			

The result is a ‘**Divest**’ object which can be summarized and plotted (Figure 3).

The uncertainty of estimation is due to sampling: The distribution of the estimators corresponds to the simulated repetitions of sampling from the original multinomial distribution of species. It ignores the remaining bias of the estimator, which is unknown. Yet, except for $q = 2$, the corrected estimators *are* biased (even though much less than the non-corrected ones), especially when q is small. New estimators to reduce the bias are included in the package regularly.

Diversity profile of a meta-community

DivProfile calculates diversity profiles, i.e., the value of diversity against its order (Figure 4). The result is a ‘**DivProfile**’ object which can be summarized and plotted.

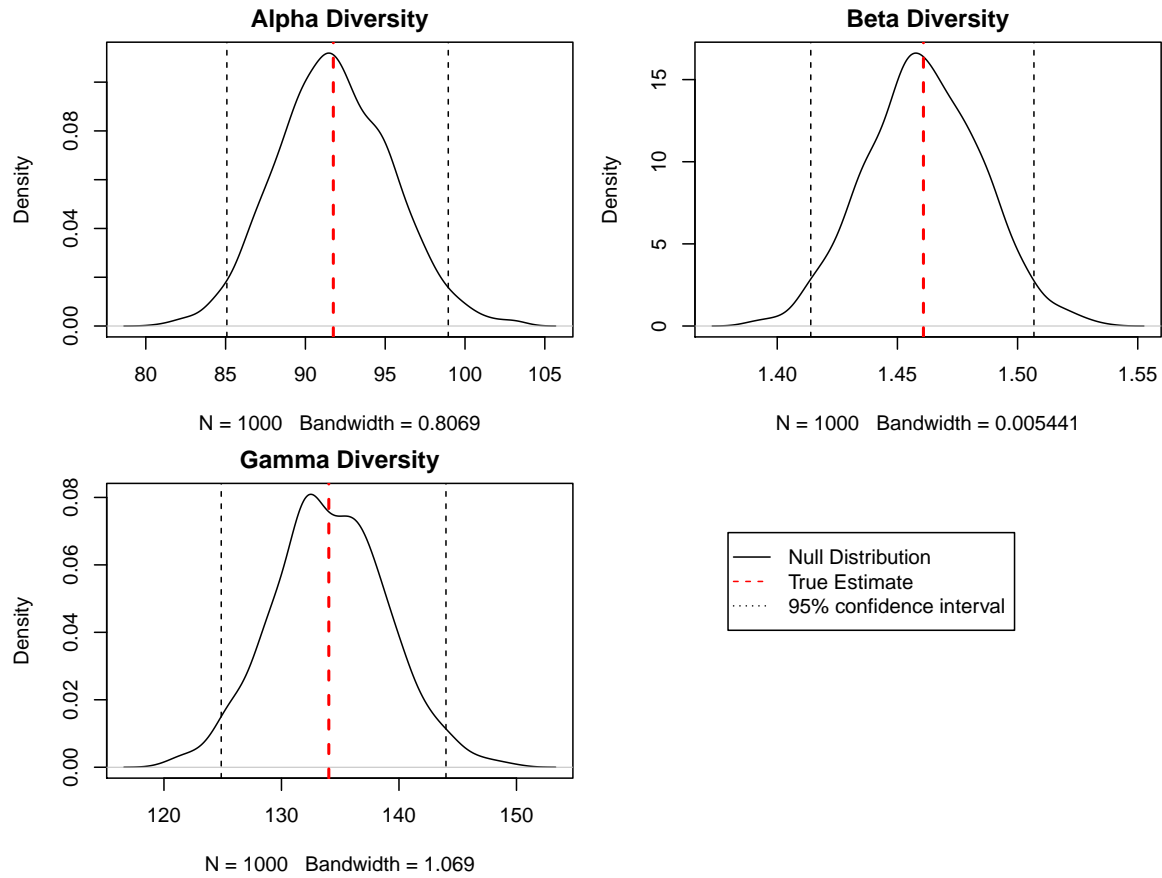


Figure 3: Plot of the diversity estimation of the meta-community Paracou618.MC. α , β and γ diversity probability densities are plotted, with a 95% confidence interval.

```
R> dp <- DivProfile(seq(0, 2, 0.2), Paracou618.MC, Biased = FALSE)
R> summary(dp)
```

Diversity profile of MetaCommunity MC

with correction: Best

Diversity against its order:

	Order	Alpha Diversity	Beta Diversity	Gamma Diversity
[1,]	0.0	205.84226	1.441996	296.82368
[2,]	0.2	181.63811	1.424471	258.73825
[3,]	0.4	157.35277	1.413780	222.46224
[4,]	0.6	133.77507	1.413903	189.14504
[5,]	0.8	111.70847	1.428705	159.59848
[6,]	1.0	91.74905	1.460828	134.02961
[7,]	1.2	75.51773	1.500587	113.32093
[8,]	1.4	63.95522	1.549024	99.06819
[9,]	1.6	55.37376	1.590012	88.04495
[10,]	1.8	48.97244	1.626123	79.63520
[11,]	2.0	44.21244	1.655569	73.19676

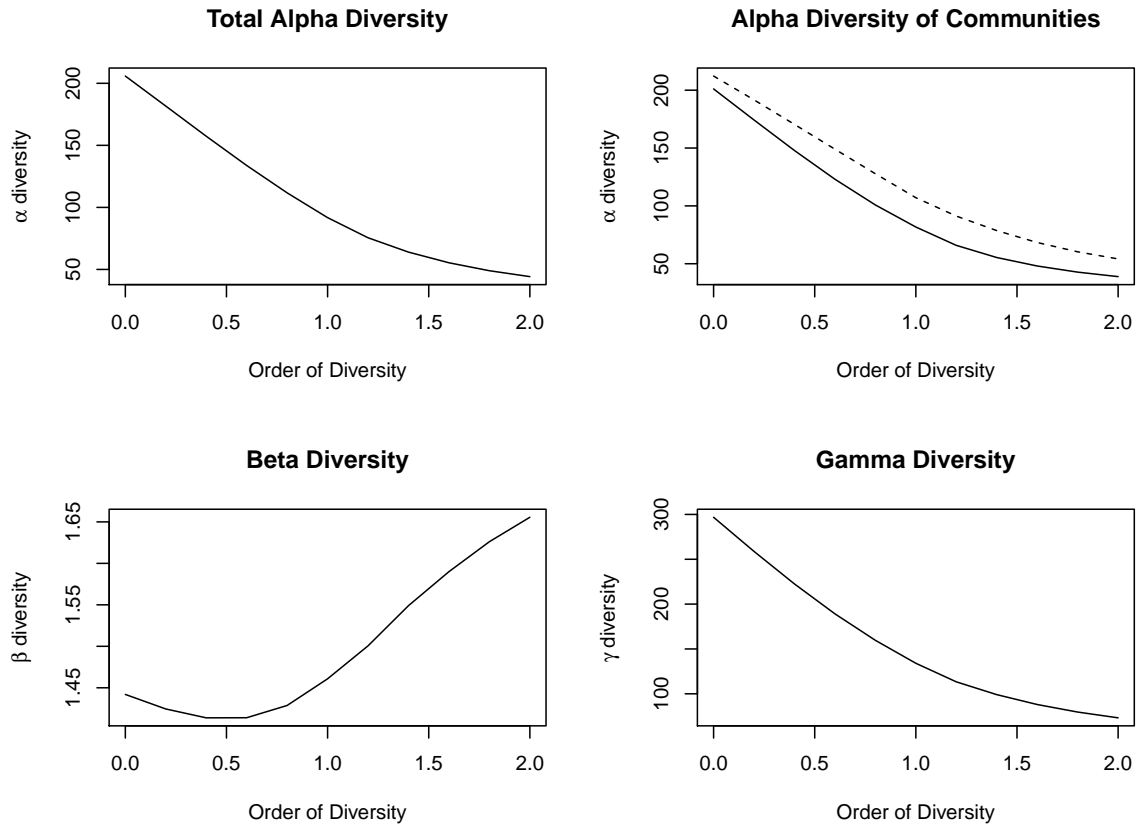


Figure 4: Diversity profile of the meta-community `Paracou618.MC`. Values are the number of effective species (α and γ diversity) and the effective number of communities (β diversity). Community P006 is represented by the solid line and community P018 by the dotted line. α and γ diversity decrease from $q = 0$ (number of species) to $q = 2$ (Simpson diversity) by construction.

Small orders of diversity give more weight to rare species. P018 can be considered more diverse than P006 because their profiles (Figure 4, top right) do not cross (Tothmeresz 1995): Its diversity is systematically higher. The shape of the β diversity profile shows that the communities are more diverse when their dominant species are considered.

Alternative functions

β entropy can also be calculated by a set of functions named after the community functions, such as `TsallisBeta`, `bcTsallisBeta`, `SimpsonBeta`, etc., which require two vectors of abundances or probabilities instead of a `MetaCommunity` object: that of the community and the expected one (usually that of the meta-community). Bias correction is currently limited to Chao and Shen's correction. The example below calculates the Shannon β entropy of the first community of `Paracou618` and the meta-community.

```
R> ShannonBeta(Paracou618.MC$Psi[, 1], Paracou618.MC$Ps)
```

```
[1] 0.3499358
```

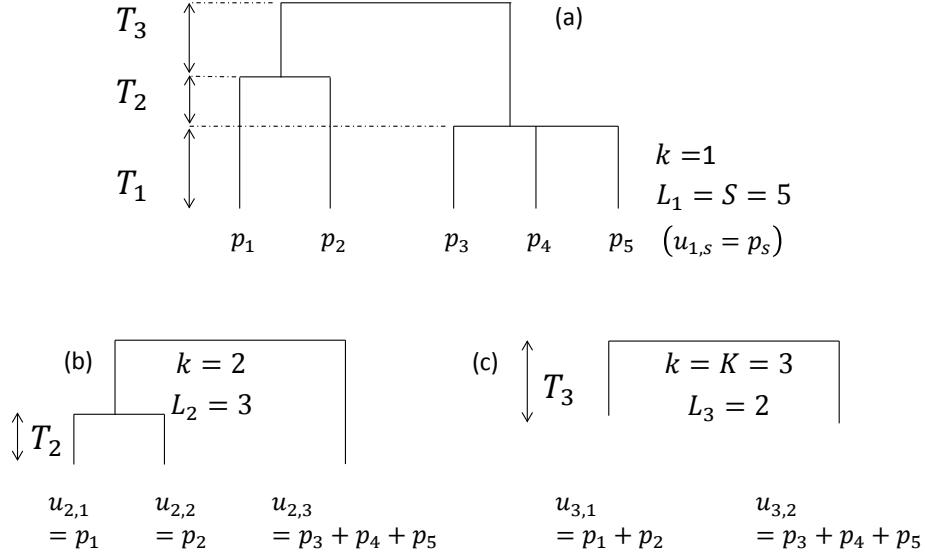


Figure 5: Hypothetical ultrametric tree. (a) The whole tree contains three slices, delimited by two nodes. The length of slices is T_k . (b) Focus on slice 2. The tree without slice 1 is reduced to 3 leaves. Frequencies of collapsed species are $u_{k,l}$. (c) Slice 3 only.

These functions are available to be particularly used, when a ‘**MetaCommunity**’ object is not available or not convenient to use (e.g., simulations). Meta-community functions are preferred in general.

4. Phylogenetic diversity

Phylogenetic or functional diversity generalizes HCDT diversity, considering the distance between species (Marcon and Hérault 2015a). Here, all species take place in an ultrametric phylogenetic or functional tree (Figure 5). The tree is cut into slices, delimited by two nodes. The first slice starts at the bottom of the tree and ends at the first node. In slice k , L_k leaves are found. The probabilities of occurrence of the species belonging to branches that were below leaf l in the original tree are summed to give the grouped probability $u_{k,l}$. HCDT entropy can be calculated in slice k :

$${}_k^q H = - \sum_l u_{k,l}^q \ln_q u_{k,l}. \quad (10)$$

Then, it is summed over the tree slices. Phyloentropy can be normalized or not. We normalize it so that it does not depend on the tree height:

$${}^q \overline{H}(T) = \sum_{k=1}^K \frac{T_k}{T} {}_k^q H. \quad (11)$$

Unnormalized values are multiplied by the tree height, such as ${}^q PD(T)$ (Chao *et al.* 2010).

Phyloentropy is calculated as HCDDT entropy along the slices of the trees applying possible corrections of estimation bias, summed, possibly normalized, and finally transformed into diversity:

$${}^q\overline{D}(T) = e_q^{{}^q\overline{H}(T)}. \quad (12)$$

4.1. Community functions

`PhyloEntropy` and the estimation bias corrected `bcPhyloEntropy` are the phylogenetic analogs of `Tsallis` and `bcTsallis`. They accept the same arguments plus an ultrametric tree of class ‘`hclust`’ or ‘`phylog`’, and `Normalize`, a Boolean to normalize the tree height to 1 (by default). Phylogenetic diversity is calculated by `PhyloDiversity` or `bcPhyloDiversity`, analogous to the species-neutral diversity functions `Diversity` and `bcDiversity`.

Results are either a ‘`PhyloDiversity`’ or a ‘`PhyloEntropy`’ object, which can be plotted (Figure 6) and summarized.

```
R> phd <- bcPhyloDiversity(Paracou618.MC$Ns, q = 1,
+   Tree = Paracou618.Taxonomy, Normalize = TRUE)
R> summary(phd)
```

```
alpha or gamma phylogenetic or functional diversity of order 1
of distribution Paracou618.MC$Ns
  with correction: Best
Phylogenetic or functional diversity was calculated according to the tree
Paracou618.Taxonomy
```

```
Diversity is  normalized
```

```
Diversity equals: 55.13383
```

The phylogenetic diversity of order 1 of the Paracou dataset is 55 effective species: 55 totally different species (only connected by the root of the tree) with equal probabilities would have the same entropy. It can be compared to its species-neutral diversity, 134 species. The latter is the diversity of the first slice of the tree. When going up the tree, diversity decreases because species collapse. On Figure 6, diversity of the second slice, between $T = 1$ and $T = 2$, is that of genera (64 effective genera) and the last slice contains (20 effective families). The phylogenetic entropy of the community is the average of the entropy along slices, weighted by the slice lengths. Diversity cannot be averaged in the same way.

A less trivial phylogeny would contain many slices, resulting in as many diversity levels with respect to T .

The `AllenH` function is similar to `PhyloEntropy`: It also calculates phyloentropy but the algorithm is that of [Allen, Kon, and Bar-Yam \(2009\)](#) for $q = 1$ and that of [Leinster and Cobbold \(2012\)](#) for $q \neq 1$. It is much faster since it does not require calculating entropy for each slice of the tree but it does not allow correction of estimation bias. `ChaoPD` calculates phylodiversity according to [Chao *et al.* \(2010\)](#), with the same advantages and limits compared to `PhyloDiversity`.

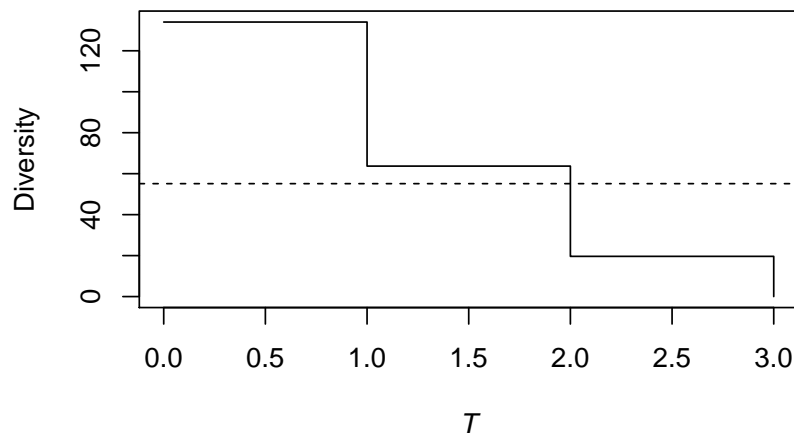


Figure 6: Plot of the γ phylodiversity estimation of the meta-community `Paracou618.MC`. The effective number of taxa of Shannon diversity is plotted against the distance from the leaves of the phylogenetic tree. Here, the tree is based on a rough taxonomy, so diversity of species, genera and families are the three levels of the curve. The dotted line represents the value of phylodiversity.

For convenience, functions `PDFD` and `Rao` are provided to calculate unnormalized phyloentropy of order 0 and 2.

4.2. Meta-community functions

Functions `DivPart`, `DivEst` and `DivProfile` return phylogenetic entropy and diversity values instead of species-neutral ones if a tree is provided in the arguments.

```
R> dp <- DivPart(q = 1, Paracou618.MC, Biased = FALSE, Correction = "Best",
+   Tree = Paracou618.Taxonomy)
R> summary(dp)
```

```
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
37.22132 51.31045
Total alpha diversity of the communities:
[1] 42.70238
Beta diversity of the communities:
[1] 1.291119
```

Gamma diversity of the metacommunity:

```
[1] 55.13383
```

The decomposition is interpreted as the species-neutral one: γ diversity is 55 effective species, made of 1.3 effective communities of 43 effective species.

Other meta-community functions, such as `AlphaEntropy` behave in the same way:

```
R> summary(BetaEntropy(Paracou618.MC, q = 2, Tree = Paracou618.Taxonomy,
+   Correction = "None", Normalize = FALSE))
```

```
HCDT beta entropy of order 2 of metaCommunity Paracou618.MC
with correction: None
```

```
Phylogenetic or functional entropy was calculated according to the tree
Paracou618.Taxonomy
```

```
Entropy is not normalized
```

```
Entropy of communities:
```

```
      P006      P018
0.04117053 0.02325883
```

```
Average entropy of the communities:
```

```
[1] 0.03350547
```

Compare with Rao's `divc` from package **ade4**:

```
R> library("ade4")
R> divc(as.data.frame(Paracou618.MC$Wi), disc(as.data.frame(
+   Paracou618.MC$Nsi), Paracou618.Taxonomy$Wdist))
```

```
              diversity
Paracou618.MC$Wi 0.03350547
```

5. Similarity-based diversity

[Leinster and Cobbold \(2012\)](#) introduced similarity-based diversity of a community ${}^qD^Z$. A matrix **Z** describes the similarity between pairs of species, defined between 0 and 1. A species' ordinariness is its average similarity with all species (weighted by species frequencies), including similarity with itself (equal to 1). Similarity-based diversity is the reciprocal of the generalized average of order q ([Hardy, Littlewood, and Pólya 1952](#)) of the community species' ordinariness.

The `Dqz` function calculates similarity-based diversity. Its arguments are the vector of probabilities of occurrences of the species, the order of diversity and the similarity matrix **Z**. The `bcDqz` function allows correction of estimation bias.

This example calculates the γ diversity of the meta-community `Paracou`. First, the similarity matrix is calculated from the distance matrix between all pairs of species as 1 minus normalized dissimilarity.

```
R> DistanceMatrix <- as.matrix(Paracou618.dist)
R> Z <- 1 - DistanceMatrix/max(DistanceMatrix)
R> bcDqz(Paracou618.MC$Ns, q = 2, Z)
```

```
[1] 1.483027
```

If \mathbf{Z} is the identity matrix, similarity-based diversity equals HCDT diversity:

```
R> Dqz(Paracou618.MC$Ps, q = 2, Z = diag(length(Paracou618.MC$Ps)))
```

```
[1] 68.7215
```

```
R> Diversity(Paracou618.MC$Ps, q = 2)
```

```
[1] 68.7215
```

Functional diversity of order 2 is only 1.48 effective species, which is very small compared to 69 effective species for Simpson diversity. 1.48 equally-frequent species with similarity equal to 0 would have the same functional diversity as the actual community (made of 425 species). This means that species are very similar from a functional point of view. The very low values returned by ${}^qD^Z$ are questioned by [Chao *et al.* \(2014\)](#) and discussed in depth by [Marcon *et al.* \(2014b\)](#): The choice of the similarity matrix is not trivial.

The similarity-based entropy of a community ${}^qH^Z$ ([Leinster and Cobbold 2012](#); [Ricotta and Szeidl 2006](#)) has the same relations with diversity as HCDT entropy and Hill numbers. The `Hqz` function calculates it:

```
R> Hqz(Paracou618.MC$Ps, q = 2, Z)
```

```
[1] 0.3208152
```

```
R> lnq(Dqz(Paracou618.MC$Ps, q = 2, Z), q = 2)
```

```
[1] 0.3208152
```

As species-neutral entropy, ${}^qH^Z$ has no straightforward interpretation beyond the average surprise of a community.

All meta-community functions can be used to estimate similarity-based diversity. Argument \mathbf{Z} must be provided:

```
R> e <- AlphaEntropy(Paracou618.MC, q = 1, Z = Z)
R> summary(e)
```

```
Similarity-based alpha entropy of order 1 of metaCommunity
Paracou618.MC with correction: Best
```

Phylogenetic or functional entropy was calculated according to the similarity

```
matrix Z
```

```
Entropy of communities:
```

```
      P006      P018
0.3945541 0.3934725
```

```
Average entropy of the communities:
```

```
[1] 0.3940912
```

The α functional entropy of the meta-community is the average entropy of communities.

6. Advanced tools

The package comes with a set of tools to realize frequents tasks: run Monte Carlo simulations on a community, quickly calculate its diversity profile, apply a function to a species distribution along a tree, and manipulate meta-communities.

6.1. Entropy of Monte Carlo simulated communities

The `EntropyCI` function is a versatile tool to simplify simulations. Simulated communities are obtained by random draws from a multinomial distribution of species and their entropy is calculated. The arguments of `EntropyCI` are an entropy function (any entropy function of the package accepting a vector of species abundances, such as `bcTsallis`), the number of simulations to run and the observed species frequencies. The result is a numeric vector containing the entropy value of each simulated community. Entropy can be finally transformed into diversity (but it is not correct to use a diversity function in simulations because the average simulated value must be calculated and only entropy can be averaged).

This example shows how to use the function. First, the distribution of the γ HCDT entropy of order 1 (Shannon entropy) of the Paracou meta-community is calculated and transformed into diversity. Then, the actual diversity is calculated and completed by the 95% confidence interval of the simulated values.

```
R> SimulatedDiversity <- expq(EntropyCI(FUN = bcTsallis,
+   Simulations = 1000, Ns = Paracou618.MC$Ns, q = 1), q = 1)
```

```
=====
```

```
R> bcDiversity(Paracou618.MC$Ns, q = 1)
```

```
[1] 134.0296
```

```
R> quantile(SimulatedDiversity, probs = c(0.025, 0.975))
```

```
      2.5%      97.5%
124.7128 144.3154
```

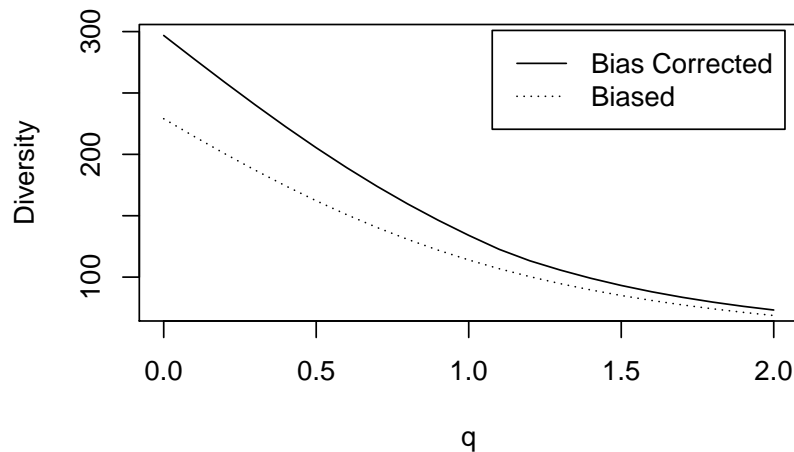


Figure 7: γ diversity profile of the the meta-community `Paracou618.MC`, without bias correction (dotted line) and with correction (solid line). .

These results are identical to those of the `DivEst` function but a single community can be addressed (`DivEst` requires a ‘`MetaCommunity`’ object).

6.2. Diversity or entropy profile of a community

This function is used to calculate diversity or entropy profiles based on community functions such as `Tsallis` or `ChaoPD`. It is similar to `DivProfile` but does not require a ‘`MetaCommunity`’ as argument. It returns a list which can be plotted.

This example evaluates bias correction on the diversity profile of the `Paracou` dataset. First, diversity profiles are calculated with and without bias correction:

```
R> bcProfile <- CommunityProfile(bcDiversity, Paracou618.MC$Ns)
R> Profile <- CommunityProfile(Diversity, Paracou618.MC$Ps)
```

Then, they can be plotted together (Figure 7):

```
R> plot(bcProfile, type = "l", main = "", xlab = "q", ylab = "Diversity")
R> lines(y ~ x, data = Profile, lty = 3)
R> legend("topright", c("Bias Corrected", "Biased"), lty = c(1, 3),
+       inset = 0.02)
```

6.3. Applying a function over a phylogenetic tree

The `PhyloApply` function is used to apply an entropy community function (generally `bcTsallis`) along a tree, the same way `lapply` works with a list.

This example shows how to calculate Shannon entropy along the tree containing the taxonomy to obtain species, genus and family entropy shown in Figure 6:

```
R> pa <- PhyloApply(Tree = Paracou618.Taxonomy, FUN = bcTsallis,
+   NorP = Paracou618.MC$Ns)
R> summary(pa)
```

```
bcTsallis applied to Paracou618.MC$Ns along the tree
Paracou618.Taxonomy
```

```
Results are normalized
```

```
The average value is: 4.009764
```

```
Values along the tree are:
```

```
      1      2      3
4.898061 4.154182 2.977048
```

```
R> exp(pa$Cuts)
```

```
      1      2      3
134.02961 63.69982 19.62979
```

```
R> exp(pa$Total)
```

```
[1] 55.13383
```

6.4. Manipulation of meta-communities

Several meta-communities, combined in a list, can be merged in two different ways. The `MergeMC` function simplifies hierarchical partitioning of diversity: It considers the aggregated data of each meta-community as a community and builds an upper-level meta-community with them. The α entropy of the new meta-community is the weighted average γ entropy of the original meta-communities.

`MergeC` combines the communities of several meta-communities to create a single meta-community containing them all. Finally, `ShuffleMC` randomly shuffles communities across meta-communities to allow simulations to test differences between meta-communities.

This example shows how to do this. First, one meta-community is created, with weights of communities proportional to their number of individuals:

```
R> (df <- data.frame(C1 = c(10, 10, 10, 10), C2 = c(0, 20, 35, 5),
+   C3 = c(25, 15, 0, 2), row.names = c("sp1", "sp2", "sp3", "sp4")))
```

```
      C1 C2 C3
sp1 10  0 25
sp2 10 20 15
sp3 10 35  0
sp4 10  5  2
```

```
R> w <- colSums(df)
```

```
R> MC1 <- MetaCommunity(Abundances = df, Weights = w)
```

Then a second one:

```
R> (df <- data.frame(C1 = c(10, 4), C2 = c(3, 4), row.names = c("sp1",
+ "sp5")))
```

```
      C1 C2
sp1 10  3
sp5  4  4
```

```
R> w <- colSums(df)
R> MC2 <- MetaCommunity(Abundances = df, Weights = w)
```

They can be merged to obtain a single meta-community containing all original communities:

```
R> mergedMC1 <- MergeC(list(MC1, MC2))
R> mergedMC1$Nsi
```

```
      MC1.C1 MC1.C2 MC1.C3 MC2.C1 MC2.C2
sp1      10      0      25      10      3
sp2      10     20      15       0      0
sp3      10     35       0       0      0
sp4      10      5       2       0      0
sp5       0      0       0       4      4
```

They can also be merged considering each of them as a community of a higher-level meta-community:

```
R> mergedMC2 <- MergeMC(list(MC1, MC2), Weights = sapply(list(MC1, MC2),
+ function(x) (x$N)))
R> mergedMC2$Nsi
```

```
      MC1 MC2
sp1  35  13
sp2  45   0
sp3  45   0
sp4  17   0
sp5   0   8
```

Hierarchical diversity partitioning can then be achieved:

```
R> dpAll <- DivPart(q = 1, MC = mergedMC2)
R> summary(dpAll)
```

HCDT diversity partitioning of order 1 of metaCommunity mergedMC2

Alpha diversity of communities:

```
      MC1      MC2
3.772161 1.943574
```

Total alpha diversity of the communities:


```
[1] 3.463277
Beta diversity of the communities:
[1] 1.236351
Gamma diversity of the metacommunity:
[1] 4.281826
```

The γ diversity of the top assemblage (MC1 and MC2) is 4.28 effective species, made of 1.24 effective meta-communities of 3.46 effective species. The α diversity of each meta-community of the top assemblage is their γ diversity when it is partitioned in turn:

```
R> dpMC1 <- DivPart(q = 1, MC = MC1)
R> summary(dpMC1)
```

HCDT diversity partitioning of order 1 of metaCommunity MC1

```
Alpha diversity of communities:
      C1      C2      C3
4.000000 2.429521 2.273918
Total alpha diversity of the communities:
[1] 2.741671
Beta diversity of the communities:
[1] 1.375862
Gamma diversity of the metacommunity:
[1] 3.772161
```

The γ diversity of MC1 is 3.77 effective species, made of 1.38 effective meta-communities of 2.74 effective species. The same decomposition can be done for MC2.

7. Conclusion

The **entropart** package allows estimating biodiversity according to the framework based on HCDT entropy, the correction of its estimation bias (Grassberger 1988; Chao and Shen 2003) and its transformation into equivalent numbers of species (Hill 1973; Jost 2006; Marcon *et al.* 2014a). Phylogenetic or functional diversity (Marcon and Hérault 2015a) can be estimated, considering phyloentropy as the average species-neutral diversity over slices of a phylogenetic or functional tree (Pavoine *et al.* 2009). Similarity-based diversity (Leinster and Cobbold 2012) can be used to estimate (Marcon *et al.* 2014b) functional diversity from a similarity or dissimilarity matrix between species without requiring building a dendrogram and thus preserving the topology of species (Pavoine, Ollier, and Dufour 2005; Podani and Schmera 2007).

The classical diversity estimators (Shannon and Simpson entropy) can be found in many R packages. Package **vegetarian** (Charney and Record 2012) allows calculating Hill numbers and partitioning them according to Jost's framework. Bias correction is never available except in the package **EntropyEstimation** (Cao and Grabchak 2015), which provides the Zhang and Grabchak's estimators of entropy and diversity and their asymptotic variances (not included in package **entropart**). Phylodiversity and similarity-based diversity are not available in any

package as far as we know. So we believe **entropart** is a useful toolbox for ecologists who need to estimate the diversity of actual, undersampled communities and to partition it.

Acknowledgments

This work has benefited from an “Investissement d’Avenir” grant managed by Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-0025).

References

- Allen B, Kon M, Bar-Yam Y (2009). “A New Phylogenetic Diversity Measure Generalizing the Shannon Index and Its Application to Phyllostomid Bats.” *American Naturalist*, **174**(2), 236–243. doi:[10.1086/600101](https://doi.org/10.1086/600101).
- Beck J, Holloway J, Schwanghart W (2013). “Undersampling and the Measurement of Beta Diversity.” *Methods in Ecology and Evolution*, **4**(4), 370–382. doi:[10.1111/2041-210x.12023](https://doi.org/10.1111/2041-210x.12023).
- Bonachela JA, Hinrichsen H, Muñoz MA (2008). “Entropy Estimates of Small Data Sets.” *Journal of Physics A: Mathematical and Theoretical*, **41**(202001), 1–9. doi:[10.1088/1751-8113/41/20/202001](https://doi.org/10.1088/1751-8113/41/20/202001).
- Cao L, Grabchak M (2015). “**EntropyEstimation**: Estimation of Entropy and Related Quantities.” R package version 1.2, URL <http://CRAN.R-project.org/package=EntropyEstimation>.
- Chao A, Chiu CH, Jost L (2010). “Phylogenetic Diversity Measures Based on Hill Numbers.” *Philosophical Transactions of the Royal Society B*, **365**(1558), 3599–3609. doi:[10.1098/rstb.2010.0272](https://doi.org/10.1098/rstb.2010.0272).
- Chao A, Chiu CH, Jost L (2014). “Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity and Related Similarity/Differentiation Measures Through Hill Numbers.” *Annual Review of Ecology, Evolution, and Systematics*, **45**, 297–324. doi:[10.1146/annurev-ecolsys-120213-091540](https://doi.org/10.1146/annurev-ecolsys-120213-091540).
- Chao A, Lee SM, Chen TC (1988). “A Generalized Good’s Nonparametric Coverage Estimator.” *Chinese Journal of Mathematics*, **16**, 189–199.
- Chao A, Shen T (2003). “Nonparametric Estimation of Shannon’s Index of Diversity When There Are Unseen Species in Sample.” *Environmental and Ecological Statistics*, **10**(4), 429–443.
- Chao A, Wang Y, Jost L (2013). “Entropy and the Species Accumulation Curve: A Novel Entropy Estimator Via Discovery Rates of New Species.” *Methods in Ecology and Evolution*, **4**(11), 1091–1100. doi:[10.1111/2041-210x.12108](https://doi.org/10.1111/2041-210x.12108).
- Charney N, Record S (2012). “**vegetarian**: Jost Diversity Measures for Community Data.” R package version 1.2, URL <http://CRAN.R-project.org/package=vegetarian>.

- Chiu CH, Chao A (2014). “Distance-Based Functional Diversity Measures and Their Decomposition: A Framework Based on Hill Numbers.” *PLOS ONE*, **9**(7), e100014. doi:[10.1371/journal.pone.0100014](https://doi.org/10.1371/journal.pone.0100014).
- Daróczy Z (1970). “Generalized Information Functions.” *Information and Control*, **16**(1), 36–51.
- Dauby G, Hardy O (2012). “Sampled-Based Estimation of Diversity Sensu Stricto by Transforming Hurlbert Diversities into Effective Number of Species.” *Ecography*, **35**(7), 661–672. doi:[10.1111/j.1600-0587.2011.06860.x](https://doi.org/10.1111/j.1600-0587.2011.06860.x).
- Dray S, Dufour A (2007). “The **ade4** Package: Implementing the Duality Diagram for Ecologists.” *Journal of Statistical Software*, **22**(4), 1–20. doi:[10.18637/jss.v022.i04](https://doi.org/10.18637/jss.v022.i04).
- Faith D (1992). “Conservation Evaluation and Phylogenetic Diversity.” *Biological Conservation*, **61**(1), 1–10. doi:[10.1016/0006-3207\(92\)91201-3](https://doi.org/10.1016/0006-3207(92)91201-3).
- Good I (1953). “On the Population Frequency of Species and the Estimation of Population Parameters.” *Biometrika*, **40**(3/4), 237–264. doi:[10.1093/biomet/40.3-4.237](https://doi.org/10.1093/biomet/40.3-4.237).
- Grassberger P (1988). “Finite Sample Corrections to Entropy and Dimension Estimates.” *Physics Letters A*, **128**(6–7), 369–373. doi:[10.1016/0375-9601\(88\)90193-4](https://doi.org/10.1016/0375-9601(88)90193-4).
- Hardy G, Littlewood J, Pólya G (1952). *Inequalities*. Cambridge University Press.
- Havrda J, Charvát F (1967). “Quantification Method of Classification Processes. Concept of Structural Alpha-Entropy.” *Kybernetika*, **3**(1), 30–35.
- Hérault B, Honnay O (2007). “Using Life-History Traits to Achieve a Functional Classification of Habitats.” *Applied Vegetation Science*, **10**(1), 73–80. doi:[10.1111/j.1654-109x.2007.tb00505.x](https://doi.org/10.1111/j.1654-109x.2007.tb00505.x).
- Hill M (1973). “Diversity and Evenness: A Unifying Notation and Its Consequences.” *Ecology*, **54**(2), 427–432. doi:[10.2307/1934352](https://doi.org/10.2307/1934352).
- Jost L (2006). “Entropy and Diversity.” *Oikos*, **113**(2), 363–375. doi:[10.1111/j.2006.0030-1299.14714.x](https://doi.org/10.1111/j.2006.0030-1299.14714.x).
- Jost L (2007). “Partitioning Diversity into Independent Alpha and Beta Components.” *Ecology*, **88**(10), 2427–2439. doi:[10.1890/06-1736.1](https://doi.org/10.1890/06-1736.1).
- Lande R (1996). “Statistics and Partitioning of Species Diversity, and Similarity Among Multiple Communities.” *Oikos*, **76**(1), 5–13. doi:[10.2307/3545743](https://doi.org/10.2307/3545743).
- Leinster T, Cobbold C (2012). “Measuring Diversity: The Importance of Species Similarity.” *Ecology*, **93**(3), 477–489. doi:[10.1890/10-2402.1](https://doi.org/10.1890/10-2402.1).
- Marcon E, Hérault B (2015a). “Decomposing Phylodiversity.” *Methods in Ecology and Evolution*, **6**(3), 333–339. doi:[10.1111/2041-210x.12323](https://doi.org/10.1111/2041-210x.12323).
- Marcon E, Hérault B (2015b). *entropart: Entropy Partitioning to Measure Diversity*. R package version 1.4.1, URL <http://CRAN.R-project.org/package=entropart>.

- Marcon E, Hérault B, Baraloto C, Lang G (2012). “The Decomposition of Shannon’s Entropy and a Confidence Interval for Beta Diversity.” *Oikos*, **121**(4), 516–522. doi:10.1111/j.1600-0706.2011.19267.x.
- Marcon E, Scotti I, Hérault B, Rossi V, Lang G (2014a). “Generalization of the Partitioning of Shannon Diversity.” *PLOS ONE*, **9**(3), e90289. doi:10.1371/journal.pone.0090289.
- Marcon E, Zhang Z, Hérault B (2014b). “The Decomposition of Similarity-Based Diversity and Its Bias Correction.” *HAL*, hal-00989454(version 1), 1–12.
- Patil G, Taillie C (1982). “Diversity as a Concept and Its Measurement.” *Journal of the American Statistical Association*, **77**(379), 548–561. doi:10.2307/2287709.
- Pavoine S, Love M, Bonsall M (2009). “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.” *Ecology Letters*, **12**(9), 898–908. doi:10.1111/j.1461-0248.2009.01344.x.
- Pavoine S, Ollier S, Dufour AB (2005). “Is the Originality of a Species Measurable?” *Ecology Letters*, **8**(6), 579–586. doi:10.1111/j.1461-0248.2005.00752.x.
- Petchey O, Gaston K (2002). “Functional Diversity (FD), Species Richness and Community Composition.” *Ecology Letters*, **5**(3), 402–411. doi:10.1046/j.1461-0248.2002.00339.x.
- Podani J, Schmera D (2007). “How Should a Dendrogram-Based Measure of Functional Diversity Function? A Rejoinder to Petchey and Gaston.” *Oikos*, **116**(8), 1427–1430. doi:10.1111/j.0030-1299.2007.16160.x.
- Rao C (1982). “Diversity and Dissimilarity Coefficients: A Unified Approach.” *Theoretical Population Biology*, **21**(1), 24–43. doi:10.1016/0040-5809(82)90004-1.
- R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- Ricotta C, Szeidl L (2006). “Towards a Unifying Approach to Diversity Measures: Bridging the Gap Between the Shannon Entropy and Rao’s Quadratic Index.” *Theoretical Population Biology*, **70**(3), 237–243. doi:10.1016/j.tpb.2006.06.003.
- Routledge R (1979). “Diversity Indices: Which Ones Are Admissible?” *Journal of Theoretical Biology*, **76**(4), 503–515. doi:10.1016/0022-5193(79)90015-8.
- Shannon C (1948a). “A Mathematical Theory of Communication.” *The Bell System Technical Journal*, **27**(3), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.
- Shannon C (1948b). “A Mathematical Theory of Communication.” *The Bell System Technical Journal*, **27**(4), 623–656. doi:10.1002/j.1538-7305.1948.tb00917.x.
- Simpson E (1949). “Measurement of Diversity.” *Nature*, **163**(4148), 688. doi:10.1038/163688a0.
- Tothmeresz B (1995). “Comparison of Different Methods for Diversity Ordering.” *Journal of Vegetation Science*, **6**(2), 283–290. doi:10.2307/3236223.

- Tsallis C (1988). “Possible Generalization of Boltzmann-Gibbs Statistics.” *Journal of Statistical Physics*, **52**(1), 479–487. doi:[10.1007/bf01016429](https://doi.org/10.1007/bf01016429).
- Tsallis C (1994). “What Are the Numbers That Experiments Provide?” *Química Nova*, **17**(6), 468–471.
- Wang JP (2011). “**SPECIES**: An R Package for Species Richness Estimation.” *Journal of Statistical Software*, **40**(9), 1–15. doi:[10.18637/jss.v040.i09](https://doi.org/10.18637/jss.v040.i09).
- Zhang Z, Huang H (2007). “Turing’s Formula Revisited.” *Journal of Quantitative Linguistics*, **14**(2–3), 222–241. doi:[10.1080/09296170701514189](https://doi.org/10.1080/09296170701514189).

Affiliation:

Eric Marcon
AgroParisTech
Campus agronomique, BP 316
97310 Kourou, French Guiana
E-mail: eric.marcon@ecofog.gf

Bruno Hérault
Cirad
Campus agronomique, BP 316
97310 Kourou, French Guiana
E-mail: bruno.herault@ecofog.gf

APPENDIX B

Tools to Characterize Point Patterns: dbmss for R

Marcon, E., S. Traissac, F. Puech et G. Lang (2015). « Tools to Characterize Point Patterns: dbmss for R ». In : Journal of Statistical Software 67.3, p. 1–15.



Tools to Characterize Point Patterns: **dbmss** for R

Eric Marcon
AgroParisTech
UMR EcoFoG

Stéphane Traissac
AgroParisTech
UMR EcoFoG

Florence Puech
RITM, Univ. Paris-Sud
Université Paris-Saclay

Gabriel Lang
AgroParisTech
UMR 518

Abstract

The **dbmss** package for R provides an easy-to-use toolbox to characterize the spatial structure of point patterns. Our contribution presents the state of the art of distance-based methods employed in economic geography and which are also used in ecology. Topographic functions such as Ripley's K , absolute functions such as Duranton and Overman's K_d and relative functions such as Marcon and Puech's M are implemented. Their confidence envelopes (including global ones) and tests against counterfactuals are included in the package.

Keywords: point patterns, spatial structure, R.

1. Introduction

Numerous researchers in various fields concern themselves with characterizing spatial distributions of objects. Amongst other questions, ecologists have been addressing the spatial attraction between species (Duncan 1991) or the non-independence of the location of dead trees in a forest (Haase, Pugnaire, Clark, and Incoll 1997). In addition of ecologists analyzing the spatial distribution of plants, economists may be concerned with the location of new entrants (Duranton and Overman 2008) or with the location of shops according to the types of good sold (Picone, Ridley, and Zandbergen 2009). In epidemiology, researchers want to identify the spatial distribution of sick individuals in comparison to the population (Diggle and Chetwynd 1991). In these research fields, the point process theory undoubtedly helps dealing with these questions. Exploratory statistics of point patterns widely rely on Ripley's seminal work (Ripley 1976, 1977), namely the K function. A recent review of similar methods is given by Marcon and Puech (2014) who called them distance-based measures of spatial concentration. We will refer to them here as spatial structures since both dispersion and concentration can be characterized. They are considered as novel and promising tools in spatial economics (Combes, Mayer, and Thisse 2008). The traditional approach to detect

localization, i.e., the degree of dissimilarity between the geographical distribution of an industry and that of a reference (Hoover 1936), relies on discrete space (a country is divided in regions for example) and measures of inequality between zones, such as the classical Gini (1912) index or the more advanced Ellison and Glaeser (1997) index. This approach suffers from several limitations, mainly the modifiable areal unit problem (MAUP): Results depend on the way zones are delimited and on the scale of observation (Openshaw and Taylor 1979). Distance-based methods have the advantage to consider space as continuous, i.e., without any zoning, allowing detecting spatial structures at all scales simultaneously and solving MAUP issues.

These methods estimate the value of a function of distance to each point calculated on a planar point pattern, typically objects on a map. They all consist in counting *neighbors* (up to or exactly at the chosen distance) around each *reference point* and transforming their number into a meaningful statistic. There are basically three possible approaches: just count neighbors, count neighbors per surface area or calculate the proportion of neighbors of interest among all neighbors. These approaches define three families of functions: absolute (how many neighbors are there?), topographic (how many neighbors per unit of area?) and relative (what is the ratio of neighbors of interest?). The function values are not the main motivation. The purpose is rather to test the point pattern against the null hypothesis that it is a realization of a known point process which does not account for a property of interest. The basic purpose of Ripley's K is to test the observed point pattern against complete spatial randomness (CSR), i.e., a homogeneous Poisson process, to detect dependence between point locations (the null hypothesis supposes independent points) assuming homogeneity (i.e., the probability to find a point is the same everywhere). Ripley-like functions, available in the proposed R (R Core Team 2015) **dbmss** package (Marcon, Lang, Traissac, and Puech 2015), can be classified in three families:

- Topographic measures such as K take space as their reference. They have been widely used in ecology (Fortin and Dale 2005). They have been built from the point process theory and have a strong mathematical background.
- Relative measures such as M (Marcon and Puech 2010) compare the structure of a point type to that of another type (they can be considered as cases and controls). They have been developed in economics, where comparing the distribution of a sector of activity to that of the whole economic activity is a classical approach (Combes *et al.* 2008), but introduced only recently in ecology (Marcon, Puech, and Traissac 2012).
- Absolute functions such as K_d (Duranton and Overman 2005) have no reference at all but their value can be compared to the appropriate null hypothesis to test it.

Relative and absolute functions have been built from descriptive statistics of point patterns, not related to the underlying point processes, so they are seen as heuristic and ignored by the statistical literature (Illian, Penttinen, Stoyan, and Stoyan 2008). Topographic functions are implemented in the **spatstat** package (Baddeley and Turner 2005) for R but absolute and relative functions are missing. We fill this gap by proposing the **dbmss** package, which is available from the Comprehensive R Archive Network (CRAN) at <http://CRAN.R-project.org/package=dbmss>. It makes the computation of the whole set of distance-based methods simple for empirical researchers by introducing measures that are not available elsewhere and

wrapping some topographic measures available in **spatstat** so that all can be used the same way.

Estimated values of the functions must be tested against a null hypothesis. The usual empirical way to characterize a spatial structure consists in computing the appropriate function and comparing it to the quantiles of a large number of simulations of the null hypothesis to reject (Kenkel 1988). We propose extended possibilities to evaluate confidence envelopes, including “global envelopes” (Duranton and Overman 2005), a goodness-of-fit test (Diggle 1983) and an analytical test (Lang and Marcon 2013; Marcon, Traissac, and Lang 2013).

Definitions of all functions and formulas for their estimation can be found in Marcon and Puech (2014) and are not repeated here, but they are summarized in Section 2 on the statistical background. Their implementation is presented in Section 3 on the package content.

2. Rationale and statistical background

We consider a map of points which often represents establishments in economic geography or trees in vegetation science. These points have two marks: a type (an industrial sector, a species, ...) and a weight (a number of employees, a basal area, ...). We want to apply to this point pattern a variety of exploratory statistics which are functions of distance between points and to test the null hypothesis of independence between point locations. These functions are either topographic, absolute or relative. They can be interpreted as the ratio between the observed number of neighbors and the expected number of neighbors if points were located independently from each other. If reference and neighbor points are of the same type, the functions are univariate and allow to study concentration or dispersion. They are bivariate, if the types differ, and allow to address the colocation of types. In the following we detail this approach.

2.1. Topographic, homogeneous functions

Topographic, homogeneous functions are Ripley’s K and its derivative g . Their null hypothesis is a Poisson homogeneous process: Rejecting it means that the process underlying the observed pattern is either not homogeneous or not independent. These functions are applied when homogeneity is assumed so independence only is tested by comparing the observed values of the function to their confidence envelope under CSR. Bivariate functions are tested against the null hypothesis of random labeling (point locations are kept unchanged but marks are redistributed randomly) or population independence (the reference point type is kept unchanged, the neighbor point type is shifted) following Goreaud and Pélissier (2003). The random labeling hypothesis considers that points preexist and their marks are the result of a process to test (e.g., are dead trees independently distributed in a forest?). The population independence hypothesis considers that points belong to two different populations with their own spatial structure and wants to test whether they are independent from each other.

Edge effect correction is compulsory to compute topographic functions: Points located close to boundaries have less neighbors because of the lack of knowledge outside the observation window. The **spatstat** package provides corrections following Ripley (1988), which we use.

2.2. Topographic, inhomogeneous functions

K_{inhom} (Baddeley, Møller, and Waagepetersen 2000) is the generalization of K to inhomogeneous processes: It tests independence of points assuming the intensity of the process is known. Empirically, it generally has to be estimated from the data where assumptions on the way to do this rely on theoretical knowledge of the process. The null hypothesis (“random position”) is that the pattern comes from an inhomogeneous Poisson process of this intensity, which can be simulated. Applying K_{inhom} to a single point type allows using the “random location” null hypothesis, following Durantón and Overman (2005): Observed points (with their marks) are shuffled among observed locations to test for independence. Bivariate K_{inhom} null hypotheses may be random labeling or population independence as defined by Marcon and Puech (2010): Reference points are kept unchanged, other points are redistributed across observed locations.

K_{mm} (Penttinen 2006; Penttinen, Stoyan, and Henttonen 1992) generalizes K to weighted points (weights are continuous marks of the points). Its null hypothesis in **dbmss** is random location. Penttinen *et al.* (1992) inferred the point process from the point pattern, and used the inferred process to simulate the null hypothesis patterns. This requires advanced spatial statistics techniques and knowledge about the process that is generally not available. The random location hypothesis is a way to draw null patterns simply, but ignores the stochasticity of the point process.

The D (Diggle and Chetwynd 1991) function compares the K function of points of interest (cases) to that of other points (controls). Its null hypothesis is random labeling.

2.3. Absolute functions

In their seminal paper, Durantón and Overman (Durantón and Overman 2005) study the distribution of industrial establishments in Great Britain. Every establishment, represented by a point, is characterized by its position (geographic coordinates), its sector of activity (point type) and its number of employees (point weight). The K_d function (Durantón and Overman 2005) is the probability density to find a neighbor a given distance apart from a point of interest in a finite point process. The K^{emp} function integrates the weights of points: It is the density probability to find an employee r apart from an employee of interest.

K_d and K^{emp} are absolute measures since their value is not normalized by the measure of space or any other reference: For a binomial process, K_d increases proportionally to r if the window is large enough to ignore edge effects (the probability density is proportional to the perimeter of the circle of radius r , Bonneu and Thomas-Agnan 2015), then edge effects make it decrease to 0 when r becomes larger than the window’s size: It is a bell-shaped curve. K_d values are not interpreted but compared to the confidence envelope of the null hypothesis, which is random location. The null hypothesis of bivariate functions is random labeling, following Durantón and Overman (2005), i.e., point types are redistributed across locations while weights are kept unchanged, or population independence (as for K_{inhom}). It is not corrected for edge effects. K_d was designed to characterize the spatial structure of an economic sector, comparing it to the distribution of the whole activity. From this point of view, it has been considered as a relative function (Marcon and Puech 2010). We prefer to be more accurate and distinguish it from strict relative functions which directly calculate a ratio or a difference between the number of points of the type of interest and the total number of points. What makes it relative is only its null hypothesis: Changing

it for random location (that of univariate K_{inhom}) would make univariate K_d behave as a topographic function (testing independence of the distribution supposing its intensity is that of the whole activity).

K_d is a leading tool in spatial economics. A great number of its applications can be found in the literature that confirms the recent interest for distance-based methods in spatial economics. A recent major study can be found in [Ellison, Glaeser, and Kerr \(2010\)](#).

2.4. Relative functions

The univariate and bivariate M function ([Marcon and Puech 2010](#)) is the ratio of neighbors of interest up to distance r normalized by its value over the whole domain. Their null hypotheses are the same as K_d 's. They do not suffer from edge effects. [Marcon and Puech \(2010\)](#) show that the M function respects most of the axioms generally accepted as the “good properties” to evaluate geographic concentration in spatial economics ([Combes and Overman 2004](#); [Duranton and Overman 2005](#)).

2.5. Unification

Empirically, all estimators can be seen as variations in a unique framework: Neighbors of each reference point are counted, their number is averaged and divided by a reference measure. Finally, this average local result is divided by its reference value, calculated over the whole point pattern instead of around each point.

Choosing reference and neighbor point types allows defining univariate or bivariate functions, counting neighbors up to or at a distance defines cumulative or density functions, taking an area or a number of points as the reference measure defines topographic or relative functions. These steps are detailed for two functions to clarify them: We focus on Ripley's g and Marcon and Puech's M bivariate function. See [Marcon and Puech \(2014\)](#) for a full review.

Reference points are denoted x_i , neighbor points are x_j . For density functions such as g , neighbors of x_i are counted at a chosen distance r :

$$n(x_i, r) = \sum_{j, i \neq j} k(\|x_i - x_j\|, r) c(i, j) \quad (1)$$

$k(\|x_i - x_j\|, r)$ is a kernel estimator, necessary to evaluate the number of neighbors at distance r , and $c(i, j)$ is an edge-effect correction (points located close to boundaries have less neighbors because of the lack of knowledge outside the observation window).

To compute the bivariate M function, reference points are of a particular type in a marked point pattern: $x_i \in \mathcal{R}$, where \mathcal{R} is the set of points of the reference type. Neighbors of the chosen type are denoted $x_j \in \mathcal{N}$. In cumulative functions such as M , neighbors are counted up to r :

$$n(x_i, r) = \sum_{x_j \in \mathcal{N}, i \neq j} \mathbf{1}(\|x_i - x_j\| \leq r) w(x_j). \quad (2)$$

Points can be weighted, i.e., $w(x_j)$ is the neighbor's weight.

The number of neighbors is then averaged. n is the number of reference points:

$$\bar{n}(r) = \frac{1}{n} \sum_{i=1}^n n(x_i, r). \quad (3)$$

The average number of neighbors is compared to a reference measure. It may be a measure of space (the perimeter of the circle of radius r for g), defining topographic functions:

$$m(r) = 2\pi r. \quad (4)$$

It may also be the number of neighbors of all types in a relative function such as M :

$$m(r) = \sum_{j, i \neq j} \mathbf{1}(\|x_i - x_j\| \leq r) w(x_j). \quad (5)$$

Finally, $\frac{\bar{n}(r)}{m(r)}$ is compared to the same ratio computed on the whole window. For g , this gives:

$$\frac{\bar{n}_0}{m_0} = \frac{n-1}{A}. \quad (6)$$

A is the area of the window, \bar{n}_0 and m_0 are the limit values of $\bar{n}(r)$ and $m(r)$ when r gets larger than the window's size. For M , it becomes:

$$\frac{\bar{n}_0}{m_0} = \frac{1}{n} \sum_{i=1}^n \frac{W_{\mathcal{N}}}{W - w(x_i)} \quad (7)$$

$W_{\mathcal{N}}$ is the total weight of neighbor points, W that of all points. Finally, despite the functions being quite different (density vs. cumulative, topographic vs. relative, univariate vs. bivariate), both estimators can be written as $\frac{\bar{n}}{m} / \frac{\bar{n}_0}{m_0}$. Their value (except for absolute functions) can be interpreted as a location quotient: $g(r) = 2$ or $M(r) = 2$ means that twice more neighbors are observed at (or up to) distance r than expected on average, i.e., ignoring the point locations in the window. The appropriate function will be chosen from the toolbox according to the question raised.

3. Package content

The **dbmss** package contains a full (within the limits of the literature reviewed in Section 2) set of functions to characterize the spatial structure of a point pattern, including tools to compute the confidence interval of the counterfactual. It allows addressing big datasets thanks to C++ code used to calculate distances between pairs of points (using **Rcpp** infrastructure, [Eddelbuettel and François 2011](#)). Computational requirements actually are an issue starting from say 10,000 points (see [Ellison et al. 2010](#), for instance). Memory requirement is $O(n)$, i.e., proportional to the number of points to store their location and type. We use loops to calculate distances and increment summary statistics rather than store a distance matrix which is $O(n^2)$, following [Scholl and Brenner \(2013\)](#). Computation time is $O(n^2)$ because $n(n-1)/2$ pair distances must be calculated. A 100,000-point set requires around 4 minutes to calculate M on a laptop computer with an i5 Intel CPU. A confidence envelope built from 1000 simulations requires about 3 days.

We consider planar point patterns (sets of points in a 2-dimensional space) with marks of a special kind: Each point comes with a continuous mark (its weight) and a discrete one (its type). We call this special type of point pattern “weighted, marked, planar point patterns” and define objects of class ‘**wmppp**’, which inherits from class ‘**ppp**’ as defined in **spatstat**.

Marks are a dataframe with two columns, `PointWeight` containing the weights of points, and `PointTypes` containing the types, as factors.

A ‘`wmppp`’ object can be created by the `wmppp()` function which accepts a dataframe as argument, or converted from a ‘`ppp`’ object by `as.wmppp()`. Starting from a CSV file containing point coordinates, their type and their weight in four columns, a ‘`wmppp`’ object can be created by just reading the file with `read.csv()` and applying `wmppp()` to the result. Options are available to specify the observation window or guess it from the point coordinates and set default weights or types to points when they are not in the dataframe, see the package help for details. The simplest code to create a ‘`wmppp`’ object with 100 points is as follows. It draws point coordinates between 0 and 1, and creates a ‘`wmppp`’ object with a default window, all points are of the same type named “All” and their weight is 1.

```
R> Pattern <- wmppp(data.frame(X = runif(100), Y = runif(100)))
R> summary(Pattern)
```

```
Marked planar point pattern: 100 points
Average intensity 106 points per square unit
Mark variables: PointWeight, PointType
Summary:
```

```
  PointWeight PointType
Min.      :1      All:100
1st Qu.:1
Median :1
Mean    :1
3rd Qu.:1
Max.    :1
```

```
Window: rectangle = [2.96e-05, 0.968919] x [0.0267366, 0.9794786] units
Window area = 0.923102 square units
```

3.1. Distance-based functions

All functions are named `Xhat` where `X` is the name of the function: Ripley’s g and K ; K ’s normalization; Besag’s L (1977); Penttinen’s K_{mm} and L_{mm} ; Diggle and Chetwynd’s D ; Baddeley et al.’s K_{inhom} and its derivative g_{inhom} ; Marcon and Puech’s M and Duranton and Overman’s K_d (including its weighted version K^{emp}). The suffix `hat` has been used to avoid confusion with other functions in R, e.g., `D` already exists in the `stats` package. Arguments are:

- A weighted, marked planar point pattern (a ‘`wmppp`’ class object). The window can be a polygon or a binary image, as in `spatstat`.
- A vector of distances.
- Optionally a reference and a neighbor point type to calculate bivariate functions, or equivalently the types of cases and controls for the D function.
- Some optional arguments, specific to some functions.

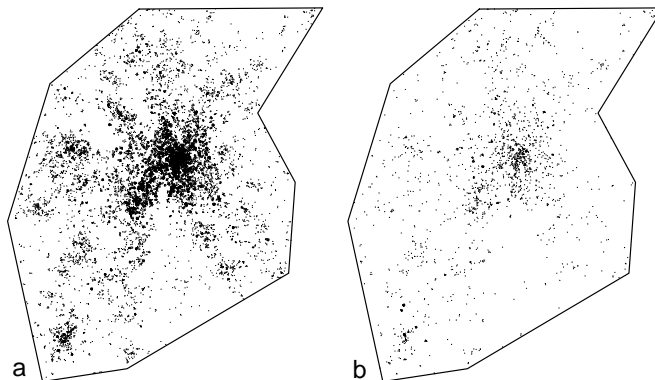


Figure 1: Map of emergencies in the urban area of Toulouse, France, during year 2004 (about 33 km from south to north). (a) 20,820 emergencies have been recorded and mapped (many points are confused at the figure scale). (b) Locations of the 10 percent most serious ones.

Topographic functions require edge-effect corrections, provided by **spatstat**: The **best** correction is systematically used. Relative functions ignore the window. Technical details are provided in the help files.

These functions return an ‘fv’ object, as defined in **spatstat**, which can be plotted.

3.2. Confidence envelopes

The classical confidence intervals, calculated by Monte Carlo simulations (Kenkel 1988) are obtained by the **XEnvelope** function, where **X** is the function’s name. Arguments are the number of simulations to run, the risk level, those of the function and the null hypothesis to simulate. These functions return a ‘**dbmssEnvelope**’ object which can be plotted.

Null hypotheses have been discussed by Goreaud and Pélissier (2003) for topographic functions such as K and by Marcon and Puech (2010) for relative functions. The null hypothesis for univariate functions is random position (points are drawn from a Poisson process for topographic functions) or random location (points are redistributed across actual locations for relative functions). Bivariate functions support random labeling and population independence as null hypotheses. The possible values of arguments are detailed in the help file of each function.

Building a confidence envelope in this way is problematic because the test is repeated at each distance. The underestimation of the risk has been discussed by Loosmore and Ford (2006). Duranton and Overman (2005) proposed a global envelope computed by the repeated elimination of simulations reaching an extreme value at any distance until the desired level is reached. The argument **Global = TRUE** is used to obtain it instead of the local one.

3.3. Examples

We illustrate the main features of the package by two examples. The first one comes from the economic literature (Bonneu 2007)¹. A point pattern is induced by data about 20,820

¹The dataset can be downloaded from: <http://publications-sfds.fr/index.php/csbigs/article/downloadSuppFile/376/69>.

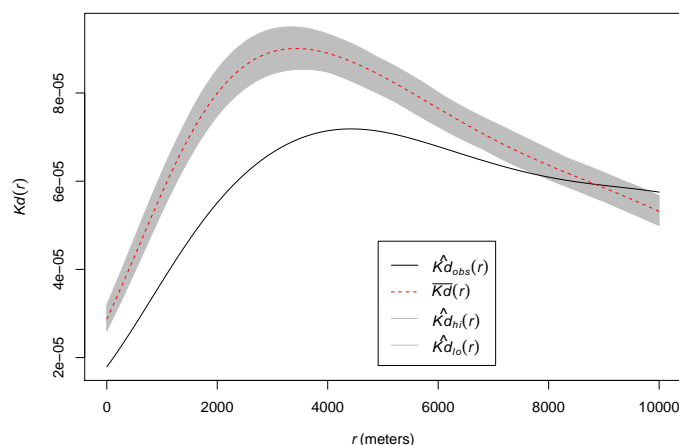


Figure 2: Representation of $K_d(r)$ values of the 10% most serious emergencies in year 2004 in the Toulouse urban area, showing their significant dispersion at all distances up to approximately 8 km. The solid, black curve is K_d . The dotted red curve is the average simulated value and the shaded area is the confidence envelope under the null hypothesis of random location. The risk level is 5%, 1000 simulations have been run. Distances are in meters.

emergencies involving the fire department of the urban area around Toulouse, France, during the year 2004 (Figure 1). The workload associated to each emergency (the number of men \times hours it required) is known. The original study tested the dependence between workload and location of emergencies: It did not exclude the null hypothesis of random labeling. We have a complementary approach here: We consider the 10 percent more serious emergencies, i.e., those which caused the highest workload. K_d may detect concentration (or dispersion) if, at a distance r from a serious emergency, the probability to find another serious emergency is greater (or lower) than that of finding an emergency regardless of its workload:

```
R> load("CSBIGS.Rdata")
R> Category <- cut(Emergencies$M, quantile(Emergencies$M, c(0, 0.9, 1)),
+   labels = c("Other", "Biggest"), include.lowest = TRUE)
R> X <- wmpvp(data.frame(X = Emergencies$X, Y = Emergencies$Y,
+   PointType = Category), win = Region)
R> KdE <- KdEnvelope(X, r = seq(0, 10000, 100), NumberOfSimulations = 1000,
+   ReferenceType = "Biggest", Global = TRUE)
R> plot(KdE)
```

The **Emergencies** data frame contains point coordinates (in meters) in columns **X** and **Y** and workload in column **M**. The second line of the code creates a vector containing a factor describing the workload to separate the 10% highest values. A ‘wmpvp’ object is created then, containing the points and their mark. The **KdEnvelope** function is run from 0 to 10 km by steps of 100 m for the most serious emergencies. Figure 2 shows that the 10% most serious emergencies are more dispersed than the distribution of all emergencies. This opens the way to discuss on the optimal location of fire stations.

The second example uses the **paracou16** point pattern (Figure 3) provided in the package. It represents the distribution of trees in a 4.1-ha tropical forest plot in the Paracou field

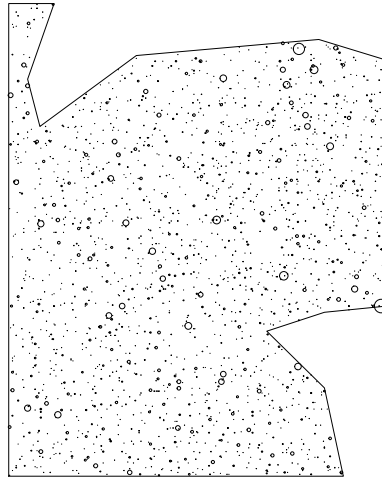


Figure 3: `paracou16` point pattern. Circles are centered on trees in a 4.1-ha forest plot (the containing rectangle is 200 m wide by 250 m long). Circle sizes are proportional to the basal areas of trees.

station in French Guiana (Gourlet-Fleury, Guehl, and Laroussinie 2004). It contains 2426 trees, where the species is either *Qualea rosea*, *Vouacapoua americana* or *Other* (one of more than 300 species). Weights are basal areas (the area of the stems virtually cut 1.3 meter above ground), measured in square centimeters.

```
R> data("paracou16", package = "dbmss")
R> plot(paracou16)
```

The question to test is dependence between the distributions of the two species of interest. Bivariate $M(r)$ is calculated for r between 0 and 30 meters. 1000 simulations are run to build the global confidence envelope.

```
R> Envelope <- MEnvelope(paracou16, r = seq(0, 30, 2),
+   NumberOfSimulations = 1000, Alpha = 0.05,
+   ReferenceType = "V. Americana", NeighborType = "Q. Rosea",
+   SimulationType = "RandomLabeling", Global = TRUE)
R> plot(Envelope)
```

The calculated function (Figure 4) is M , showing the repulsion between *V. Americana* and *Q. rosea* up to 30 m. Significance is unclear, since the observed values of the function are very close to the lower bound of the envelope. The complete study, with a larger dataset giving significant results, can be found in Marcon *et al.* (2012).

3.4. Goodness-of-fit test

A goodness-of-fit test for K has been proposed by Diggle (1983), applied to K by Loosmore and Ford (2006) and to M by Marcon *et al.* (2012). It calculates the distance between the actual values of the function and its average value obtained in simulations of the null hypothesis. The same distance is calculated for each simulated point pattern, and the returned

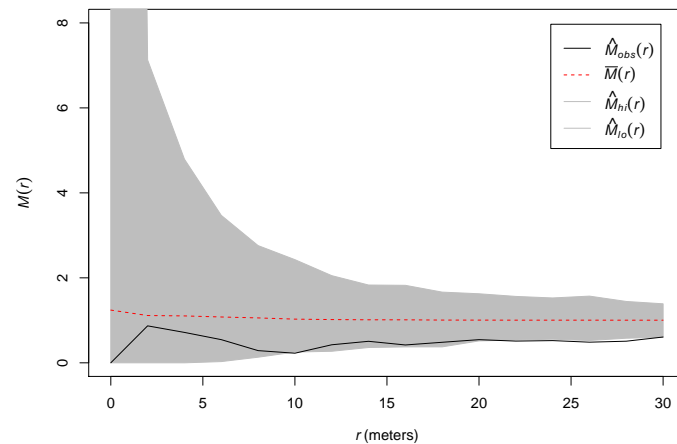


Figure 4: Representation of $M(r)$ values of *Qualea rosea* around *Vouacapoua Americana* trees in the `paracou16` point pattern. The solid, black curve is M . The dotted red curve is the average simulated value. The shaded area is the confidence envelope. $M = 1$ is expected if points are independently distributed. The risk level is 5%, 1000 simulations have been run. Distances are in meters.

p value of the test by the ratio of simulations whose distance is larger than that of the real point pattern. The test is performed by the `GoFtest` function whose argument is the envelope previously calculated (actually, the function uses the simulation values). Applied to the example of Paracou trees, the p value is:

```
R> GoFtest(Envelope)
```

```
[1] 0.273
```

3.5. Ktest

The *Ktest* has been developed by Lang and Marcon (Lang and Marcon 2013; Marcon *et al.* 2013). It does not rely on simulations and returns the p value to erroneously reject CSR given the values of K . It relies on the exact variance of K calculated with edge-effect corrections. It only works in a rectangular window.

The following example tests a 1.5-ha subset of `paracou16` (100 m by 150 m, origin at the South Western corner). It rejects CSR ($p = 0.0027$).

```
R> data("paracou16", package = "dbmss")
R> RectWindow <- owin(c(300, 400), c(0, 150))
R> X <- paracou16[RectWindow]
R> plot(X)
R> Ktest(X, seq(5, 50, 5))
```

```
[1] 0.002682576
```

4. Conclusion

We built this package to provide an easy-to-use toolbox for users of spatial statistics mainly in economic geography and ecology. We wrapped up some **spatstat** functions to allow using them similarly to our original functions to build a rather complete set of tools, including topographic, absolute and relative functions. The analysis is limited to testing a point pattern against an appropriate null hypothesis, according to the framework developed in the economic literature (Combes *et al.* 2008) but we believe **dbmss** is a useful extension of **spatstat** for researchers who are motivated by empirical results more than by the tools themselves, regardless of their scientific field. Full features for point pattern analysis can be found in **spatstat** for those who want to go further, including the simulation of many point processes as alternate null hypotheses and model fitting beyond exploratory statistics.

Future developments include the use of distance matrices as input of the distance-based functions to allow addressing road distance or geographic coordinates. We will also develop subsampling techniques to be able to manage huge datasets (several million points) whose distances cannot all be calculated in a reasonable time.

Acknowledgments

We thank Florent Bonneu who kindly allowed us to use his published fire emergency data.

This work has benefited from an “Investissement d’Avenir” grant managed by Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-25-01).

References

- Baddeley A, Møller J, Waagepetersen R (2000). “Non- And Semi-Parametric Estimation of Interaction in Inhomogeneous Point Patterns.” *Statistica Neerlandica*, **54**(3), 329–350. doi:10.1111/1467-9574.00144.
- Baddeley A, Turner R (2005). “**spatstat**: An R Package for Analyzing Spatial Point Patterns.” *Journal of Statistical Software*, **12**(6), 1–42. doi:10.18637/jss.v012.i06.
- Besag J (1977). “Comments on Ripley’s Paper.” *Journal of the Royal Statistical Society B*, **39**(2), 193–195.
- Bonneu F (2007). “Exploring and Modeling Fire Department Emergencies with a Spatio-Temporal Marked Point Process.” *Case Studies in Business, Industry and Government Statistics*, **1**(2), 139–152.
- Bonneu F, Thomas-Agnan C (2015). “Measuring and Testing Spatial Mass Concentration of Micro-Geographic Data.” *Spatial Economic Analysis*, **10**(3), 289–316. doi:10.1080/17421772.2015.1062124.
- Combes PP, Mayer T, Thisse JF (2008). *Economic Geography*. Princeton University Press, Princeton.

- Combes PP, Overman H (2004). “The Spatial Distribution of Economic Activities in the European Union.” In JV Henderson, JF Thisse (eds.), *Handbook of Urban and Regional Economics*, volume 4, chapter 64, pp. 2845–2909. Elsevier, Amsterdam.
- Diggle P (1983). *Statistical Analysis of Spatial Point Patterns*. Academic Press, London.
- Diggle P, Chetwynd A (1991). “Second-Order Analysis of Spatial Clustering for Inhomogeneous Populations.” *Biometrics*, **47**(3), 1155–1163. doi:10.2307/2532668.
- Duncan R (1991). “Competition and the Coexistence of Species in a Mixed Podocarp Stand.” *Journal of Ecology*, **79**(4), 1073–1084. doi:10.2307/2261099.
- Duranton G, Overman H (2005). “Testing for Localisation Using Micro-Geographic Data.” *Review of Economic Studies*, **72**(4), 1077–1106. doi:10.1111/0034-6527.00362.
- Duranton G, Overman H (2008). “Exploring the Detailed Location Patterns of UK Manufacturing Industries Using Microgeographic Data.” *Journal of Regional Science*, **48**(1), 213–243. doi:10.1111/j.1365-2966.2006.0547.x.
- Eddelbuettel D, François R (2011). “**Rcpp**: Seamless R and C++ Integration.” *Journal of Statistical Software*, **40**(8), 1–18. doi:10.18637/jss.v040.i08.
- Ellison G, Glaeser E (1997). “Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach.” *Journal of Political Economy*, **105**(5), 889–927. doi:10.1086/262098.
- Ellison G, Glaeser E, Kerr W (2010). “What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns.” *The American Economic Review*, **100**(3), 1195–1213. doi:10.1257/aer.100.3.1195.
- Fortin MJ, Dale M (2005). *Spatial Analysis. A Guide for Ecologists*. Cambridge University Press, Cambridge.
- Gini C (1912). *Variabilità e Mutabilità*, volume 3 of *Studi Economico-Giuridici dell’Università di Cagliari*. Università di Cagliari.
- Goreaud F, Péliissier R (2003). “Avoiding Misinterpretation of Biotic Interactions with the Intertype K_{12} -Function: Population Independence vs. Random Labelling Hypotheses.” *Journal of Vegetation Science*, **14**(5), 681–692.
- Gourlet-Fleury S, Guehl J, Laroussinie O (2004). *Ecology & Management of a Neotropical Rainforest. Lessons Drawn from Paracou, a Long-Term Experimental Research Site in French Guiana*. Elsevier, Paris, France.
- Haase P, Pugnaire F, Clark S, Incoll L (1997). “Spatial Pattern in Anthyllis Cytisoides Shrubland on Abandoned Land in Southeastern Spain.” *Journal of Vegetation Science*, **8**(5), 627–634. doi:10.2307/3237366.
- Hoover E (1936). “The Measurement of Industrial Localization.” *The Review of Economics and Statistics*, **18**(4), 162–171. doi:10.2307/1927875.
- Illian J, Penttinen A, Stoyan H, Stoyan D (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. John Wiley & Sons, Chichester, England. doi:10.1002/9780470725160.

- Kenkel N (1988). “Pattern of Self-Thinning in Jack Pine: Testing the Random Mortality Hypothesis.” *Ecology*, **69**(4), 1017–1024. doi:[10.2307/1941257](https://doi.org/10.2307/1941257).
- Lang G, Marcon E (2013). “Testing Randomness of Spatial Point Patterns with the Ripley Statistic.” *ESAIM: Probability and Statistics*, **17**, 767–788. doi:[10.1051/ps/2012027](https://doi.org/10.1051/ps/2012027).
- Loosmore N, Ford E (2006). “Statistical Inference Using the G or K Point Pattern Spatial Statistics.” *Ecology*, **87**(8), 1925–1931. doi:[10.1890/0012-9658\(2006\)87\[1925:siutgo\]2.0.co;2](https://doi.org/10.1890/0012-9658(2006)87[1925:siutgo]2.0.co;2).
- Marcon E, Lang G, Traissac S, Puech F (2015). *dbmss: Distance-Based Measures of Spatial Structures*. R package version 2.2.3, URL <http://CRAN.R-project.org/package=dbmss>.
- Marcon E, Puech F (2010). “Measures of the Geographic Concentration of Industries: Improving Distance-Based Methods.” *Journal of Economic Geography*, **10**(5), 745–762. doi:[10.1093/jeg/lbp056](https://doi.org/10.1093/jeg/lbp056).
- Marcon E, Puech F (2014). “A Typology of Distance-Based Measures of Spatial Concentration.” *Working Paper hal-00679993*, HAL SHS. Version 2.
- Marcon E, Puech F, Traissac S (2012). “Characterizing the Relative Spatial Structure of Point Patterns.” *International Journal of Ecology*, **2012**(Article ID 619281), 1–11. doi:[10.1155/2012/619281](https://doi.org/10.1155/2012/619281).
- Marcon E, Traissac S, Lang G (2013). “A Statistical Test for Ripley’s Function Rejection of Poisson Null Hypothesis.” *ISRN Ecology*, **2013**(Article ID 753475), 1–9. doi:[10.1155/2013/753475](https://doi.org/10.1155/2013/753475).
- Openshaw S, Taylor P (1979). “A Million or so Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem.” In N Wrigley (ed.), *Statistical Applications in the Spatial Sciences*, pp. 127–144. Pion, London.
- Penttinen A (2006). “Statistics for Marked Point Patterns.” In *The Yearbook of the Finnish Statistical Society*, pp. 70–91. The Finnish Statistical Society, Helsinki.
- Penttinen A, Stoyan D, Henttonen H (1992). “Marked Point Processes in Forest Statistics.” *Forest Science*, **38**(4), 806–824.
- Picone G, Ridley D, Zandbergen P (2009). “Distance Decreases with Differentiation: Strategic Agglomeration by Retailers.” *International Journal of Industrial Organization*, **27**(3), 463–473. doi:[10.1016/j.ijindorg.2008.11.007](https://doi.org/10.1016/j.ijindorg.2008.11.007).
- R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Ripley B (1976). “The Second-Order Analysis of Stationary Point Processes.” *Journal of Applied Probability*, **13**(2), 255–266. doi:[10.2307/3212829](https://doi.org/10.2307/3212829).
- Ripley B (1977). “Modelling Spatial Patterns.” *Journal of the Royal Statistical Society B*, **39**(2), 172–212.
- Ripley B (1988). *Statistical Inference for Spatial Processes*. Cambridge University Press. doi:[10.1017/cbo9780511624131](https://doi.org/10.1017/cbo9780511624131).

Scholl T, Brenner T (2013). “Optimizing Distance-Based Methods for Big Data Analysis.”
Technical Report 2013-09, Philipps University Marburg.

Affiliation:

Eric Marcon, Stéphane Traissac
AgroParisTech, UMR EcoFoG
Campus agronomique, BP 316
97310 Kourou, French Guiana
E-mail: Eric.Marcon@ecofog.gf, Stephane.Traissac@ecofog.gf

Florence Puech
RITM
Univ. Paris-Sud, Université Paris-Saclay
92330, Sceaux, France
E-mail: Florence.Puech@u-psud.fr

Gabriel Lang
AgroParisTech, INRA, UMR 518 Math. Info. Appli.
16 rue Claude Bernard
75005 Paris, France
E-mail: Gabriel.Lang@agroparistech.fr

APPENDIX C

Landscape patterns influence communities of medium- to large-bodied vertebrate in undisturbed terra firme forests of French Guiana

Richard-Hansen C., Jaouen G., Denis T., Brunaux O., Marcon E.
et Guitet, S. (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.

Landscape patterns influence communities of medium- to large-bodied vertebrates in undisturbed terra firme forests of French Guiana

Cécile Richard-Hansen¹, Gaëlle Jaouen, Thomas Denis, Olivier Brunaux, Eric Marcon and Stéphane Guitet

ONCFS (Office National de la Chasse et de la Faune Sauvage), EcoFoG, ONCFS Campus agronomique, BP 316, 97379 Kourou cedex French Guiana, France
(Received 27 May 2014; revised 13 May 2015; accepted 14 May 2015)

Abstract: Whereas broad-scale Amazonian forest types have been shown to influence the structure of the communities of medium- to large-bodied vertebrates, their natural heterogeneity at smaller scale or within the terra firme forests remains poorly described and understood. Diversity indices of such communities and the relative abundance of the 21 most commonly observed species were compared from standardized line-transect data across 25 study sites distributed in undisturbed forests in French Guiana. We first assessed the relevance of a forest typology based on geomorphological landscapes to explain the observed heterogeneity. As previously found for tree beta-diversity patterns, this new typology proved to be a non-negligible factor underlying the beta diversity of the communities of medium- to large-bodied vertebrates in French Guianan terra firme forests. Although the species studied are almost ubiquitous across the region, they exhibited habitat preferences through significant variation in abundance and in their association index with the different landscape types. As terra firme forests represent more than 90% of the Amazon basin, characterizing their heterogeneity – including faunal communities – is a major challenge in neotropical forest ecology.

Key Words: animal communities, diversity, environmental heterogeneity, French Guiana, landscape ecology, species-habitat association

INTRODUCTION

Although they are often iconic and well known to forest dwellers, precise information is lacking on the distribution and ecological preferences of most vertebrate species in neotropical forests. In central Amazonia, previous studies revealed that the structure of communities of medium- to large-bodied vertebrates varies according to the two major forest types: seasonally inundated forests (várzea) and terra firme forests (Haugaasen & Peres 2005a, b, 2008). According to these studies, seasonally inundated forests appeared to be less diverse but carry higher densities and biomass of primates compared with the well-drained uplands (terra firme). However, at finer geographic scale (i.e. within each category), the inherent heterogeneity of these faunal communities remains poorly documented, with the exception of some mainly descriptive studies focused on primate communities (Buchanan-Smith *et al.* 2000, Freese *et al.* 1982, Heymann *et al.* 2002, Sussman

& Phillips-Conroy 1995), and a more recent and detailed analysis in western Amazonia (Palminteri *et al.* 2011). According to these authors, although hunting pressure and/or human impact are often the best predictors of primate community structure, biogeographic and environmental factors also drive community structure. The main descriptive parameter for forest types was still flooded vs. unflooded areas, but this parameter was refined as gradient. The same authors also pointed out that the drivers may be more a combination of environmental factors rather than any one factor.

In French Guiana, the whole territory was until recently considered as apparently homogeneous terra firme forest. However, recent research demonstrated the existence of several types of terra firme forest across Amazonia (Anderson *et al.* 2009) or within the Guiana Shield (Fayad *et al.* 2014, Gond *et al.* 2011). Even in a regional context where environmental gradients are quite weak, as is the case of the Guiana Shield, the hyper-diversified tropical rain forest shows a significant gradient of tree composition and strong subregional patterns (Guitet *et al.* 2015). The best factor identified

¹ Corresponding author. Email: cecile.richard-hansen@ecofog.gf

to explain these broad-scale patterns in the floristic and structural diversity of the terra firme rain forest was the geomorphological landscape type (Guitet *et al.* 2013). In the Amazon region, other studies have also linked geomorphological landscape type with forest physiognomy (Anderson *et al.* 2009) and/or biological diversity or community structure (Deichmann *et al.* 2011, Figueiredo *et al.* 2014, Sombroek 2000). Such an integrative variable is thus a good candidate to combine local ecological conditions and to approximate forest structure and composition, but its influence on vertebrate communities has never been tested to date.

In French Guiana, abundance data on medium- to large-bodied vertebrates revealed strong differences across undisturbed forest sites (Richard-Hansen 2006). This study scale is below that typically used for turnover in most Amazonian large-vertebrate species, thereby focusing the analysis of community heterogeneity on niche differentiation and community structure (abundances) rather than dispersal limitation and species replacement (http://www.iucnredlist.org/mammals/data_types; Patterson *et al.* 2005). We therefore hypothesized that environmental parameters and forest types can partially explain this heterogeneity in French Guiana, as documented in other forested environments of Amazonia. The influence of the landscape type on the forest structure has been proved (Guitet *et al.* 2015), and the aim of the present study was to assess the relevance of this classification as an underlying driver of the distribution patterns of the communities of medium- to large-bodied vertebrates, with respect to its ability to describe the combination of local environmental factors.

METHODS

Study area: French Guiana

French Guiana covers about 85 000 km² in the east of the Guiana shield between Suriname and the Brazilian state of Amapá (4°N, 53°W). Altitude generally ranges between 0 and 200 m asl (mean 140 m asl) with few mountain peaks exceeding 800 m. The climate is equatorial with annual rainfall ranging from 3600 mm in the north-east to 2000 mm in the south and the west, with a mean annual temperature of about 26°C. The number of consecutive months with less than 100 mm precipitation (dry season) ranges from two in the north to three in the south with high interannual variation (Sombroek 2001). Savannas and mangroves occur only in the coastal sedimentary plain, while the evergreen rain forest covers more than 90% of French Guiana (<http://www.fao.org>, Guitet *et al.* 2015). Natural habitats show slight variability and high species diversity, with a complex tree community and

often more than 150–200 species ha⁻¹ (Sabatier *et al.* 1997).

Overall human density is below 3 inhabitants km⁻², and 75% of the population is restricted to the five major towns, with the remaining population living in a few small villages and settlements (<http://www.insee.fr>) mainly along the two main rivers that form the borders with Suriname and Brazil (Figure 1). A National Park covers 34 million ha, 20 million ha of which comprise the core area where only the resident population is allowed to hunt for subsistence. Roads are limited to a less than 50 km-wide northern coastal strip, while the rest of the country is accessible only by boat or by small airplane from Cayenne to a few main settlements. Timber harvesting and agriculture are contained in subcoastal areas, covering currently around 2 million ha, close to the biggest towns and main roads. Consequently, most of the hunting pressure is applied on the northern coastal strip, along main rivers and streams and around the scattered villages.

Animal abundance

Standardized line transect surveys (Buckland *et al.* 1993) were conducted at 25 different study sites across French Guiana. The study sites are very isolated and most can be accessed only by helicopter or several days walking, so we consider that there was no strong or recent hunting pressure, even by autochthonous populations. The same field design was implemented at each site, consisting of four 3-km long trails radiating from a central place (campsite). This design makes it possible to account for small local variations in the environment, including topographic features or scattered resources (fruiting trees), within a single global abundance index, characterizing a similar area for each site surveyed. Transects were walked at less than 1 km h⁻¹ every morning (7h00–11h00) and afternoon (14h30–18h00) by only one observer per trail, systematically alternating transects on consecutive days to avoid observer bias. All encounters with focal species and their localization on the trail were systematically recorded and the perpendicular distance between the animal and the transect was measured to the nearest metre with a laser range finder. Transects were surveyed an average \pm SD of 13.7 ± 1.9 times each, during an 8-d field session. Total survey effort per site ranged from 140 to 210 km (average \pm SD = 163 ± 17.7 km), with a cumulative survey effort of 4073 km across 99 individual transects at 25 sites. The minimum effort required for reliable estimates of abundance and richness in this environment was estimated at 100 km (de Thoisy *et al.* 2008). The surveys were all conducted during the dry season (September–December) to avoid interference with potential seasonal variation.

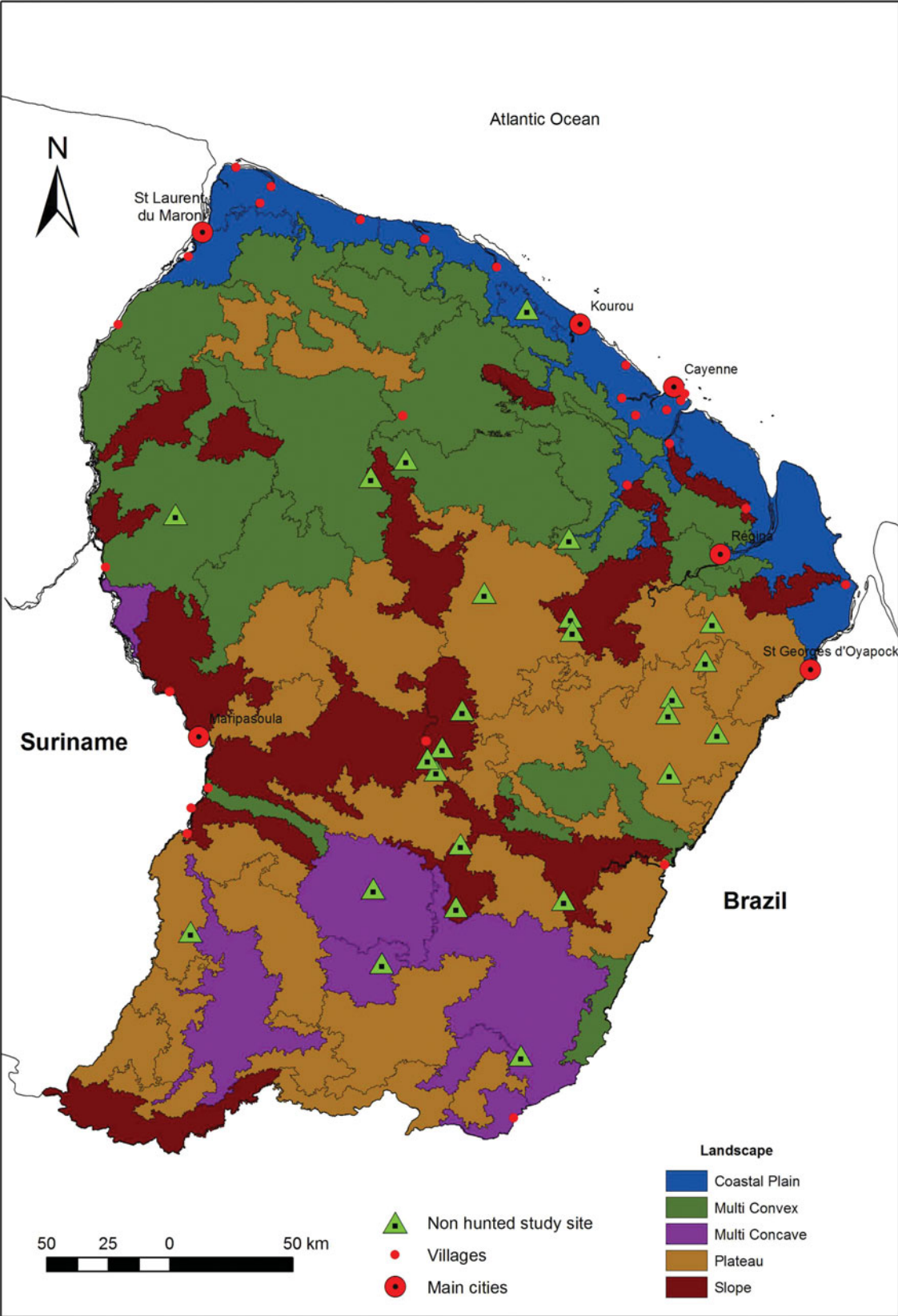


Figure 1. Location of 25 undisturbed study sites in French Guiana, and their distribution within the five landscape types, characterized from a geomorphological analysis based on a digital elevation model.

Thirty-seven species were recorded (mammals weighing > 0.5 kg and large terrestrial birds), and diversity estimates were based on this pool of species. For abundance comparisons, we focused on the 21 most frequent species, including primates, ungulates, caviomorph rodents, large terrestrial birds (cracids, tinamous, trumpeters, guans) and tortoises, for which a reliable index of abundance could be calculated. Tinamidae species (*Crypturellus* spp. and *Tinamus major*) were grouped because many observations lacked clear identification.

Environmental characteristics of the study sites

The environment was characterized by the geomorphological landscape type defined by Guitet *et al.* (2013). This typology was developed from a multi-scale geomorphological diversity analysis based on a digital elevation model computed from a fine Shuttle Radar Topography Mission images (SRTM, 30 m resolution). Variations in micro-relief defined 12 landform types whose spatial distribution drew 82 different patches classified in 10 landscape types that can be grouped under five main categories: (1) coastal plain, (2) plateau, (3) mountain, (4) multi-convex and (5) multi-concave landscape. The joint-valleys are considered with the multi-convex category (Guitet *et al.* 2013). Recent results showed that the structure and composition of the forest is clearly influenced by these landscape types (Guitet *et al.* 2015). Coastal plains ($N = 2$ sites in this study), located in the northern part of French Guiana, are lowland forests on Quaternary marine sediments. They are characterized by a relatively low canopy (28 m in height), high density of small trees, and relative high abundance of Clusiaceae, Caesalpinioideae and Lecythidaceae. The plateau category ($N = 8$ sites) includes several types of relatively flat relief of moderate elevation dissected to a varying extent by rivers, exclusively covered by well-drained ferralsols with very localized hydromorphic soils. Burseraceae, Mimosoideae and Caesalpinioideae are dominant tree families, but high abundances of palms are also found. Small inselbergs are also frequent. Sloping areas ($N = 9$ sites), locally called mountains despite their modest altitudes (<840 m asl), are characterized by higher relief with many slopes. The dominant forest type is characterized by a high canopy (35–40 m), high basal-area values and the abundance of very large trees, with high diversity and much more infrequent families such as Vochysiaceae, Malvaceae and Annonaceae being more abundant compared with other forest types. The multi-convex landscape ($N = 3$ sites) is dominated by more or less regular hills with a dense hydrographic network, and dominance of Lecythidaceae and Caesalpinioideae. The soil cover is more diversified mixing clayic ferralsols with more sandy or loamy soils acrisols. The multi-

concave landscape ($N = 3$ sites) corresponds to large peneplains in the south, characterized by very flat relief, covered by leached and partially inundated soils during the wet season, although the water levels never rise as high as in the Amazonian várzea forests. The canopy is low (30 m high) and discontinuous, and vegetation is characterized by the dominance of Burseraceae, Mimosoideae and Myristicaceae with relatively few large trees and dense understorey with few palms. Finding undisturbed sites was harder in some landscapes types because of proximity of human settlements (coastal plain) or difficult access (multi-concave landscape), thus explaining the unbalanced sampling.

Six other broad-scale environmental variables were also tested: the biogeographic region (Paget 1999), the vegetation type based on remotely sensed landscape classes (RSLC) from the VEGETATION sensor of the SPOT-4 satellite (Gond *et al.* 2011), annual rainfall (Meteo France, unpubl. data), the proportion of hydromorphic soils, the mean slope and the mean differences in altitude for the area. The last three variables were extracted from a digital elevation model computed from fine-resolution Shuttle Radar Topography Mission images (SRTM, 30 m resolution). All these data were computed for a circle with a 4-km radius encompassing the survey transects.

Data analysis

An index of abundance of groups encountered per 10 km walked (elsewhere referred to as encounter rate, sensu Buckland *et al.* 1993) was calculated to control for overall differences in sampling effort (Peres 1997). Perpendicular distances (PD) were recorded, but not enough observations of each species were made at each site to correctly estimate the detection function for all of them and hence to calculate densities. However, we assumed that this index of abundance (hereafter, abundance) of different species could be compared between sites because, except for agouti (*Dasyprocta leporina*), the distributions of the distances of observation were not statistically different (ANOVA on $\log(PD)$, $P > 0.5$).

The dissimilarity between faunal communities in different landscape types was first tested by permutational multivariate analysis of variance on the site \times species tables of raw counts of the 21 most common species, using χ^2 distance matrices. The Adonis test was selected because it is more robust and less sensitive to dispersion effects (within-group variation) than some of its alternatives (ANOSIM, etc.) (Anderson 2001). We also tested the pertinence of the landscape typology as a potential explanatory variable in this variation using a between-class correspondence analysis (BCA), which is a particular case of correspondence analysis

on instrumental variable (i.e. canonical correspondence analysis) with only one categorical variable (Dolédec & Chessel 1989, Dray & Dufour 2007, Dray *et al.* 2012, Péliissier *et al.* 2003). A correspondence analysis was first performed on the site \times species tables of raw counts of the 21 most common species, and between-class analysis was then performed on the results (site coordinates), with the landscape type of each site as categorical variable. From this analysis, the between-class inertia is the proportion of total inertia of the table explained by the landscape variable, while the within-class inertia is the proportion of total inertia not explained by this variable. The statistical significance of this portion of initial variance captured by this instrumental variable was tested with Monte Carlo row permutation tests against the null hypothesis of no relation between species assemblage and landscape type (Couteron *et al.* 2003). The same analysis was made for the six other variables. These analyses were performed with the ade4 (Dray & Dufour 2007) and vegan-packages in R.

Diversity of communities and meta-communities. Crude richness of a study site is the number of species recorded during the survey, within the fixed maximum of 37 focal species. We calculated the diversity profile for each site community, and for each meta-community created by pooling the sites belonging to the same landscape type. The diversity profile plots the value of Hill numbers (Hill 1973) against the order of diversity q (Kindt *et al.* 2006, Patil & Taillie 1982). Hill numbers are the transformation of Tsallis entropy values into an effective number of species, i.e. the number of species of equal frequency that would yield the same diversity as real data (Jost 2006). Tsallis entropy qH (Tsallis 1988) generalizes the classical indices of diversity in a parameterized measure, where the choice of the parameter gives more or less importance to rare species: 0H is the number of species minus 1, 1H is Shannon entropy (Shannon 1948) and 2H is Simpson index (Simpson 1949). All values of diversity were corrected for estimation bias (Marcon *et al.* 2014): the Chao & Shen (2003) estimator applies to small values of q , that of Grassberger (1988) to high values.

We tested the relevance of landscape type as a diversity predictor. We first pooled sites within one landscape type, and then pooled all landscape types together, allowing the measurement of β diversity across both levels (Marcon *et al.* 2012). We tested the observed ratio of β diversity between landscapes over β diversity within landscapes against its distribution under the null hypothesis of independence between sites and landscapes: we shuffled sites among landscapes and calculated the ratio of β diversity 1000 times. A result of the test was considered significant if the actual ratio was in the last five percentiles of the distribution of the simulated values, showing

that β diversity between landscapes was higher (relative to β diversity within landscapes) than under the null hypothesis. An alternative, more intuitive test would address the ratio of β entropies. Although it is more similar to a classical analysis of variance (since the total β entropy is the sum of within and between landscape β entropies), it suffers from the drawbacks discussed by Jost (2008). β entropy is constrained by the value of α entropy, thereby invalidating the test. Diversity estimates and comparison were made with R package entropart (Marcon & Hérault 2015).

Finally, we looked for species-landscape associations using the set of indices initially proposed by Dufrêne & Legendre (1997) to study species assemblages and habitat types. Our aim here focused on the relative abundance of the 21 most common species occurring in most sites rather than that of rare or indicative species. Following De Caceres & Legendre (2009), we thus selected the point-biserial correlation coefficient (r_{pb}), which is the Pearson correlation computed between a quantitative vector (i.e. the vector containing the species abundance values at the various sites) and a binary vector (i.e. the vector of site membership values) rather than the better known indicator value index (IndVal). To account for the unequal number of sites in the different landscape types, we used the corrected group-equalized index (r_{pb}^g), (De Caceres & Legendre 2009). The significance of these associations was tested by Monte Carlo permutation tests. We also tested the difference in species abundance in sites belonging to one particular landscape compared with sites located in different landscapes by permutation tests, after Sidak's correction for multiple testing. We then considered whether combining basic landscape types would better match species preferences (De Caceres *et al.* 2010). It may also happen that a particular site group has no indicator or associated species even if its sites have a community composition that is clearly distinct from the sites of other site groups (De Caceres *et al.* 2012). In these cases, the joint occurrence of two or more species has a higher positive predictive value for the site group than the two species taken independently, so we also explored correlation values for combinations of species (De Caceres *et al.* 2012). All analyses mentioned in this section were computed with the R package indicpecies.

RESULTS

Abundances of common species varied greatly across French Guiana, even in areas with no strong or recent human influence of hunting, logging or gold mining (Table 1). Nine out of 21 species were present in each of the 25 sites, 15 were present in at least 90% of sites (more than 21) and 12 showed a null abundance at least once. These 12 species may be totally absent from the site

Table 1. Index of abundance (number of observations per 10 km) recorded for 21 species in 25 undisturbed sites in French Guiana, and according to the different landscape types (MCV: multi-concave; MCX: multi-convex; PLA: plateau; PLN: coastal plains; SLO: sloping areas). Abundance significantly higher or lower compared with all other sites: * $P \leq 0.05$; abundance significantly higher or lower compared with other landscapes: † $P \leq 0.05$ (permutation test, corrected P-value for multiple comparisons).

	General mean \pm SD	Landscape				
		MCV	MCX	PLA	PLN	SLO
Primates						
<i>Alouatta macconnelli</i> (Linnaeus, 1976)	0.56 \pm 0.30	0.71	0.42	0.61	0.45	0.52
<i>Ateles paniscus</i> (Linnaeus, 1758)	1.19 \pm 0.76	0.81	1.31	0.96	0.36	†1.66
<i>Cebus apella</i> (Linnaeus, 1758)	0.85 \pm 0.46	0.96	1.04	0.61	1.69*	0.79
<i>Cebus olivaceus</i> (Schomburgk, 1848)	0.24 \pm 0.24	0.19	†0.45	0.21	†0.00*	0.25
<i>Pithecia pithecia</i> (Linnaeus, 1766)	0.06 \pm 0.08	0.16	0.00	0.06	0.07	0.04
<i>Saguinus midas</i> (Linnaeus, 1758)	0.41 \pm 0.31	0.53	0.55	0.32	0.92	0.30
<i>Saimiri sciureus</i> (Linnaeus, 1758)	0.04 \pm 0.09	0.15	0.00	0.00	0.10	0.03
Ungulates						
<i>Mazama americana</i> (Erxleben, 1777)	0.43 \pm 0.29	0.33	0.32	0.53	0.30	0.43
<i>Mazama nemorivaga</i> (F.Cuvier, 1817)	0.44 \pm 0.29	0.39	0.34	0.51	0.59	0.39
<i>Pecari tajacu</i> (Linné, 1758)	0.29 \pm 0.20	0.34	0.07	0.41	0.41	0.22
<i>Tayassu pecari</i> (Link, 1795)	0.03 \pm 0.06	0.02	0.00	0.02	0.00	0.05
<i>Tapirus terrestris</i> (Linnaeus, 1758)	0.05 \pm 0.07	0.00	0.06	0.04	0.07	0.06
Rodents						
<i>Dasyprocta leporina</i> (Linné, 1758)	1.48 \pm 0.75	1.66	2.26	1.27	2.50	††1.11
<i>Myoprocta acouchy</i> (Erxleben, 1777)	0.57 \pm 0.33	0.72	0.50	0.52	0.65	0.57
Birds						
<i>Crax alector</i> (Linnaeus, 1776)	0.57 \pm 0.33	0.33	0.48	0.60	0.49	0.66
<i>Odontophorus gujanensis</i> (J.F. Gmelin, 1789)	0.31 \pm 0.31	0.54	0.04	0.42	0.00	0.30
<i>Ortalis motmot</i> (Linnaeus, 1766)	0.02 \pm 0.07	0.13	0.02	0.01	0.00	0.00
<i>Penelope marail</i> (S. Müller, 1776)	0.33 \pm 0.17	†0.59*	0.11*	0.32	0.42	0.31
<i>Psophia crepitans</i> (Linnaeus, 1758)	1.05 \pm 0.66	1.44	0.87	0.97	1.29	1.01
Tinamidae	2.20 \pm 0.89	3.33*	2.11	2.12	2.29	1.92
Reptile						
<i>Chelonoidis denticulata</i> (Linnaeus, 1766)	0.19 \pm 0.17	0.45*	0.20	0.12	0.24	0.16

or present in densities that were too low to be detected with our sampling protocol.

correlation index. All the other environmental variables tested explained a smaller proportion of total inertia with both analyses (Table 2).

Structure of animal communities in various landscapes

The permutational multivariate analysis of variance (Adonis test) on animal communities according to the various environmental variables showed that the proportion of variance explained by the landscape variable was the highest ($R^2 = 0.24$), and significant according to permutation test (Table 2). The between-class analysis also revealed that 24.3% of the total inertia of the data was explained by the instrumental variable of landscape typology. The Monte Carlo row permutation test for this unique environmental variable was significant ($P = 0.007$). Moreover, the graphic representation of the results of this between-class analysis showed that multi-convex and multi-concave landscapes presented the most distinct vertebrate assemblages, while plateau and mountain communities were less clearly distinguished (Figure 2). The main structuring species are shown on the graph, and their affinities with the various landscapes were tested subsequently with the

Diversity of landscape communities

For each individual site community, Simpson diversity varied from eight to 16 effective species, and richness ($q = 0$) estimated with Chao and Shen's bias correction (approximately equal to the Jackknife 1 or Chao 1 estimators) was between 18 and 31 (Table 3). With a few exceptions, the highest richness values corresponded to sites in multi-concave landscapes and the lowest richness values to sites in multi-convex ones, with values for plains and mountainous sites between the two. Considering Simpson diversity, however, mountain sites were among the lowest values. The beta diversity between landscape meta-communities was significantly different ($P < 0.05$) from the β diversity between random meta-communities for q values of between 0.2 and 1.9. Common species were more evenly distributed in the various landscapes, and were present everywhere: less common species made the difference between landscapes; ignoring them (choosing

Table 2. Analysis of variance between the communities of medium- to large-bodied vertebrates in 25 study sites in French Guiana, according to seven environmental variables. Partial R-square from permutational multivariate analysis of variance (Adonis test), tested with permutation test with pseudo-F. Between-class inertia from a principal component analysis with respect to the instrumental variable (PCAIV) performed on the coordinates of a correspondence analysis, tested by Monte Carlo test. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

	Landscape	Vegetation type	Biogeography	% Hydromorphic soil	Mean Slope	Difference in altitude	Annual rainfall
Partial R^2 (Adonis test)	0.24**	0.17	0.15**	0.16***	0.14**	0.13*	0.12*
% between-class inertia	0.24**	0.20*	0.15***	0.15**	0.14*	0.13*	0.13*

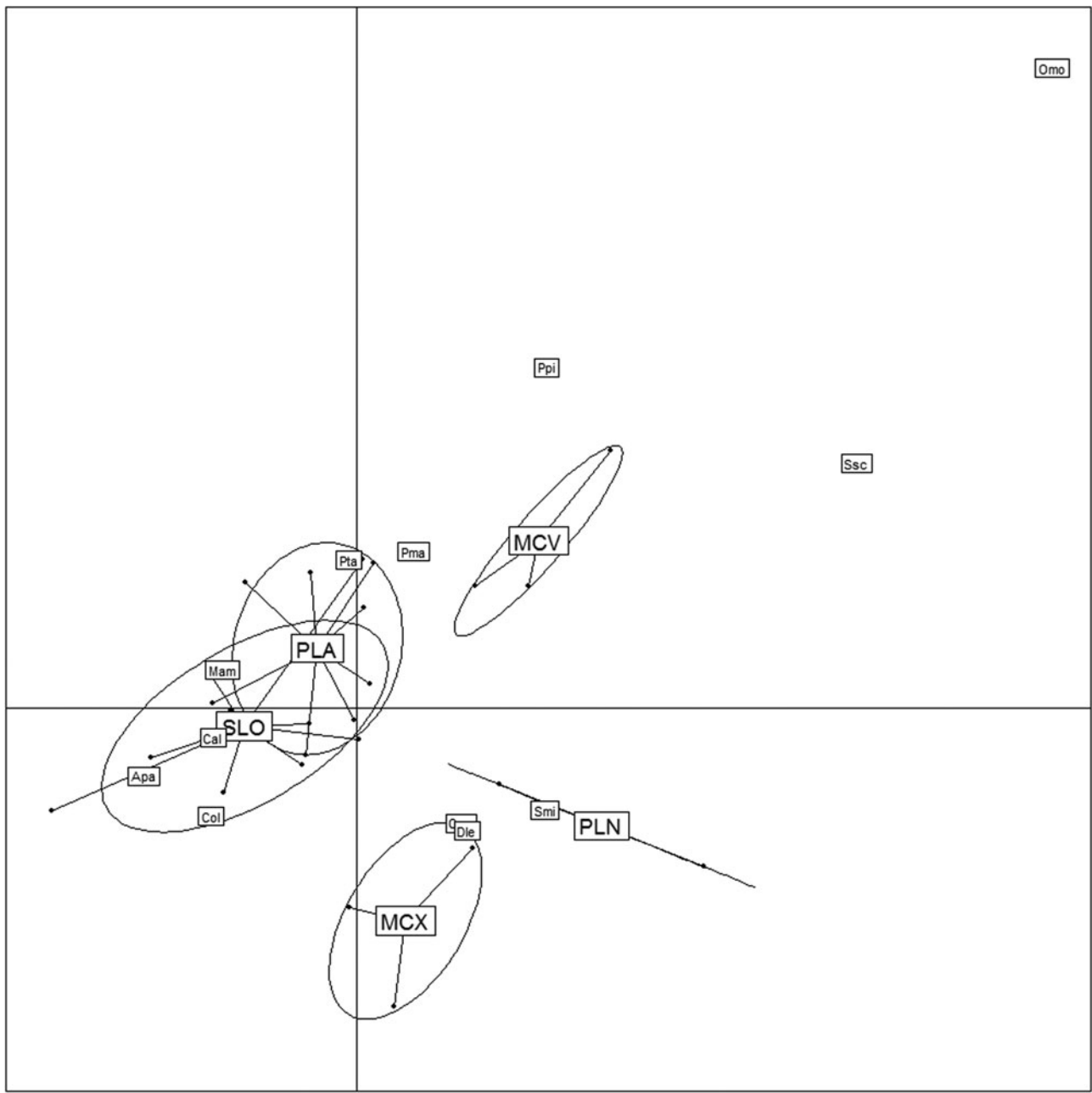


Figure 2. Between-class analysis of the communities of medium- to large-bodied vertebrates in 25 study sites in five landscapes types in terra firme forests of French Guiana. The ellipses graphically sum up each landscape type (MCX = multi-convex; MCV = multi-concave; PLA = plateau; SLO = sloping areas; PLN = coastal plain) by covering 67% of the sites belonging to the landscape type; the centre of each ellipse is the centre of gravity of these sites. Main structuring species are indicated (Omo: *Ortalis motmot*, Ssc: *Saimiri sciureus*, Ppi: *Pithecia pithecia*, Pma: *Penelope marail*, Pla: *Pecari tajacu*, Mam: *Mazama americana*, Cal: *Crax alector*, Apa: *Ateles paniscus*, Col: *Cebus olivaceus*, Smi: *Saguinus midas*, Dle: *Dasyprocta leporina*).

Table 3. Main diversity indices, corresponding to three entropy values (q), for the medium- to large-bodied vertebrate communities in 25 study sites in terra firme forests of French Guiana, according to their landscape type. Values correspond to effective number of species. Landscape types : MCV = multi-concave, three sites; MCX = multi-convex, three sites; PLA = plateau, eight sites; PLN = coastal plains, two sites; SLO = sloping areas, nine sites.

Site	Diversity index		
	Richness ($q = 0$)	Shannon ($q = 1$)	Simpson ($q = 2$)
MCV.1	31.3	18.1	14.1
MCV.2	21.7	13.8	11.0
MCV.3	26.3	18.2	15.3
MCX.1	22.4	13.6	10.1
MCX.2	22.6	11.9	09.0
MCX.3	23.5	14.1	11.1
PLA.1	20.9	14.9	12.3
PLA.2	25.7	16.5	12.7
PLA.3	22.6	14.2	11.3
PLA.4	23.7	15.5	12.3
PLA.5	22.0	15.0	13.1
PLA.6	24.0	17.6	15.4
PLA.7	26.6	19.0	16.3
PLA.8	18.9	13.8	10.8
PLN.1	23.3	12.8	08.6
PLN.2	17.8	14.6	12.2
SLO.1	23.3	12.2	07.9
SLO.2	19.7	13.9	11.5
SLO.3	23.8	13.8	09.3
SLO.4	23.7	15.2	12.4
SLO.5	23.3	13.5	09.5
SLO.6	24.6	16.2	13.7
SLO.7	22.9	16.9	14.6
SLO.8	20.2	14.7	13.0
SLO.9	23.6	15.1	11.6

high values of q) made the test inconclusive. For small values of q , a lack of power of the test was involved: bias correction was more important, and so was the variance of the estimator of diversity.

The diversity profiles of the five meta-communities (γ diversity) corresponding to the five landscape types differed, whatever the order of entropy considered ($0 \leq q \leq 2$, i.e. from the number of species to Simpson diversity, Figure 3). The most diverse meta-community is encountered in the multi-concave landscape, despite the small sample size in this category, and the least diverse in the plain and multi-convex landscapes. Plateaux and mountainous areas were intermediate in terms of diversity, the steeper-sloped areas were more diverse than plateaux when rare species were considered ($q = 0$), and the reverse when only common species were considered ($q = 2$).

Characterization of landscape communities

The multi-concave landscape was positively associated with the largest number of species (Table 4). Six

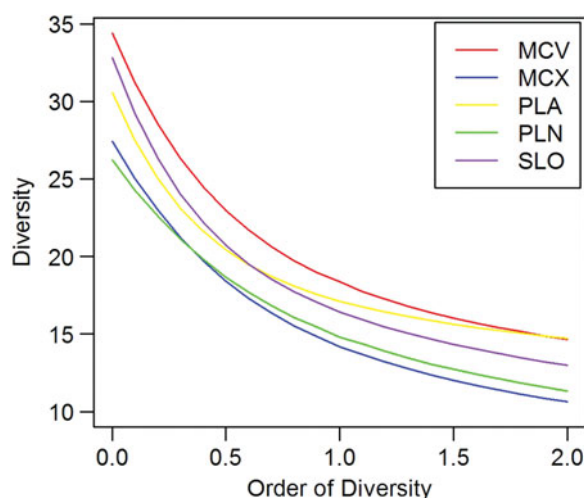


Figure 3. Gamma diversity profiles of the communities of medium- to large-bodied vertebrates in the five landscape types (MCV = multi-concave; MCX = multi-convex; PLA = plateau; PLN = coastal plain; SLO = sloping areas), as estimated by diurnal line-transects conducted in 25 non-disturbed study sites in terra firme forest in French Guiana.

species had a correlation coefficient $r_{pb}^g \geq 0.5$ for this landscape category. *Penelope marail*, *Ortalis motmot*, Tinamidae and the tortoise *Chelonoidis denticulata* were the most characteristic species, and *Saimiri sciureus* and *Pithecia pithecia* were the most typically associated primates. Moreover, despite lower scores and no statistical significance, four more species had their maximum correlation coefficient in multi-concave landscapes (*Alouatta macconnelli*, *Psophia crepitans*, *Odontophorus guyanensis* and *Myoprocta acouchi*). These results on association tendencies between species and landscapes are confirmed by comparisons of abundance. The abundance of *S. sciureus*, *O. guyanensis*, *O. motmot*, *P. marail*, *C. denticulata* and tinamidae were significantly higher in multi-concave landscapes than in other landscapes and/or other sites combined (Table 1). In contrast, two species had negative r_{pb}^g in these areas: *Tapirus terrestris* and *Crax alector* ($r_{pb}^g = -0.4$ and -0.3 respectively) (Table 4). Finally, two of the three top-ranked sites in terms of crude richness were also located in a multi-concave landscape, and they also belonged to the three top-ranked sites regarding total abundance (total abundance, all species combined).

Cebus apella was clearly associated with coastal plains ($r_{pb}^g = 0.7$, $P < 0.05$, Table 4). The abundance of this species was significantly higher there than at all the other sites combined (Table 1) ($P < 0.05$). *Saguinus midas* also reached its maximum levels in this plain landscape. In contrast, *Ateles paniscus* and *Cebus olivaceus* had their lowest and negative coefficient there ($r_{pb}^g = -0.5$), and the abundance of *C. olivaceus* was significantly lower than in other landscape types.

Table 4. Association of 21 medium- to large-bodied vertebrate species with five landscape types in French Guianan pristine rainforest, as estimated by point-biserial correlation coefficient, corrected for unequal sampling in different landscapes (r_{pb}^g). MCV: Multi-concave; MCX: Multi-convex; PLA: Plateau; PLN: Coastal plains; SLO: sloping area. Monte Carlo Permutation test: *: $P \leq 0.05$; **: $P \leq 0.01$.

	Landscape				
	MCV	MCX	PLA	PLN	SLO
Primates					
<i>Alouatta macconnelli</i>	0.3	−0.2	0.1	−0.2	0
<i>Ateles paniscus</i>	−0.2	0.2	0	−0.5	0.5
<i>Cebus apella</i>	−0.1	0	−0.4	0.7**	−0.2
<i>Cebus olivaceus</i>	−0.1	0.5	0	−0.5	0.1
<i>Pithecia pithecia</i>	0.5	−0.3	0	0	−0.1
<i>Saguinus midas</i>	0	0	−0.3	0.5	−0.3
<i>Saimiri sciureus</i>	0.5	−0.3	−0.3	0.2	−0.1
Ungulates					
<i>Mazama americana</i>	−0.1	−0.1	0.3	−0.2	0.1
<i>Mazama nemorivaga</i>	−0.1	−0.2	0.1	0.3	−0.1
<i>Pecari tajacu</i>	0.1	−0.6	0.3	0.3	−0.2
<i>Tayassu pecari</i>	0	−0.2	0	−0.2	0.4
<i>Tapirus terrestris</i>	−0.4	0.1	0	0.2	0.1
Rodents					
<i>Dasyprocta leporina</i>	−0.1	0.3	−0.3	0.4	−0.4
<i>Myoprocta acouchy</i>	0.2	−0.1	−0.1	0.1	0
Birds					
<i>Crax alector</i>	−0.3	−0.1	0.2	−0.1	0.3
<i>Odontophorus guyanensis</i>	0.4	−0.3	0.2	−0.4	0.1
<i>Ortalis motmot</i>	0.6	−0.1	−0.1	−0.2	−0.2
<i>Penelope marail</i>	0.6**	−0.6	−0.1	0.2	−0.1
<i>Psophia crepitans</i>	0.3	−0.2	−0.1	0.1	−0.1
Tinamidae	0.5	−0.1	−0.1	0	−0.2
Reptiles					
<i>Chelonoidis denticulata</i>	0.6	−0.1	−0.3	0	−0.2

The associations between all species and the multi-convex, mountainous or plateau landscapes were all weaker ($r_{pb}^g \leq 0.5$), and none was statistically significant. *Cebus olivaceus* was the only species showing some association with multi-convex areas ($r_{pb}^g = 0.5$) and a higher abundance than in other landscapes, while nine species showed a negative association with this landscape, among which most conspicuously *Pecari tajacu* and *Penelope marail* ($r_{pb}^g = -0.6$) (Table 1). *Ateles paniscus* tended to show a maximum association with the mountainous landscape ($r_{pb}^g = 0.5$; abundance significantly higher than in other landscapes and other sites ($P < 0.05$)), whereas *Dasyprocta leporina* and the small primate *Saguinus midas* showed their minimum and negative values in this landscape type (Table 1). The abundance of *D. leporina* was significantly lower in mountainous landscapes than in other landscapes ($P < 0.05$) (Table 1). *Mazama americana* was the species most associated with plateaux ($r_{pb}^g = 0.3$) and *Cebus apella* the least ($r_{pb}^g = -0.4$, Table 4).

Another analysis considered if combining landscapes matched species preferences better. Whereas several species remained more strongly associated with a single

landscape type, some species turned out to be more strongly associated with a combination of landscapes. For example, *Penelope marail* appeared to be associated with the combination of smoothed landscapes, i.e. multi-concave + plain ($r_{pb}^g = 0.7$, $P < 0.05$).

Finally, another analysis looked for associations between combinations of two or more species and various landscapes. Multi-concave landscape appeared to be characterized by a large multi-species community, mainly comprising birds (*Odontophorus guyanensis*, *Penelope marail*, *Ortalis motmot*, Tinamidae), the small primate *Saimiri sciureus* and the tortoise *Chelonoidis denticulata*; the plateau landscape by the simultaneous abundance of *Pecari tajacu* and *Mazama americana*, and the multi-convex landscape by the combined high abundance of *Cebus olivaceus* and *Dasyprocta leporina*.

DISCUSSION

We found that the geomorphological typology of landscapes is a non-negligible factor driving the structure and the beta-diversity patterns of medium- to

large-bodied vertebrate communities in terra firme forests in French Guiana. The geomorphological landscapes combine effects of geology, climate, relief and history in one descriptive variable. As previously found for tree beta-diversity patterns, this integrated parameter better explains the differences between animal communities than some simple environmental parameters separately.

Habitat preference results in the disproportionate use of some resources and/or conditions over others. Habitat selection can be considered at various scales, previously defined as four selection orders (Johnson 1980). At small spatial and temporal scales, animals select different local resources or conditions. As both scales increase, these individual behavioural decisions result in survival and reproductive performances at the levels of individuals and populations. Over evolutionary time, these habitat choices contribute to the species' environmental niche or functional habitat (Gaillard *et al.* 2010). In the same way, De Caceres & Legendre (2009) stated that the r^2_{pb} value, computed from relative abundance, indicates the degree of preference of species for a target landscape compared with the other landscapes, and that 'negative correlation values tell us when a species "avoids" the target site group' (also referred to as 'negative fidelity' by phytosociologists). Following these assumptions, we interpreted the higher abundance of species in a particular habitat as a preference of this species for this habitat, resulting in higher abundance.

Some species, such as the howler monkey *Alouatta macconnelli*, appeared to be generalists or ubiquitous, and were not associated with any particular landscapes. This is consistent with other studies generally considering howler monkey as a generalist plastic species, with a varied diet (Julliot & Sabatier 1993, Simmen *et al.* 2001) and few particular requirements (Lehman 2004, Schwarzkopf & Rylands 1989). Some other species appear to have more restricted distribution: *Saimiri sciureus* were only detected in three study sites and *Pithecia pithecia* in 12. This may be related to very special habitat requirements leading to a true patchy distribution, or to very low densities in the other sites, in both cases denoting some habitat preferences although no significant results were highlighted in this study. In contrast, *Cebus apella* is a very common species encountered all over the country, but our results showed a clear preference for plain landscape type, in which they are particularly abundant. Among birds, the smallest species are characteristic of the low-altitude southern area, while *Penelope marail* is more generally associated with all the flat relief areas (northern plains and southern multi-concave area). On the other hand, *Crax alector* appears to favour steeper areas. In French Guiana, the distribution of *Crax alector* in various habitats and with respect to environmental parameters has been analysed more precisely, showing a

clear positive relationship between *C. alector* densities and the mean slope of the prospected site (Denis 2012).

Few species appeared really specialized, but although most species taken separately do not demonstrate strong habitat preferences, their assemblages produced typical communities in the various landscapes types.

Landscape communities characteristics

The multi-concave forest type appears to be the preferred habitat of a large set of species. These relatively low-elevation forests also host higher diversities of both rare and common species. We hypothesize that the lower and fragmented canopy provides a better-lit environment, with vertical strata and a greater diversity of niches. The flat environment at lower elevations can also be considered as less constraining. However one site appears to be quite different from the others with respect to most of the parameters considered, in particular for its much lower diversity. However, this site (the Waki basin) is also considered to be a very particular forest habitat type, and should probably be considered and characterized separately (Guitet *et al.* 2013, 2015).

In contrast, the other landscapes were the preferred habitat of only one or two species, and the α diversities of these sites were also lower. For example, the correlation coefficients of all animal species with montane environments were generally low, and very often negative, and only two species tended to be associated (*Ateles paniscus* and *Crax alector*). The α richness ($q = 0$) of each mountainous site was rather low (20–24), even if the estimated richness of the meta-community of whole mountainous landscape (γ diversity) was among the highest, and was similar to that of the multi-concave meta-community (32.8). These two results may indicate that our mountain sample is rather heterogeneous (greater turnover), or that many less abundant species are present in these environments, but were difficult to detect and hence only randomly detected by our sampling method. However, the larger number of study sites in this category may also explain this higher γ diversity. *Cebus olivaceus* and *Dasyprocta leporina* were the only species to be positively associated with multi-convex landscapes. These areas are generally characterized by high abundance of the tree families Lecythidaceae and Caesalpinioidae, and of several species of palm tree ($>200 \text{ ha}^{-1}$), which could explain the high abundance of this rodent (Cid *et al.* 2013). As for mountainous or multi-convex areas, few animal species clearly showed preference for plateaux, but the combined abundance of red brocket deer *Mazama americana* and the collared peccary *Pecari tajacu* is nevertheless characteristic of these environments. Like for mountainous areas, the mean α diversity was relatively low whereas the global γ diversity was higher (for $q = 0$), which could also be linked with

the large sample size in this category. Moreover, the definition of 'plateau' used in this study was probably too broad, and combined habitats that were too dissimilar. A finer-scale landscape typology identified three different types of plateaux (Guitet *et al.* 2013), but we lacked sufficient replicates to analyse the potential differences in the vertebrate community in these subcategories. In the same way, the two study sites considered in this study in the 'plain' category are in fact quite different and belonged to different types in the finer typology (Guitet *et al.* 2013). The coastal plain is the most extensively inhabited and consequently hunted area (de Thoisy *et al.* 2010), so finding replicates in undisturbed localities is challenging.

In all cases, it should be kept in mind that the diversity values estimated here depend on the methodology used, which mainly concerns the large diurnal species potentially detected during line transects. Some taxa may be underrepresented by this method, particularly nocturnal species and felids.

Relevance of the landscape typology for communities of medium- to large-bodied vertebrates

Our results highlight the influence of broad habitat categories on medium- to large-sized vertebrate communities in upland terra firme forests of French Guiana. An integrative parameter, the geomorphological landscapes proposed by Guitet *et al.* (2013), explains this heterogeneity better than most of the single parameters related to it. This is congruent with the conclusions drawn by Palminteri *et al.* (2011) that each environmental variable examined appeared to contribute to some component of the heterogeneity in primate communities in Peru, none of them being an outstanding contributor. In some cases, however, the geographic scale inherent to this classification (and used in this study) may not match field reality. For example, a medium-sized valley within a larger sloping environment was included in the mountain landscape category, whereas its faunal community was not characteristic of this landscape type (low to medium abundances of *Ateles* and *Crax*, for example). However, the overall floristic composition of this particular site matched the expected one better, according to the classification, than the faunal community (Guitet *et al.* 2015). It is likely that the temporal and geographic scales of these two biodiversity components differ. The vegetation reflects long-term climatic and geomorphological influences, whereas the large-fauna community should react more rapidly to local conditions and present filter-effects. On the other hand, some species presented affinities with two different landscapes, which for them, probably share key environmental features. For example, *Penelope marail* and *Saimiri sciureus* were associated with both the multi-concave

landscapes located in the southern part of French Guiana and with the plains located in the northern part. The common pertinent parameter may be flat relief and low elevations, independently of other parameters. The landscape classification used here permitted sufficient replicates within each type. A finer classification exists, identifying 12 different landscape types instead of five (Guitet *et al.* 2013), including three different forms of plateau, and three types of forest in the coastal plains, but additional sampling is needed to correctly analyse vertebrate assemblages at this finer scale.

A priori classifications of structural habitats do not focus on the meaning of the species distributions, with respect to active habitat selection or to environmental parameter selection by the different species. However, it corresponds to the approach used when designing legislation or policy to manage species in geographic space. Although still rough, our results may help guide territorial management of highly sensitive species, and help analyse the impacts of hunting while accounting for natural variation in abundance in various environments. More generally, the geomorphological-based typology of landscapes could be used in other countries and/or regions to characterize and predict animal community distribution throughout their territory. Coblenz & Riitters (2004) already pointed out that topography plays a primary role in regional to continental-scale biodiversity, and the landscape level is becoming more and more popular in analysis and/or resource management (Arroyo-Rodriguez & Fahrig 2014, Bonnot *et al.* 2013, Clark & Clark 2000, Hawes *et al.* 2012, Melo *et al.* 2013, Mockrin *et al.* 2011, Priego-Santander *et al.* 2013). The terra firme forests are generally known as oligotrophic forests typically sustaining low biomass densities of primates and other medium-sized to large-sized vertebrates (Emmons 1984, Haugaasen & Peres 2005a, Palacios & Peres 2005). However, they represent approximately 95% of the Amazon basin (Palacios & Peres 2005), so it is a major challenge to be able to characterize their heterogeneity, including the faunal assemblages with which they are associated.

ACKNOWLEDGEMENTS

Funding was provided for many years by ONCFS, and came from other external sources including European Funds 'HABITAT' and 'CHASSE' programmes, French Overseas Ministry, French Ministry of environment (ECOTROP programme), *Parc Amazonien de Guyane* PAG, ONF *Office National des Forêts*, CNRS Nouragues funds. EM, GJ and TD were supported by an 'Investissement d'Avenir' grant managed by Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-0025). We are very grateful to all participants in transect surveys,

including ONCFS and PAG staff and some hardworking and passionate volunteers.

LITERATURE CITED

- ANDERSON, L. O., MALHI, Y., LADLE, R. J., ARAGÃO, L. E. O. C., SHIMABUKURO, Y., PHILLIPS, O. L., BAKER, T. R., COSTA, A. C. L., ESPEJO, J. S., HIGUCHI, N., LAURANCE, W. F., LOPEZ-GONZALEZ, G., MONTEAGUDO, A. L., NUÑEZ-VARGAS, P., PEACOCK, J., QUESADA, C. A., ALMEIDA, S. & VASQUEZ, R. 2009. Influence of landscape heterogeneity on spatial patterns of wood productivity, wood specific density and above ground biomass in Amazonia. *Biogeosciences* 6:1883–1902.
- ANDERSON, M. J. 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 26:32–46.
- ARROYO-RODRIGUEZ, V. & FAHRIG, L. 2014. Why is a landscape perspective important in studies of primates? *American Journal of Primatology* 76:901–909.
- BONNOT, T. W., THOMPSON, F. R., MILLSPAUGH, J. & JONES-FARRAND, D. T. 2013. Landscape-based population viability models demonstrate importance of strategic conservation planning for birds. *Biological Conservation* 165:104–114.
- BUCHANAN-SMITH, H. M., HARDIE, S. M., CACERES, C. & PRESCOTT, M. J. 2000. Distribution and forest utilization of *Saguinus* and other primates of the Pando Department, Northern Bolivia. *International Journal of Primatology* 21:353–379.
- BUCKLAND, S. T., ANDERSON, D. R., BURNHAM, K. P. & LAAKE, J. L. 1993. *Distance sampling: estimating abundance of biological populations*. Chapman & Hall, London. 446 pp.
- CHAO, A. & SHEN, T. J. 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics* 10:429–443.
- CID, B., OLIVEIRA-SANTOS, L. R. & MOURAO, G. 2013. Seasonal habitat use of agoutis (*Dasyprocta azarae*) is driven by the palm *Attalea phalerata* in Brazilian Pantanal. *Biotropica* 45:380–385.
- CLARK, D. B. & CLARK, D. A. 2000. Landscape-scale variation in forest structure and biomass in a tropical rain forest. *Forest Ecology and Management* 137:185–198.
- COBLENTZ, D. D. & RIITERS, K. H. 2004. Topographic controls on the regional-scale biodiversity of the south-western USA. *Journal of Biogeography* 31:1125–1138.
- COUTERON, P., PÉLISSIER, R., MAPAGA, D., MOLINO, J.-F. & TEILLIER, L. 2003. Drawing ecological insights from a management-oriented forest inventory in French Guiana. *Forest Ecology and Management* 172:89–108.
- DE CACERES, M. & LEGENDRE, P. 2009. Associations between species and groups of sites: indices and statistical inference. *Ecology* 90:3566–3574.
- DE CACERES, M., LEGENDRE, P. & MORETTI, M. 2010. Improving indicator species analysis by combining groups of sites. *Oikos* 119:1674–1684.
- DE CACERES, M., LEGENDRE, P., WISER, S. K. & BROTON, L. 2012. Using species combinations in indicator value analyses. *Methods in Ecology and Evolution* 3:973–982.
- DETHOISY, B., BROSSE, S. & DUBOIS, M.-A. 2008. Assessment of large-vertebrate species richness and relative abundance in Neotropical forest using line-transect censuses: what is the minimal effort required? *Biodiversity and Conservation* 17:2627–2644.
- DE THOISY, B., RICHARD-HANSEN, C., GOGUILLON, B., JOUBERT, P., OBSTANCIAS, J., WINTERTON, P. & BROSSE, S. 2010. Rapid evaluation of threats to biodiversity: human footprint score and large vertebrate species responses in French Guiana. *Biodiversity and Conservation* 19:1567–1584.
- DEICHMANN, J. L., LIMA, A. P. & WILLIAMSON, G. B. 2011. Effects of geomorphology and primary productivity on Amazonian leaf litter herpetofauna. *Biotropica* 43:149–156.
- DENIS, T. 2012. *Caractérisation et sélection de l'habitat chez le Hocco alector (Crax alector) en Guyane française*. MSc thesis, AgroParisTech, Montpellier.
- DOLÉDEC, S. & CHESSEL, D. 1989. Rythmes saisonniers et composantes stationnelles en milieu aquatique. II Prise en compte et élimination d'effets dans un tableau faunistique. *Acta Oecologica* 10:207–232.
- DRAY, S. & DUFOUR, A.-B. 2007. The ade4 Package: implementing the duality diagram for ecologists. *Journal of Statistical Software* 22.
- DRAY, S., PÉLISSIER, R., COUTERON, P., FORTIN, M.-J., LEGENDRE, P., PERES-NETO, P. R., BELLIER, E., BIVAND, R., BLANCHET, F. G., DE CACERES, M., DUFOUR, A.-B., HEEGAARD, E., JOMBART, T., MUNOZ, F., OKSANEN, J., THIOULOUSE, J. & WAGNER, H. 2012. Community ecology in the age of multivariate multiscale spatial analysis. *Ecological Monographs* 82:257–275.
- DUPRÊNE, M. & LEGENDRE, P. 1997. Species assemblages and indicator species : the need for a flexible asymmetrical approach. *Ecological Monographs* 67:345–366.
- EMMONS, L. H. 1984. Geographic variation in densities and diversities of non-flying mammals in Amazonia. *Biotropica* 16:210–222.
- FAYAD, I., BAGHDADIA, N., GOND, V., BAILLY, J. S., BARBIER, N., EL HAJJ, M. & FABRE, F. 2014. Coupling potential of ICESat/GLAS and SRTM for the discrimination of forest landscape types in French Guiana. *International Journal of Applied Earth Observation and Geoinformation* 33:21–31.
- FIGUEIREDO, F. O. G., COSTA, F. R. C., NELSON, B. W. & PIMENTEL, T. P. 2014. Validating forest types based on geological and land-form features in central Amazonia. *Journal of Vegetation Science* 25:198–212.
- FREESE, C., HELTNE, P. G., CASTRO, N. & WHITESIDES, G. 1982. Patterns and determinants of monkey densities in Peru and Bolivia, with notes on distributions. *International Journal of Primatology* 3:53–90.
- GAILLARD, J. M., HEBBLEWHITE, M., LOISON, A., FULLER, M., POWELL, P., BASILE, M. & VAN MOORTER, B. 2010. Habitat-performance relationships: finding the right metric at a given spatial scale. *Philosophical Transactions of the Royal Society* 365:2255–2265.
- GOND, V., FREYCON, V., MOLINO, J.-F., BRUNAUX, O., INGRASSIA, F., JOUBERT, P., PEKEL, J.-F., PRÉVOST, M. F., THIERRON, V., TROMBE, P. J. & SABATIER, D. 2011. Broad scale pattern of forest landscape types in Guiana Shield. *International Journal of Applied Earth Observation and Geoinformation* 13:357–367.

- GRASSBERGER, P. 1988. Finite sample corrections to entropy and dimension estimates. *Physics Letters A* 128:369–373.
- GUITET, S., CORNU, J. F., BRUNAU, O., BETBEDER, J., CAROZZA, J. M. & RICHARD-HANSEN, C. 2013. Landform and landscape mapping, French Guiana (South America). *Journal of Maps* 9:325–335.
- GUITET, S., PELISSIER, R., BRUNAU, O., JAOUEN, G. & SABATIER, D. 2015. Geomorphological landscape features explain floristic patterns in French Guiana rainforest. *Biodiversity and Conservation*. doi: 10.1007/s10531-014-0854-8.
- HAUGAASEN, T. & PERES, C. A. 2005a. Mammal assemblage structure in Amazonian flooded and unflooded forests. *Journal of Tropical Ecology* 21:133–145.
- HAUGAASEN, T. & PERES, C. A. 2005b. Primate assemblage structure in amazonian flooded and unflooded forests. *American Journal of Primatology* 67:243–258.
- HAUGAASEN, T. & PERES, C. A. 2008. Population abundance and biomass of large-bodied birds in Amazonian flooded and unflooded forests. *Bird Conservation International* 18:87–101.
- HAWES, J. E., PERES, C. A., RILEY, L. B. & HESS, L. 2012. Landscape-scale variation in structure and biomass of Amazonian seasonally flooded and unflooded forests. *Forest Ecology and Management* 281:163–176.
- HEYMANN, E. W., ENCARNACION, F. & CANAQUIN, J. 2002. Primates of the Rio Curaray, Northern Peruvian Amazon. *International Journal of Primatology* 23:191–201.
- HILL, M. O. 1973. Diversity and evenness: a unifying notation and its consequences. *Ecology* 54:427–432.
- JOHNSON, D. H. 1980. The comparison of usage and availability measurements for evaluating resource preference. *Ecology* 61:65–71.
- JOST, L. 2006. Entropy and diversity. *Oikos* 113:363–375.
- JOST, L. 2008. GST and its relatives do not measure differentiation. *Molecular Ecology* 17:4015–4026.
- JULLIOT, C. & SABATIER, D. 1993. Diet of red howler monkey (*Alouatta seniculus*) in French Guiana. *International Journal of Primatology* 14:527–549.
- KINDT, R., VAN DAMME, P. & SIMONS, A. J. 2006. Tree diversity in western Kenya: using profiles to characterise richness and evenness. *Biodiversity and Conservation* 15:1253–1270.
- LEHMAN, S. M. 2004. Biogeography of the primates of Guyana: effects of habitat use and diet on geographic distribution. *International Journal of Primatology* 25:1225–1242.
- MARCON, E. & HÉRAULT, B. 2015. Entropart, an R package to partition diversity. *Journal of Statistical Software*, in press.
- MARCON, E., HÉRAULT, B., BARALOTO, C. & LANG, G. 2012. The decomposition of Shannon's entropy and a confidence interval for beta diversity. *Oikos* 121:516–522.
- MARCON, E., SCOTTI, I., HÉRAULT, B., ROSSI, V. & LANG, G. 2014. Generalization of the partitioning of Shannon diversity. *PLoS ONE* 9. e90289.
- MELO, F. P. L., ARROYO-RODRIGUEZ, V., FAHRIG, L., MARTINEZ-RAMOS, M. & TABARELLI, M. 2013. On the hope for biodiversity-friendly tropical landscapes. *Trends in Ecology and Evolution* 28:462–468.
- MOCKRIN, M. H., ROCKWELL, R. F., REDFORD, K. H. & KEULER, N. S. 2011. Effects of landscape features on the distribution and sustainability of ungulate hunting in northern Congo. *Conservation Biology* 25:514–525.
- PAGET, D. 1999. *Etude de la diversité spatiale des écosystèmes forestiers Guyanais*. Ph.D. thesis, Ecole Nationale du Génie Rural des Eaux et Forêts. 188 pp.
- PALACIOS, E. & PERES, C. A. 2005. Primate population densities in three nutrient-poor Amazonian terra firme forests of south-eastern Colombia. *Folia Primatologica* 76:135–145.
- PALMINTERI, S., POWELL, G. & PERES, C. A. 2011. Regional-scale heterogeneity in primate community structure at multiple undisturbed forest sites across south-eastern Peru. *Journal of Tropical Ecology* 27:181–194.
- PATIL, G. P. & TAILLIE, C. 1982. Diversity as a concept and its measurement. *Journal of the American Statistical Association* 77:548–561.
- PATTERSON, B. D., CEBALLOS, G., SECHREST, W., TOGNELLI, M. F., BROOKS, T., LUNA, L., ORTEGA, P., SALAZAR, I. & YOUNG, B. E. 2005. *Digital distribution maps of the mammals of the western hemisphere, version 2.0*. NatureServe, Arlington.
- PÉLISSIER, R., COUTERON, P., DRAY, S. & SABATIER, D. 2003. Consistency between ordination techniques and diversity measurements : two strategies for species occurrence data. *Ecology* 84:242–251.
- PERES, C. A. 1997. Primate community structure at twenty western Amazonian flooded and unflooded forest. *Journal of Tropical Ecology* 13:381–405.
- PRIEGO-SANTANDER, A. G., CAMPOS, M., BOCCO, G. & RAMÍREZ-SÁNCHEZ, L. G. 2013. Relationship between landscape heterogeneity and plant species richness on the Mexican Pacific coast. *Applied Geography* 40:171–178.
- RICHARD-HANSEN, C. 2006. Biodiversité et paysages en forêt guyanaise. Développement d'une méthodologie de caractérisation et de spatialisation des habitats à l'usage des gestionnaires des milieux naturels forestiers. Pp. 153–154 in Nivet, C., McKey, D. & Legris, C. (eds.) *Ecosystèmes tropicaux. 2ème colloque de restitution du programme de recherche*. Ministère de l'écologie et du développement durable, Paris.
- SABATIER, D., GRIMALDI, M., PREVOST, M. F., GUILLAUME, J., GODRON, M., DOSSO, M. & CURMI, P. 1997. The influence of soil cover organization on the floristic and structural heterogeneity of a Guianan rain forest. *Plant Ecology* 131:81–108.
- SCHWARZKOPF, L. & RYLANDS, A. B. 1989. Primate species richness in relation to habitat structure in Amazonian rainforest fragments. *Biological Conservation* 48:1–12.
- SHANNON, C. E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27:379–423.
- SIMMEN, B., JULLIOT, C., BAYART, F. & PAGÈS-FEUILLE, E. 2001. Diet and population densities of the primate community in relation to fruit supplies. Pp. 89–101 in Bongers, F., Charles-Dominique, P., Forget, P.-M. & Théry, M. (eds.) *Nouragues. Dynamics and plant–animal interactions in a neotropical rainforest*. Kluwer Academic Publishers, Dordrecht.
- SIMPSON, J. 1949. Measurements of diversity. *Nature* 163:688.

- SOMBROEK, W. 2000. Amazon landforms and soils in relation to biological diversity. *Acta Amazonica* 30:81–100.
- SOMBROEK, W. 2001. Spatial and temporal patterns of Amazon rainfall – consequences for the planning of agricultural occupation and the protection of primary forests. *Ambio* 30:388–396.
- SUSSMAN, R. W. & PHILLIPS-CONROY, J. E. 1995. A survey of the distribution and density of the primates of Guyana. *International Journal of Primatology* 16:761–791.
- TSALLIS, C. 1988. Possible generalization of Boltzmann–Gibbs statistics. *Journal of Statistical Physics* 52:479–487.

APPENDIX D

Decomposing phylodiversity

Marcon, E. et B. Hérault (2015). « Decomposing phylodiversity ».
In : *Methods in Ecology and Evolution* 6.3, p. 333–339.

Decomposing phylodiversity

Eric Marcon^{1*} and Bruno Hérault²

¹AgroParisTech, UMR EcoFoG, BP 709, F-97310 Kourou, French Guiana; and ²Cirad, UMR EcoFoG, BP 709, F-97310 Kourou, French Guiana

Summary

1. Measuring functional or phylogenetic diversity is the object of an active literature. The main issues to address are relating measures to a clear conceptual framework, allowing unavoidable estimation-bias correction and decomposing diversity along spatial scales.
2. We provide a general mathematical framework to decompose measures of species-neutral, phylogenetic or functional diversity into α and β components. We first unify the definitions of phylogenetic and functional entropy and diversity as a generalization of HCDT entropy and Hill numbers when an ultrametric tree is considered. We then derive the decomposition of diversity. We propose a bias correction of the estimates allowing meaningful computation from real, often undersampled communities. Entropy can be transformed into true diversity, that is an effective number of species or communities.
3. Estimators of α - and β -entropy, phylogenetic and functional entropy are provided.
4. Proper definition and estimation of diversity is the first step towards better understanding its underlying ecological and evolutionary mechanisms.

Key-words: entropy, biodiversity, phylogenetic diversity, functional diversity

Introduction

The species-neutral approach of diversity measurement is based on Hill numbers, that is the effective number of species (Jost 2006). It is now being completed by far more interesting conceptual frameworks taking into account the species relatedness, that is either their functional or their phylogenetic proximity. This is what has been called, in the first case, 'functional diversity' (Tilman *et al.* 1997) and, in the second one, 'phylogenetic diversity' or 'phylodiversity' (Webb, Losos & Agrawal 2006). When both relative abundance and degree of relatedness between species (or individuals) are quantified, Pielou (1975) suggested that diversity measures should be generalized, integrating taxonomic differences between species. A little later, Rao (1982) proposed that the average of the species differences can be used as a measure of biodiversity. Despite some attempts to take into account taxonomic distinctness into a taxic diversity measure (Vane-Wright, Humphries & Williams 1991), this 'avant-garde' idea has been hardly applied in ecology (e.g. Warwick & Clarke 1995; Crozier 1997). During the last decade, increasing interests into the evolutionary history of communities (Webb 2000) as well as the need for conservation strategies taking phylogenetic risks into account (Faith 2008) revived the interest in phylodiversity partitioning.

Phylogenetic trees are built upon the genetic similarities among various biological individuals or other superior taxa. In a given local assemblage, phylogenetic diversity aims to quantify the evolutionary history shared among individuals since the time of the most recent common ancestor (Faith 1992;

Chao, Chiu & Jost 2010). All else being equal, an assemblage of phylogenetically divergent species is often seen as more diverse than a local assemblage of closely related species (Velend *et al.* 2010). There is an increasing interest to partition this phylogenetic diversity not only between local communities but also between time periods in order to elucidate community assembly rules (Pavoine, Love & Bonsall 2009) and investigate what is commonly called the phylogenetic structure of communities (e.g. Cavender-Bares *et al.* 2004). For instance, Hardy & Senterre (2007) argued that a proper partitioning of phylodiversity is a necessary step prior to deciphering phylogenetic clustering (either due to local speciation of allopatric clades or habitat filtering of phylogenetically conserved traits) from phylogenetic overdispersion (allopatric speciation of two ancestral sympatric species, habitat filtering of phylogenetically convergent traits, competitive exclusion of related species).

Functional diversity was often defined as the extent of functional differences among individuals or species in a local community (Tilman 2001), an important determinant of ecosystem processes (Loreau *et al.* 2001). Functional diversity based on functional trees is a great tool to estimate the complementarity among individuals' or species' trait values by estimating their dispersion in trait space at all hierarchical scales simultaneously, avoiding discretization of continuous trait variation into functional groups (Petchey & Gaston 2002). Functional trees differ from phylogenetic trees as phylogenetic trees reflect evolutionary constraints whilst functional trees also take into account functional convergence (Hérault 2007). Each time a 'proper' functional tree can be constructed from a functional trait-based distance matrix (Podani & Schmera 2007), it should be possible to estimate and partition functional diversity in a manner similar to phylogenetic diversity (Petchey & Gaston

*Correspondence author. E-mail: Eric.Marcon@ecofog.gf

2002). However, functional differences among species or individuals and, *in fine*, the functional diversity value itself will depend strongly on the *a priori* choice of important functional traits (Weiher *et al.* 1999).

In this paper, we consider that all individuals or species of a given local community are placed in an ultrametric phylogenetic or functional tree. The distance between two species is measured as the length of the branches between them and their first common node. Our methods apply regardless from which biological information and how the tree is constructed, but phylogenetic diversity is the main target, as we will discuss it. We will write *phyloentropy* and *phyloentropy* for short when presenting the methods, and *phylogenetic* or *functional diversity* when we are more specific. The last two terms are also existing measures of diversity, PD (Faith 1992) and FD (Petchey & Gaston 2002). We will show that they are special cases of our measures (Table 1) and we will write PD and FD explicitly when considering them.

Chao, Chiu & Jost (2010) generalized Hill numbers to measure phyloentropy. Pavoine, Love & Bonsall (2009) generalized HCDT entropy to measure phyloentropy (Shimatani 2001; Ricotta 2005 had already done it, but for Rao's quadratic entropy only). We first show here their equivalence: phyloentropy is transformed into phyloentropy the same way HCDT entropy is transformed into diversity *sensu stricto*. Then, we derive phyloentropy partitioning as a straightforward generalization of that of HCDT diversity. We discuss the difference between our approach and that of Chiu, Jost & Chao (2014). Finally, we provide estimation-bias corrections for phyloentropy in order to obtain bias-corrected measures of phyloentropy.

Partitioning phyloentropy

TSALLIS ENTROPY

Tsallis entropy, also known as HCDT entropy (Havrda & Charvát 1967; Daróczy 1970; Tsallis 1988), has proven to be a powerful tool to measure diversity, generalizing the classical indices of diversity, including the number of species, Shannon and Simpson indices (Jost 2006). The order of diversity q gives more or less importance to rare species. Entropy can be converted into diversity *sensu stricto* (Hill 1973; Jost 2006), which is easy to interpret and compare. Statistical estimators of

diversity measures are intrinsically biased because of unseen species and also because they are not linear functions of probabilities (Marcon *et al.* 2014a). This is a serious issue (Dauby & Hardy 2012; Beck, Holloway & Schwanghart 2013), even if some bias corrections are available for HCDT entropy estimators (Grassberger 1988; Chao & Shen 2003; Marcon *et al.* 2014a).

SPECIES-NEUTRAL DIVERSITY

We first recall some features of HCDT diversity partitioning (Marcon *et al.* 2014a). Consider a metacommunity made of several local communities. Abundances of species in each local community are denoted $n_{s,i}$ ($s = 1, 2, \dots, S$ is the index of species, i the index of communities). n_s is the number of individuals of species s in the metacommunity, n_i the number of individuals sampled in local community i and n the total number. The same notations are used for probabilities of occurrence $p_{s,i}$ whose population values are unknown but estimated with $\hat{p}_{s,i} = n_{s,i}/n_i$. Community weights are w_i ; they may be equal to n_i/n , but any positive values summing to 1 are allowed. Probabilities in the metacommunity depend on these weights: $p_s = \sum_i w_i p_{s,i}$. Diversity of the metacommunity is γ -diversity. Diversity of local communities is α -diversity. The formalism of deformed logarithms is appropriated: it allows elegant and intuitive algebra. The logarithm of order q is defined as follows:

$$\ln_q x = \frac{x^{1-q} - 1}{1 - q} \quad \text{eqn 1}$$

Its inverse function is the deformed exponential given as follows:

$$e_q^x = [1 + (1 - q)x]^{1/(1-q)} \quad \text{eqn 2}$$

Note that

$$e_q^{x+y} = e_q^x e_q^y \quad \text{eqn 3}$$

Tsallis entropy of the metacommunity, ${}^q H_\gamma$, can be written as follows:

$${}^q H_\gamma = \frac{1 - \sum_s p_s^q}{q - 1} = - \sum_s p_s^q \ln_q p_s \quad \text{eqn 4}$$

Last, diversity is the deformed exponential of entropy, ${}^q D_\gamma = e_q^{q H_\gamma}$, and entropy is the deformed logarithm of diversity: ${}^q H_\gamma = \ln_q {}^q D_\gamma$.

Table 1. Many usual measures of diversity are special cases of phyloentropy, either reducing it to species-neutral diversity or limiting it to values of q equal to 0, 1 or 2

	Diversity of order q	Special values of q
Phylogenetic or functional entropy/diversity	Entropy: ${}^q \bar{H}(T)$ Diversity: ${}^q \bar{D}(T)$	$T[{}^0 \bar{H}(T) + 1]$ equals PD (Faith 1992) and FD (Petchey & Gaston 2002) $T[{}^1 \bar{H}(T)]$ equals H_p , the phylogenetic generalization of Shannon's index (Allen, Kon & Bar-Yam 2009) $T[{}^2 \bar{H}(T)]$ equals Rao's quadratic entropy
Species-neutral diversity	Entropy: ${}^q H$ Diversity: ${}^q D$	${}^0 H + 1$ is species richness ${}^1 H$ is Shannon entropy ${}^2 H$ is Simpson entropy

PHYLOENTROPY AND PHYLODIVERSITY

Consider a phylogenetic or functional ultrametric tree (Fig. 1) partitioned into depth intervals delimited by slices passing through the internal nodes. Following Chao, Chiu & Jost (2010), the first slice starts at the bottom of the tree and ends at the lowest node. In slice k , L_k leaves are found. The probabilities of occurrence of the species belonging to the branches that were below leaf l in the original tree are summed to give the grouped probability $u_{k,l}$.

We follow Pavoine, Love & Bonsall (2009) to define phyloentropy as the sum of the entropies in each tree slice, weighted by the slice height. However, we normalize it by the total tree height, $T = \sum_{k=1}^K T_k$. We denote it as ${}^q\bar{H}(T)$:

$${}^q\bar{H}(T) = \sum_{k=1}^K \frac{T_k}{T} {}^qH_k \quad \text{eqn 5}$$

qH_k is HCDT entropy in slice k . It is calculated as ${}^qH_k = -\sum_{s,i} u_{k,i}^q \ln_q u_{k,i}$.

Chao, Chiu & Jost (2010) generalized Hill numbers to phylogenetic diversity, defined as follows:

$${}^q\bar{D}(T) = \left(\sum_{k=1}^K \frac{T_k}{T} \sum_{l=1}^{L_k} u_{k,l}^q \right)^{\frac{1}{1-q}} \quad \text{eqn 6}$$

Simple algebra shows that

$${}^q\bar{D}(T) = e_q^{q\bar{H}(T)} \quad \text{eqn 7}$$

This relation is exactly the same as the relation between HCDT entropy and diversity. In other words, phyloentropy is the weighted average of entropy along the tree, and phylodiversity is the corresponding Hill number. Entropy is linear, it can be summed over slices, but diversity is not: phylodiversity is not the weighted average of diversity along the tree.

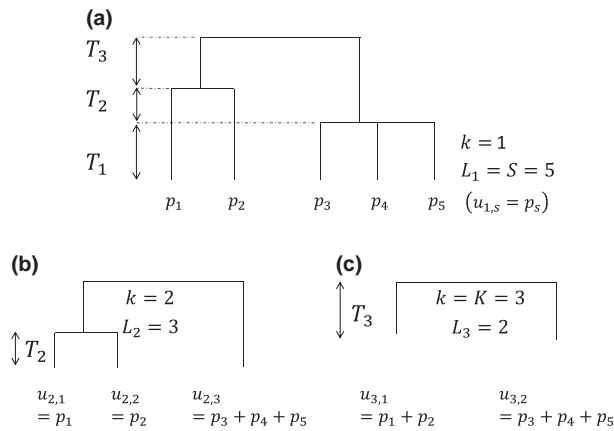


Fig. 1. Hypothetical ultrametric tree. (a) The whole tree contains three slices, delimited by two nodes. The length of slices is T_k . (b) Focus on slice 2. The tree without slice 1 is reduced to three leaves. Frequencies of collapsed species are $u_{k,l}$. (c) Slice 3 only.

DECOMPOSITION

Marcon *et al.* (2014a) derived the decomposition of HCDT entropy, generalizing Shannon entropy partitioning (Rao & Nayak 1985; Marcon *et al.* 2012), based on Patil and Taillie's concept of diversity of a mixture (Patil & Taillie 1982). Note that it differs from Jost's (2007) non-additive partitioning when community weights are unequal; see Marcon *et al.* (2014a) for a full discussion.

$${}^qH_\gamma = {}^qH_\alpha + {}^qH_\beta = \sum_i w_{i,i}^q H_\alpha + \sum_i w_{i,i}^q H_\beta \quad \text{eqn 8}$$

α - and β -entropies are the weighted sums of local community entropies ${}^qH_\alpha$ and ${}^qH_\beta$:

$$\begin{aligned} {}^qH_\alpha &= -\sum_s p_{s,i}^q \ln_q p_{s,i} \\ {}^qH_\beta &= \sum_s p_{s,i}^q \ln_q \frac{p_{s,i}}{p_s} \end{aligned} \quad \text{eqn 9}$$

Since phyloentropy is a linear transformation of generalized entropy, its decomposition is identical and follows equation (8). In slice k , HCDT γ -entropy is denoted ${}^qH_{\gamma,k}$, and the contributions of local community i to α - and β -entropy are ${}^qH_{\alpha,k,i}$ and ${}^qH_{\beta,k,i}$. This can be summed over slices and rearranged to obtain the decomposition of γ -phyloentropy:

$$\begin{aligned} \sum_k \frac{T_k}{T} {}^qH_{\gamma,k} &= \sum_k \frac{T_k}{T} \sum_i w_{i,k,i}^q H_{\alpha,k,i} + \sum_k \frac{T_k}{T} \sum_i w_{i,k,i}^q H_{\beta,k,i} \\ \Leftrightarrow {}^q\bar{H}_\gamma(T) &= \sum_i w_i \sum_k \frac{T_k}{T} {}^qH_{\alpha,k,i} + \sum_i w_i \sum_k \frac{T_k}{T} {}^qH_{\beta,k,i} \\ \Leftrightarrow {}^q\bar{H}_\gamma(T) &= \sum_i w_i {}^q\bar{H}_{\alpha,i}(T) + \sum_i w_i {}^q\bar{H}_{\beta,i}(T) = {}^q\bar{H}_\alpha(T) + {}^q\bar{H}_\beta(T) \end{aligned} \quad \text{eqn 10}$$

The deformed exponential of equation (8) is the decomposition of phylodiversity given as follows:

$$\begin{aligned} {}^q\bar{D}(T) &= {}^q\bar{D}_\alpha(T) {}^q\bar{D}_\beta(T) \\ {}^q\bar{D}(T) &= e_q^{q\bar{H}_\gamma(T)}; {}^q\bar{D}_\alpha(T) = e_q^{q\bar{H}_\alpha(T)}; {}^q\bar{D}_\beta(T) = e_q^{\frac{q\bar{H}_\beta(T)}{1+(1-q)\bar{H}_\alpha(T)}} \end{aligned} \quad \text{eqn 11}$$

α - and γ -phylodiversities can be interpreted as an equivalent number of species, that is to say the number of species equally different from each other (i.e. in an ultrametric tree made of a single slice), with the same probability of occurrence, that would give the same measure of diversity. β -phylodiversity is an equivalent number of communities, that is to say the number of completely distinct, equally weighted communities that would yield the same β -diversity as the actual metacommunity.

BIAS CORRECTION

α - and γ -HCDT entropies can be corrected following Marcon *et al.* (2014a). When q is low, unobserved species are the main issue that can be corrected according to Chao & Shen (2003). When q is high, the contribution of rare species to entropy is small, so the bias they cause is little, but entropy is less linear with respect to probabilities, requiring the correction of Grass-

berger (1988). The limit between low and high values of q is reached when both estimators are equal, empirically above $q = 1$ (Marcon *et al.* 2014a). Bias correction relies on the number of sampled individuals (probabilities are not enough) and can be computed for positive values of q . The unbiased estimators are denoted ${}^q\tilde{H}$ instead of ${}^q\hat{H}$. Their formulas are in Marcon *et al.* (2014a) and are not repeated here.

Phyloentropy can be corrected by summing the bias-corrected estimators of HCDT entropy in each slice of the tree. Bias-corrected α -entropy, ${}^q\tilde{H}_\alpha(T)$, relies on values of ${}^q\tilde{H}_{k,i}$, the bias-corrected estimators of HCDT α -entropy in slice k in local community i .

$${}^q\tilde{H}_\alpha(T) = \sum_i w_i \sum_k \frac{T_k}{T} {}^q\tilde{H}_{k,i} \quad \text{eqn 12}$$

Since the number of individuals in some leaves $u_{k,i}$ increases in slices close to the root of the tree, the bias decreases with k .

${}^q\tilde{H}_\gamma(T) = \sum_k \frac{T_k}{T} {}^q\tilde{H}_\gamma$ is calculated in the same way. β -phyloentropy is obtained as the difference between ${}^q\tilde{H}_\gamma(T)$ and ${}^q\tilde{H}_\alpha(T)$ because Grassberger's correction is not available to allow direct calculation.

EXAMPLE

We used the tropical forest data set already investigated by Marcon *et al.* (2012, 2014a). Two 1-ha plots were fully inventoried in the Paracou field station in French Guiana. 1124 individual trees (diameter at breast height over 10 cm) have been sampled among 229 species. The phylogenetic tree was built introducing a rough taxonomy of the 229 species in the analysis: distance between species of the same genus is set to 1, 2 for different genera of the same family and 3 for different families. The functional tree was based on species relatedness using four key functional traits, each of them related to one axis of the leaf-height-seed-stem economic spectra of tropical trees (Baraloto *et al.* 2010b): seed mass and tree maximum height (Hérault *et al.* 2011) plus specific leaf area and wood specific gravity (Baraloto *et al.* 2010a). The functional tree was built from a Gower's similarity matrix agglomerated using Ward's method (full details in Hérault & Honnay 2007). Diversity was calculated with the *entropart* package (Marcon & Hérault 2014) under R (R Development Core Team 2014): bias-corrected entropy was calculated first, summed and finally transformed into diversity. Necessary R codes are in the supporting information, Appendix S1.

We first calculated the species-neutral, phylogenetic and functional diversity of order 1 of the metacommunity (the two plots) and partitioned it (each plot is considered as a local community, weights are proportional to the numbers of individuals). The γ -species-neutral diversity (Hill number of Shannon entropy) is 134 effective species, partitioned into α -diversity equal to 92 effective species (82 and 107 in each plot) and β -diversity equal to 1.46 equivalent communities. Phylogenetic and functional diversity values, respectively, are: ${}^1\tilde{D}(T) = 55$ and 5.9 , ${}^1_\alpha\tilde{D}(T) = 42$ and 5.5 with ${}^1_\beta\tilde{D}(T) = 1.29$ and 1.06 . Considering the taxonomy of Paracou species, γ -phylodiversity is around 2.5 times smaller than species-neutral diversity.

Functional diversity is only six equivalent species, showing an extreme redundancy according to the functional tree: FD (Petchev & Gaston 2002), that is functional diversity of order 0, is estimated equal to 18 whilst the number of estimated species is 297.

Since γ -diversity is the product of α by β , they can be represented as nested rectangles (Fig. 2). The rectangle of size ${}^q_\beta\tilde{D}(T)$ by ${}^q_\alpha\tilde{D}(T)$ has the same area as that of size ${}^q_\gamma\tilde{D}(T)$ by 1. Plotting species-neutral and phylodiversity together summarizes the essential information: the reduction of diversity due to the consideration of species phylogenetic or functional proximity.

Profiles (Fig. 3) can be drawn for species-neutral, phylogenetic and functional diversities.

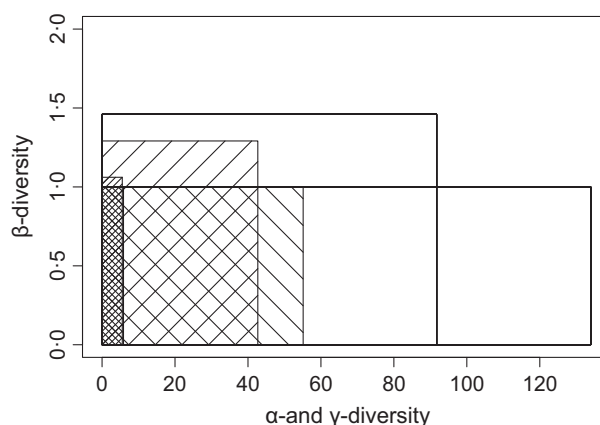


Fig. 2. Graphical representation of the diversity of order 1 in Paracou plots. Transparent rectangles represent species-neutral diversity, hatched rectangles phylogenetic diversity and shaded rectangles functional diversity. In each case, the horizontal rectangle of height 1 represents γ -diversity (respectively, 134, 55 and 6 effective species). The other rectangle has the same area, but its size is α -diversity by β -diversity.

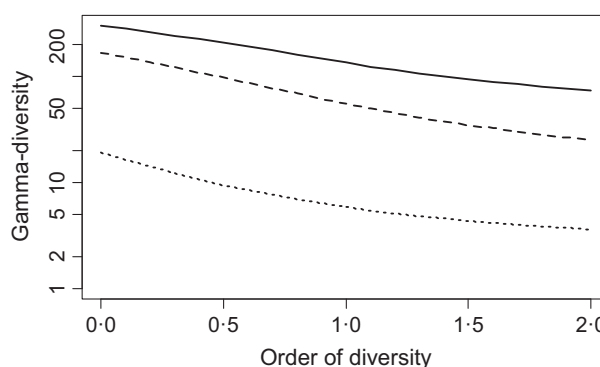


Fig. 3. γ -diversity profile of Paracou plots. Species-neutral diversity (solid line), phylogenetic diversity (dashed line) and functional diversity (dotted line) are plotted against the order of diversity, between 0 (number of species, PD and FD) and 2 (Simpson diversity and Rao's quadratic entropy transformed into diversity), with estimation-bias correction. Diversity scale is logarithmic for readability.

Discussion

UNIFICATION OF MEASURES OF DIVERSITY

Phyloentropy generalizes many previous indices of diversity. Rao's (1982) quadratic entropy is phyloentropy of order 2 multiplied by T , the tree height. It has been explored in depth and several results obtained here were already known in this special case. It has been partitioned early by Rao himself, weighting communities according to their number of individuals, as Villeger & Moullot (2008) whilst Hardy & Senterre (2007) or Pavoine *et al.* (2013) used equal weights. Hardy & Jost (2008) validated both weightings but a general framework allowing the additive partitioning of Rao's entropy was missing (Guiasu & Guiasu 2011). We showed that arbitrary weights are acceptable.

Other indices of diversity can be considered as special cases of phyloentropy (Table 1).

ALTERNATIVE PARTITIONING

Chiu, Jost & Chao (2014) propose a different partitioning of phylodiversity (Chao, Chiu & Jost 2010) focusing on the independence between α and β components, following Jost (2007). It requires a particular definition of α -diversity (in Chiu, Jost & Chao 2014; equation 6 for neutral diversity and (8) for phylodiversity), whilst we adopt Routledge's (1979) definition: α -entropy is the weighted average entropy of communities, see equation (8). Chiu *et al.*'s approach is completely different from ours, as we will show it with a simple example. Consider N communities containing a single species, no species is shared between communities. Whatever q , the entropy of each community is 0, its diversity is 1 effective species. In our framework, α -entropy equals 0 and α -diversity is 1. More generally, whatever the weights and whatever q , if all communities have the same diversity, α -diversity equals it.

In Chiu *et al.*'s framework, β -diversity must be N since no species are shared, so α -diversity is γ -diversity divided by N . Species-neutral α -diversity of our example is not 1 but $\frac{1}{N} \left(\sum_{i=1}^N w_i^q \right)^{\frac{1}{1-q}}$. Community weights and species frequencies play a similar role: low-weighted communities, as rare species, have a lower influence when q increases, and inversely, α -diversity is driven by rare species of low-weighted communities when q decreases. We consider in this paper that community weights are arbitrary, such as sampling unit sizes, so Chiu *et al.*'s α -diversity is not suitable here.

We believe that Routledge's definition of α -diversity is more appropriate. Entropy is the average information in each community so it can meaningfully be averaged between communities according to their weight to define α -entropy. Adding an infinitesimal community (with weight close to 0) does not change the metacommunity's diversity, whilst it changes discontinuously in Chiu *et al.*'s framework (β -diversity jumps from N to $N + 1$, for example).

The price to pay is α - and β -diversities are not independent, as discussed more thoroughly in Marcon *et al.* (2014a). The real consequences of this dependence will have to be studied in depth.

NON-ULTRAMETRIC DISTANCES BETWEEN SPECIES

Our framework relies on ultrametric trees, since entropy must be calculated slice by slice. Phylogenetic data are usually organized as a tree, but not necessarily ultrametric. Chao, Chiu & Jost (2010) calculate $q\bar{D}(T)$ as a sum over the branches rather than other slices of the trees, allowing them to address non-ultrametric trees. Although it is defined mathematically, such a value of phylodiversity faces several issues. Pavoine & Bonsall (2009) discuss its inconsistency in the special case of $q = 2$, for example, the fact that the species distribution maximizing diversity is not unique then. Leinster & Cobbold (2012) show that the distance between species used to calculate $q\bar{D}(T)$ depends on species frequencies, questioning the very sense of what is measured. For these two reasons, we conclude that non-ultrametric trees are not appropriate to measure phylodiversity in our framework, not only for technical issues (only ultrametric trees can be sliced to allow estimation-bias correction) but for conceptual ones.

Functional diversity is more frequently calculated as a non-ultrametric matrix of distances between species, whose transformation into a dendrogram causes deformations (Pavoine, Ollier & Dufour 2005). The choice of the clustering method influences the shape of the tree and may lead to inconsistent results (Podani & Schmera 2006), although appropriate methods, applied to the example above, reduce these issues (Podani & Schmera 2007). A more appropriate way to address functional diversity is probably using directly the distance matrix between species or its transformation into a similarity matrix. Similarity-based diversity (Leinster & Cobbold 2012) may be preferred to evaluate functional diversity. We derive its decomposition and propose reduced-bias estimators elsewhere (Marcon, Zhang & Hérault 2014b).

Conclusion

In this paper, we provide a general, consistent and operational framework to decompose measures of species-neutral, phylogenetic or even functional diversity into α (within local communities) and β (between local communities) components. We show that entropy can be calculated and its estimation bias corrected in each slice of the phylogenetic or functional tree, summed over slices and finally transformed into diversity. In fact, phylodiversity can be analysed without using any species concept (i.e. diversity of individuals without categorizing them into a set of species) provided that phylogenetic or functional distance between individuals can be assessed, for example using molecular data or functional trait measured for each individual member of a metacommunity (Paine *et al.* 2011). Being able to properly partition phylodiversity is a necessary step towards deciphering the ecological and evolutionary mechanisms that underlie the structure and assembly of communities. Moreover, diversity partitioning will improve our assessment of human-driven modifications of ecosystem functioning in conservation studies.

Acknowledgements

This work has benefited from an 'Investissement d'Avenir' grant managed by Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-0025). We thank Dr David Nipperess and an anonymous referee for their helpful comments and suggestions.

Data accessibility

The R scripts used to work the examples are available in the online supplement of the paper. They rely on the entropart package (Marcon & Hérault 2014) for R, which contains the data.

References

- Allen, B., Kon, M. & Bar-Yam, Y. (2009) A new phylogenetic diversity measure generalizing the Shannon index and its application to phyllostomid bats. *American Naturalist*, **174**, 236–243.
- Baraloto, C., Paine, C.E.T.P., Patiño, S., Bonal, D., Hérault, B. & Chave, J. (2010a) Functional trait variation and sampling strategies in species rich plant communities. *Functional Ecology*, **24**, 208–216.
- Baraloto, C., Paine, C.E.T., Poorter, L., Beauchêne, J., Bonal, D., Domenach, A.M. *et al.* (2010b) Decoupled leaf and stem economics in rain forest trees. *Ecology Letters*, **13**, 1338–1347.
- Beck, J., Holloway, J.D. & Schwanghart, W. (2013) Undersampling and the measurement of beta diversity. *Methods in Ecology and Evolution*, **4**, 370–382.
- Cavender-Bares, J., Ackerly, D.D., Baum, D.A. & Bazzaz, F.A. (2004) Phylogenetic overdispersion in Floridian oak communities. *The American Naturalist*, **163**, 823–843.
- Chao, A., Chiu, C.-H. & Jost, L. (2010) Phylogenetic diversity measures based on Hill numbers. *Philosophical Transactions of the Royal Society B*, **365**, 3599–3609.
- Chao, A. & Shen, T.J. (2003) Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*, **10**, 429–443.
- Chiu, C.H., Jost, L. & Chao, A. (2014) Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers. *Ecological Monographs*, **84**, 21–44.
- Crozier, R.H. (1997) Preserving the information content of species: genetic diversity, phylogeny, and conservation worth interaction with reasons justifying the preservation. *Annual Review of Ecology and Systematics*, **28**, 243–268.
- Daróczy, Z. (1970) Generalized information functions. *Information and Control*, **16**, 36–51.
- Dauby, G. & Hardy, O.J. (2012) Sampled-based estimation of diversity sensu stricto by transforming Hurlbert diversities into effective number of species. *Ecography*, **35**, 661–672.
- Faith, D.P. (1992) Conservation evaluation and phylogenetic diversity. *Biological Conservation*, **61**, 1–10.
- Faith, D.P. (2008) Threatened species and the potential loss of phylogenetic diversity: conservation scenarios based on estimated extinction probabilities and phylogenetic risk analysis. *Conservation Biology: The Journal of the Society for Conservation Biology*, **22**, 1461–1470.
- Grassberger, P. (1988) Finite sample corrections to entropy and dimension estimates. *Physics Letters A*, **128**, 369–373.
- Guiasu, R.C. & Guiasu, S. (2011) The weighted quadratic index of biodiversity for pairs of species: a generalization of Rao's index. *Natural Science*, **3**, 795–801.
- Hardy, O.J. & Jost, L. (2008) Interpreting and estimating measures of community phylogenetic structuring. *Journal of Ecology*, **96**, 849–852.
- Hardy, O.J. & Senterre, B. (2007) Characterizing the phylogenetic structure of communities by an additive partitioning of phylogenetic diversity. *Journal of Ecology*, **95**, 493–506.
- Havrda, J. & Charvát, F. (1967) Quantification method of classification processes. Concept of structural a-entropy. *Kybernetika*, **3**, 30–35.
- Hérault, B. (2007) Reconciling niche and neutrality through the Emergent Group approach. *Perspectives in Plant Ecology, Evolution and Systematics*, **9**, 71–78.
- Hérault, B. & Honnay, O. (2007) Using life-history traits to achieve a functional classification of habitats. *Applied Vegetation Science*, **10**, 73–80.
- Hérault, B., Bachelot, B., Poorter, L., Rossi, V., Bongers, F., Chave, J., Paine, C.E.T., Wagner, F. & Baraloto, C. (2011) Functional traits shape ontogenetic growth trajectories of rain forest tree species. *Journal of Ecology*, **99**, 1431–1440.
- Hill, M.O. (1973) Diversity and evenness: a unifying notation and its consequences. *Ecology*, **54**, 427–432.
- Jost, L. (2006) Entropy and diversity. *Oikos*, **113**, 363–375.
- Jost, L. (2007) Partitioning diversity into independent alpha and beta components. *Ecology*, **88**, 2427–2439.
- Leinster, T. & Cobbold, C. (2012) Measuring diversity: the importance of species similarity. *Ecology*, **93**, 477–489.
- Loreau, M., Naeem, S., Inchausti, P., Bengtsson, J., Grime, J.P., Hector, A. *et al.* (2001) Biodiversity and ecosystem functioning: current knowledge and future challenges. *Science*, **294**, 804–808.
- Marcon, E. & Hérault, B. (2014). entropart, an R package to partition diversity. *Journal of Statistical Software*, in press.
- Marcon, E., Zhang, Z. & Hérault, B. (2014b) The decomposition of similarity-based diversity and its bias correction. *HAL*, **00989454**, 1–12.
- Marcon, E., Hérault, B., Baraloto, C. & Lang, G. (2012) The decomposition of Shannon's entropy and a confidence interval for beta diversity. *Oikos*, **121**, 516–522.
- Marcon, E., Scotti, I., Hérault, B., Rossi, V. & Lang, G. (2014a) Generalization of the partitioning of Shannon diversity. *PLoS One*, **9**, e90289.
- Paine, C.E.T., Baraloto, C., Chave, J. & Hérault, B. (2011) Functional traits of individual trees reveal ecological constraints on community assembly in tropical rain forests. *Oikos*, **120**, 720–727.
- Patil, G.P. & Taillie, C. (1982) Diversity as a concept and its measurement. *Journal of the American Statistical Association*, **77**, 548–561.
- Pavoine, S. & Bonsall, M.B. (2009) Biological diversity: distinct distributions can lead to the maximization of Rao's quadratic entropy. *Theoretical Population Biology*, **75**, 153–163.
- Pavoine, S., Love, M.S. & Bonsall, M.B. (2009) 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. *Ecology Letters*, **12**, 898–908.
- Pavoine, S., Ollier, S. & Dufour, A.-B. (2005) Is the originality of a species measurable? *Ecology Letters*, **8**, 579–586.
- Pavoine, S., Blondel, J., Dufour, A.B., Gasc, A. & Bonsall, M.B. (2013) A new technique for analysing interacting factors affecting biodiversity patterns: crossed-DPCoA. *PLoS One*, **8**, e54530.
- Petchey, O.L. & Gaston, K.J. (2002) Functional diversity (FD), species richness and community composition. *Ecology Letters*, **5**, 402–411.
- Pielou, E.C. (1975) *Ecological Diversity*. Wiley, New York.
- Podani, J. & Schmera, D. (2006) On dendrogram-based measures of functional diversity. *Oikos*, **115**, 179–185.
- Podani, J. & Schmera, D. (2007) How should a dendrogram-based measure of functional diversity function? A rejoinder to Petchey and Gaston. *Oikos*, **116**, 1427–1430.
- R Development Core Team. (2014). *R: A Language and Environment for Statistical Computing*. R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria.
- Rao, C.R. (1982) Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology*, **21**, 24–43.
- Rao, C. & Nayak, T. (1985) Cross entropy, dissimilarity measures, and characterizations of quadratic entropy. *Information Theory, IEEE Transactions on*, **31**, 589–593.
- Ricotta, C. (2005) On hierarchical diversity decomposition. *Journal of Vegetation Science*, **16**, 223–226.
- Routledge, R.D. (1979) Diversity indices: which ones are admissible? *Journal of Theoretical Biology*, **76**, 503–515.
- Shimatan, K. (2001) Multivariate point processes and spatial variation of species diversity. *Forest Ecology and Management*, **142**, 215–229.
- Tilman, D. (2001). Functional diversity. *Encyclopedia of Biodiversity* (ed. S. Levin), pp. 109–121. Academic Press, San Diego.
- Tilman, D., Knops, J., Wedin, D., Reich, P., Ritchie, M. & Siemann, E. (1997) The influence of functional diversity and composition on ecosystem processes. *Science*, **277**, 1300–1302.
- Tsallis, C. (1988) Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, **52**, 479–487.
- Vane-Wright, R.I., Humphries, C.J. & Williams, P.H. (1991) What to protect? – Systematics and the agony of choice. *Biological Conservation*, **55**, 235–254.
- Vellend, M., Cornwell, W.K., Magnuson-Ford, K. & Moers, A.Ø. (2010). Measuring phylogenetic biodiversity. *Biological Diversity: Frontiers in Measurement and Assessment* (eds A.E. Magurran & B.J. McGill), pp. 194–207. Oxford University Press, Oxford.
- Villeger, S. & Moullot, D. (2008) Additive partitioning of diversity including species differences: a comment on Hardy & Senterre (2007). *Journal of Ecology*, **96**, 845–848.

- Warwick, R.M. & Clarke, K.R. (1995) New “biodiversity” measures reveal a decrease in taxonomic distinctness with increasing stress. *Marine Ecology Progress Series*, **129**, 301–305.
- Webb, C.O. (2000) Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *American Naturalist*, **156**, 145–155.
- Webb, C.O., Losos, J.B. & Agrawal, A.A. (2006) Integrating phylogenies into community ecology. *Ecology*, **87**, S1–S2.
- Weiher, E., van der Werf, A., Hompson, K., Roderick, M., Garnier, E. & Eriksson, O. (1999) Challenging Theophrastus: a common core list of plant traits for functional ecology. *Journal of Vegetation Science*, **10**, 609–620.

Received 11 March 2014; accepted 25 November 2014
Handling Editor: Robert Freckleton

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. R code of the example.

APPENDIX E

Generalization of the Partitioning of Shannon Diversity

Marcon, E., I. Scotti, B. Hérault, V. Rossi et G. Lang (2014).
« Generalization of the Partitioning of Shannon Diversity ». In :
Plos One 9.3, e90289.

Generalization of the Partitioning of Shannon Diversity

Eric Marcon^{1*}, Ivan Scotti², Bruno Hérault³, Vivien Rossi³, Gabriel Lang^{4,5}

1 AgroParisTech, UMR Écologie des Forêts de Guyane, Kourou Cedex, France, **2** INRA, UMR Écologie des Forêts de Guyane, Kourou Cedex, France, **3** CIRAD, UMR Écologie des Forêts de Guyane, Kourou Cedex, France, **4** AgroParisTech, UMR 518 Math. Info. Appli., Paris, France, **5** INRA, UMR 518 Math. Info. Appli., Paris, France

Abstract

Traditional measures of diversity, namely the number of species as well as Simpson's and Shannon's indices, are particular cases of Tsallis entropy. Entropy decomposition, *i.e.* decomposing gamma entropy into alpha and beta components, has been previously derived in the literature. We propose a generalization of the additive decomposition of Shannon entropy applied to Tsallis entropy. We obtain a self-contained definition of beta entropy as the information gain brought by the knowledge of each community composition. We propose a correction of the estimation bias allowing to estimate alpha, beta and gamma entropy from the data and eventually convert them into true diversity. We advocate additive decomposition in complement of multiplicative partitioning to allow robust estimation of biodiversity.

Citation: Marcon E, Scotti I, Hérault B, Rossi V, Lang G (2014) Generalization of the Partitioning of Shannon Diversity. PLoS ONE 9(3): e90289. doi:10.1371/journal.pone.0090289

Editor: Jean Thioulouse, CNRS - Université Lyon 1, France

Received: January 25, 2013; **Accepted:** January 31, 2014; **Published:** March 6, 2014

Copyright: © 2014 Marcon et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work has benefited from an "Investissement d'Avenir" grant managed by Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-25-01). Funding came from the project Climfor (Fondation pour la Recherche sur la Biodiversité). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Co-author Bruno Hérault is a PLOS ONE Editorial Board member. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: eric.marcon@ecofog.gf

Introduction

Diversity partitioning means that, in a given area, the gamma diversity D_γ of all individuals found may be split into within (alpha diversity, D_α) and between (beta diversity, D_β) local assemblages. Alpha diversity reflects the diversity of individuals in local assemblages whereas beta diversity reflects the diversity of the local assemblages. The latter, D_β , is commonly derived from D_α and D_γ estimates [1]. Recently, a prolific literature has emerged on the problem of diversity partitioning, because it addresses the issue of quantifying biodiversity at large scale. Jost's push [2–5] has helped to clarify the concepts behind diversity partitioning but mutually exclusive viewpoints have been supported, in particular in a forum organized by Ellison [6] in *Ecology*. A recent synthesis by Chao *et al.* [7] wraps up the debate and attempts to reach a consensus. Traditional measures of diversity, namely the number of species as well as Simpson's and Shannon's indices, are all special cases of the Tsallis entropy [8,9]. The additive decomposition [10] of these diversity measures does not provide independent components but Jost [3] derived a non-additive partitioning of entropy which does.

A rigorous vocabulary is necessary to avoid confusion. *Unrelated* or *independent* (sensu [7]) means that the range of values of $^qD_\beta$ is not constrained by the value of $^qD_\alpha$, which is a desirable property. *Unrelated* is more pertinent than *independent* since diversity is not a random variable here, but *independent* is widely used, by [3] for example. We will write *independent* throughout the paper for convenience. We will write *partitioning* only when independent components are obtained and *decomposition* in other cases.

Tsallis entropy can be easily transformed into Hill numbers [11]. Jost [3] called Hill numbers *true diversity* because they are homogeneous to a number of species and have a variety of desirable properties that will be recalled below. We will call *diversity*

true diversity only, and *entropy* Simpson and Shannon indices as well as Tsallis entropy. The multiplicative partitioning of true γ diversity allows obtaining independent values of α and β diversity when local assemblages are equally weighted.

However, we believe that the additive decomposition of entropy still has something to tell us. In this paper, we bring out an appropriate mathematical framework that allows us to write Tsallis entropy decomposition. We show its mathematical equivalence to the multiplicative partition of diversity. This is simply a generalization of the special case of Shannon diversity [12]. Doing so, we establish a self-contained (*i.e.* it does not rely on the definitions of α and γ entropies) definition of β entropy, showing it is a generalized Jensen-Shannon divergence, *i.e.* the average generalized Kullback-Leibler divergence [13] between local assemblages and their average distribution. Beyond clarifying and making explicit some concepts, we acknowledge that this decomposition framework largely benefits from a consistent literature in statistical physics. In particular, we rely on it to propose bias corrections that can be applied to Tsallis entropy in general. After bias correction, conversion of entropy into true diversity provides independent, easy-to-interpret components of diversity. Our findings complete the well-established non-additive (also called pseudo-additive) partitioning of Tsallis entropy. We detail their differences all along the paper.

Methods

Consider a meta-community partitioned into several local communities (let $i = 1, 2, \dots, I$ denote them). n_i individuals are sampled in community i . Let $s = 1, 2, \dots, S$ denote the species that compose the meta-community, n_{si} the number of individuals of species s sampled in the local community i , $n_s = \sum_i n_{si}$ the total number of individuals of species s , $n = \sum_s \sum_i n_{si}$ the total number

of sampled individuals. Within each community i , the probability p_{si} for an individual to belong to species s is estimated by $\hat{p}_{si} = n_{si}/n_i$. The same probability for the meta-community is p_s . Communities may have a weight, w_i , satisfying $p_s = \sum_i w_i p_{si}$. The commonly-used $w_i = n_i/n$ is a possible weight, but the weighting may be arbitrary (e.g. the sampled areas).

We now define precisely entropy. Given a probability distribution $\mathbf{p}_s = \{p_1; p_2; \dots; p_s; \dots; p_S\}$, we choose an information function $\mathcal{I}(p_s)$, which is a decreasing function of p_s having the property $\mathcal{I}(1) = 0$: information is much lower when a frequent species is found. Entropy is defined as the average amount of information obtained when an individual is sampled [14]:

$$H = \sum_s p_s \mathcal{I}(p_s) \quad (1)$$

The best-known information function is $\mathcal{I}(p_s) = -\ln(p_s)$. This defines the entropy of Shannon [15]. $\mathcal{I}(p_s) = (1-p_s)/p_s$ yields the number of species minus 1 and $\mathcal{I}(p_s) = 1-p_s$, Simpson's [16] index. Relative entropy is defined when the information function quantifies how different an observed distribution \mathbf{p}_s is different from the expected distribution \mathbf{p}'_s . The Kullback-Leibler [17] divergence is the best-known relative entropy, equal to $\sum_s p_s \ln(p_s/p'_s)$. Shannon's beta entropy has been shown to be the weighted sum of the Kullback-Leibler divergence of local communities, where the expected probability distribution of species in each local community is that of the meta-community [12,18]:

$${}^1H_\beta = \sum_i w_i \sum_s p_{si} \ln\left(\frac{p_{si}}{p_s}\right) \quad (2)$$

Let us define γ as the meta-community's diversity, α as local communities' diversities, and β as diversity between local communities. Tsallis γ entropy of order q is defined as:

$${}^qH_\gamma = \frac{1 - \sum_s p_s^q}{q-1} \quad (3)$$

and the corresponding α entropy in the local community i is:

$${}^qH_\alpha = \frac{1 - \sum_s p_{si}^q}{q-1} \quad (4)$$

The natural definition of the total α entropy is the weighted average of local community's entropies, following Routledge [19]:

$${}^qH_\alpha = \sum_i w_i {}^qH_{\alpha_i} \quad (5)$$

This is the key difference between our decomposition framework and the non-additive one. Jost [3] proposed another definition, ${}^qH_\alpha = \sum_i (w_i^q / \sum_i w_i^q) {}^qH_{\alpha_i}$, i.e. the normalized q -expectation of the entropy of communities [20] rather than their weighted mean. It is actually a derived result, see the discussion below. Our results rely on Routledge's definition (see Appendix S1).

α and γ diversity values are given by Hill numbers qD , called "numbers equivalent" or "effective number of species", i.e. the number of equally-frequent species that would give the same level of diversity as the data [14]:

$${}^qD_\gamma = \left(\sum_s p_s^q \right)^{\frac{1}{1-q}} \quad (6)$$

Routledge α diversity is:

$${}^qD_\alpha = \left(\sum_i w_i \sum_s p_{si}^q \right)^{\frac{1}{1-q}} \quad (7)$$

Combining (3) and (6) yields:

$${}^qD_\gamma = (1 - (q-1) {}^qH_\gamma)^{\frac{1}{1-q}} \quad (8)$$

We also use the formalism of deformed logarithms, proposed by Tsallis [21] to simplify manipulations of entropy. The deformed logarithm of order q is defined as:

$$\ln_q x = \frac{x^{1-q} - 1}{1-q} \quad (9)$$

It converges to \ln when $q \rightarrow 1$.

The inverse function of $\ln_q x$ is the deformed exponential:

$$e_q^x = [1 + (1-q)x]^{\frac{1}{1-q}} \quad (10)$$

The basic properties of deformed logarithms are:

$$\ln_q(xy) = \ln_q x + \ln_q y - (q-1)(\ln_q x)(\ln_q y) \quad (11)$$

$$\ln_q \frac{1}{x} = -x^{q-1} \ln_q x \quad (12)$$

$$e_q^{x+y} = e_q^x e_q^{\frac{y}{1-(q-1)x}} \quad (13)$$

Tsallis entropy can be rewritten as:

$${}^qH_\gamma = \frac{1 - \sum_s p_s^q}{q-1} = - \sum_s p_s^q \ln_q p_s \quad (14)$$

Diversity and Tsallis entropy are transformations of each other:

$${}^qH_\gamma = \ln_q {}^qD_\gamma \quad (15)$$

$${}^qD_\gamma = e_q^{qH_\gamma} \quad (16)$$

$${}^qH_\beta = \sum_s p_s^q \ln_q \frac{p_{si}}{p_s} \quad (23)$$

Decomposing diversity of order q

We start from the multiplicative partitioning of true diversity.

$${}^qD_\gamma = {}^qD_\alpha {}^qD_\beta \quad (17)$$

If community weights are equal, β diversity is independent of α diversity (it is whatever the weights if α diversity is weighted according to Jost, but this is not our choice). We will consider the unequal weight case later.

β diversity is the equivalent number of communities, *i.e.* the number of equally-weighted, non-overlapping communities that would have the same diversity as the observed ones.

We want to explore the properties of entropy decomposition. We calculate the deformed logarithm of equation (17):

$$\ln_q {}^qD_\gamma = \ln_q {}^qD_\alpha + \ln_q {}^qD_\beta - (q-1)(\ln_q {}^qD_\alpha)(\ln_q {}^qD_\beta) \quad (18)$$

$$\Leftrightarrow {}^qH_\gamma = {}^qH_\alpha + \ln_q {}^qD_\beta - (q-1)({}^qH_\alpha)(\ln_q {}^qD_\beta) \quad (19)$$

Equation (19) is Jost's partitioning framework (equation 8f in [3]). Jost retains $H_B = \ln_q {}^qD_\beta$ as the β component of entropy partitioning. It is independent of ${}^qH_\alpha$ (they are respective transformations of independent ${}^qD_\beta$ and ${}^qD_\alpha$), contrarily to the β component of the additive decomposition [10,22] defined as ${}^qH_\gamma - {}^qH_\alpha$.

After some algebra requiring Routledge's definition of α diversity detailed in Appendix S1, we obtain from equation (19):

$${}^qH_\gamma - {}^qH_\alpha = \frac{\sum_i w_i \sum_s p_{si}^q - \sum_s p_s^q}{q-1} \quad (20)$$

The right term of equation (20) is a possible definition of the β component of additive decomposition. It can be much improved if we consider $\sum_s p_s^q = \sum_s p_s^{q-1} \sum_i w_i p_{si}$ and rearrange equation (20) to obtain:

$${}^qH_\gamma - {}^qH_\alpha = \sum_i w_i \sum_s p_{si}^q \ln_q \frac{p_{si}}{p_s} \quad (21)$$

We obtained the β entropy of order q . It is the weighted average of the generalized Kullback-Leibler divergence of order q (previously derived by Borland *et al.* [13] in thermostatics) between each community and the meta-community:

$${}^qH_\beta = \sum_i w_i {}^qH_{\beta i} \quad (22)$$

${}^qH_\beta$ converges to the Kullback-Leibler divergence when $q \rightarrow 1$.

The average Kullback-Leibler divergence between several distributions and their mean is called Jensen-Shannon divergence [23], so our β entropy ${}^qH_\beta$ can be called *generalized Jensen-Shannon divergence*. It is different from the non-logarithmic Jensen-Shannon divergence [24] which measures the difference between the equivalent of our α entropy and $-\sum_i w_i p_{si}^q \ln_q p_s^q$ (the latter is not Tsallis γ entropy).

Our results are summarized in Table 1, including transformation of entropy into diversity. The partition of entropy of order q is formally similar to that of Shannon entropy. It is in line with Patil and Taillie's [14] conclusions: ${}^qH_\beta$ is the information gain attributable to the knowledge that individuals belong to a particular community, beyond belonging to the meta-community.

Information content of generalized entropy

Both ${}^qH_\gamma$ and ${}^qH_\beta$ must be rearranged to reveal their information function and explicitly write them as entropies. Straightforward algebra yields:

$${}^qH_\gamma = - \sum_s p_s \frac{p_s^{q-1} - 1}{q-1} \quad (24)$$

$${}^qH_\beta = \sum_s p_{si} \frac{p_{si}^{q-1} - p_s^{q-1}}{q-1} \quad (25)$$

The information functions respectively tend to those of Shannon entropy when $q \rightarrow 1$.

Properties of generalized β entropy

${}^qH_\beta$ is not independent of ${}^qH_\alpha$. Only Jost's H_B is an independent β component of diversity indices. But ${}^qH_\beta$ takes place in a generalized decomposition of entropy. Its limit when $q \rightarrow 1$ is Shannon β entropy, and in this special case only ${}^qH_\beta$ is independent of ${}^qH_\alpha$.

${}^qH_\beta$ is interpretable and self-contained (*i.e.* it is not just a function of γ and α entropies): it is the information gain brought by the knowledge of each local community's species probabilities related to the meta-community's probabilities. It is an entropy, defined just as Shannon β entropy but with a generalized information function.

${}^qH_\beta$ is always positive (proof in [25]), so entropy decomposition is not limited to equally-weighted communities.

Bias correction

Estimation bias (we follow the terminology of Dauby and Hardy [26]) is a well-known issue. Real data are almost always samples of larger communities, so some species may have been missed. The induced bias on Simpson entropy is smaller than on Shannon entropy because the former assigns lower weights to rare species, *i.e.* the sampling bias is even more important when q decreases.

We denote ${}^q\hat{H}$ the naive estimators of entropy, obtained by applying the above formulas to estimators of probabilities (such as ${}^q\hat{H}_\beta = \sum_s \hat{p}_{si}^q \ln_q (\hat{p}_{si}/\hat{p}_s)$). Let ${}^q\tilde{H}$ denote the estimation-bias corrected estimators. Chao and Shen's [27] correction can be

Table 1. Values of entropy and diversity for generalized entropy of order q and Shannon entropy.

Diversity measure	Generalized entropy	Shannon
γ entropy	${}^qH_\gamma = -\sum_s p_s^q \ln_q p_s$	${}^1H_\gamma = -\sum_s p_s \ln p_s$
β entropy	${}^qH_\beta = \sum_i w_i \sum_s p_{si}^q \ln_q \frac{p_{si}}{p_s}$	${}^1H_\beta = \sum_i w_i \sum_s p_{si} \ln \frac{p_{si}}{p_s}$
True γ diversity (Hill number)	${}^qD_\gamma = e_q^{qH_\gamma}$	${}^1D_\gamma = e^{H_\gamma}$
True β diversity (numbers equivalent)	${}^qD_\beta = e_q^{\frac{qH_\beta}{1-(q-1)^{1/H_\beta}}}$	${}^1D_\beta = e^{H_\beta}$

The deformed logarithm formalism allows presenting all orders of entropy as a generalization of Shannon entropy. Generalized β entropy is a generalized Kullback-Leibler divergence, i.e. the information gain obtained by the knowledge of each community's composition beyond that of the meta-community. Robust estimation of the entropy of real communities requires estimation bias correction introduced in the text.

doi:10.1371/journal.pone.0090289.t001

applied to all of our estimators. It relies on the Horvitz-Thomson [28] estimator which corrects a sum of measurements for missing species by dividing each measurement by $1 - (1 - \hat{p}_{si})^n$, i.e. the probability for each species to be present in the sample. Next, the sample coverage of community i , denoted C_i , is the sum of probabilities the species of the sample represent in the whole community. It is easily estimated [29] from the number of singletons (species observed once) of the sample, denoted S_i^1 , and the sample size n_i :

$$\hat{C}_i = 1 - \frac{S_i^1}{n_i} \quad (26)$$

The sample coverage of the meta-community is estimated the same way: $\hat{C} = 1 - S^1/n$. An unbiased estimator of p_{si} is $\tilde{p}_{si} = \hat{C}_i \hat{p}_{si}$, and $\tilde{p}_s = \hat{C} \hat{p}_s$. Combining sample coverage, Horvitz-Thomson and equation (23) estimator yields:

$${}^q\tilde{H}_\gamma = -\sum_s \frac{(\hat{C} \hat{p}_s)^q \ln_q \hat{C} \hat{p}_s}{1 - (1 - \hat{C} \hat{p}_s)^n} \quad (27)$$

$${}^q\tilde{H}_\beta = \sum_s \frac{(\hat{C}_i \hat{p}_{si})^q \ln_q \frac{\hat{C}_i \hat{p}_{si}}{\hat{C} \hat{p}_s}}{1 - (1 - \hat{C}_i \hat{p}_{si})^n} \quad (28)$$

Another estimation bias has been widely studied by physicists. The latter generally consider that all species of a given community are known and their probabilities quantified. Their main issue is not at all missing species but the non-linearity of entropy measures (see [30] for a short review). Probabilities p_s are estimated by \hat{p}_s . For $q > 0$, estimating p_s^q by $(\hat{p}_s)^q$ is an important source of underestimation of entropy. Grassberger [31] derived an unbiased estimator \tilde{p}_s^q under the assumption that the number of observed individuals of a species along successive samplings follows a Poisson distribution, as in Fisher's model [32] although arguments are different. Grassberger shows that:

$$\tilde{p}_s^q \approx n_s^{-q} \left(\frac{\Gamma(n_s + 1)}{\Gamma(n_s - q + 1)} + \frac{(-1)^n \Gamma(1 + q) \sin \pi q}{\pi(n + 1)} \right) \quad (29)$$

where $\Gamma(\cdot)$ is the gamma function ($\Gamma(n) = (n-1)!$ if n is an integer). Practical computation of $\Gamma(n_s + 1)$ is not possible for large samples so the first term of the sum must be rewritten as:

$\Gamma(n_s + 1)/\Gamma(n_s - q + 1) = \Gamma(q)/\mathcal{B}(n_s - q + 1, q)$ where \mathcal{B} is the beta function. This estimator can be plugged into the formula of Tsallis γ entropy to obtain:

$${}^q\tilde{H}_\gamma = \frac{1 - \sum_s \tilde{p}_s^q}{q - 1} \quad (30)$$

Other estimations of p_s^q are readily detailed here. Holste *et al.* [33] derived the Bayes estimator of p_s^q (with a uniform prior distribution of probabilities not adapted to most biological systems) and, recently, Hou *et al.* [34] derived ${}^2\tilde{H}_\gamma = n/(n-1)(1 - \sum_s \hat{p}_s^2)$, namely the bias correction proposed by Good [29] and Lande [10]. Bonachela *et al.* [30] proposed a balanced estimator for not too small probabilities p_s which do not follow a Poisson distribution. This may be applied to low-diversity communities. In summary, the estimation of p_s^q requires assumptions about the distribution of p_s and Grassberger's correction is recognized by all these authors as the best up-to-date for very diverse communities. Better corrections exist but are available for special values of q only, such as the recent Chao *et al.*'s estimator of Shannon entropy [35].

The correction for missing species by Chao and Shen and that for non-linearity by Grassberger ignore each other. Chao and Shen's bias correction is important when q is small and becomes negligible for $q=2$ while Grassberger's correction increases with q , vanishing for $q=0$. A rough but pragmatic estimation-bias correction is the maximum value of the two corrections. It cannot be applied when $q < 0$ (Grassberger's correction is limited to positive values of q) neither to β entropy (Chao and Shen's correction can but Grassberger's can't). An estimator of β entropy will be obtained as the difference between unbiased γ and α entropy.

We illustrate this method with a tropical forest dataset already investigated by [12]. Two 1-ha plots were fully inventoried in the Paracou field station in French Guiana. This results in 1124 individual trees (diameter at breast height over 10 cm) belonging to 229 species. Figure 1 shows diversity values calculated for q between 0 and 2, with and without correction. Chao and Shen's bias correction is inefficient for $q > 1.5$ and can even be worse than the naive estimator. In contrast, Grassberger's correction is very good for high values of q , but ignores the missed species and decreases when $q \rightarrow 0$. The maximum value offers an efficient correction. By nature, α and γ diversity values decrease with q (proof in [36]): around 300 species are estimated in the meta-community ($q=0$, Figure 1), but the equivalent number of species is only 73 for $q=2$.

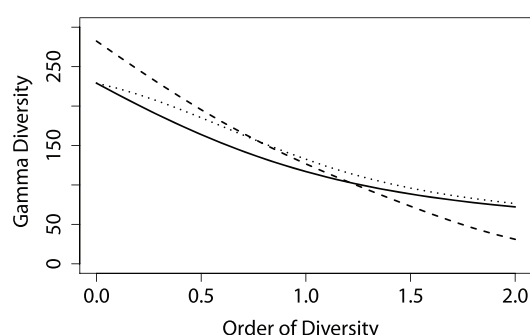


Figure 1. Profile of the γ diversity in a tropical forest meta-community. Data from French Guiana, Paracou research station, 2 ha inventoried, 1124 individual trees, and 229 observed species. Solid line: without estimation bias correction; dotted line: Grassberger correction; dashed line: Chao and Shen correction. The maximum value is our bias-corrected estimator of diversity.
doi:10.1371/journal.pone.0090289.g001

Converting unbiased entropy into diversity introduces a new bias issue because of the non-linear transformation by the deformed exponential of order q . We follow Grassberger's argument: this bias can be neglected because the transformed quantity (i.e. the entropy) is an average value (the information) over many independent terms, so it has little fluctuations (contrarily to the species probabilities whose non-linear transformation causes serious biases, as we have seen above).

We used Barro Colorado Island (BCI) tropical forest data [37] available in the vegan package [38] for R [39] to show the convergence of the estimators to the real value of diversity. 21457 trees were inventoried in a 50 hectare plot. They belong to 225 species. Only 9 species are observed a single time, so the sample coverage is over 99.99%. The inventory can be considered as almost exhaustive and used to test bias correction. We subsampled the BCI community by drawing chosen size samples (from 100 to 5000 trees) in a multinomial distribution respecting the global species frequencies. We drew 100 samples of each size, calculated their entropy, averaged it and transformed the result into diversity before plotting it in Figure 2. For low values of q , Chao and Shen's correction is the most efficient. It is close to the Chao1 estimator [40] of the number of species for $q=0$ (not shown). A correct estimation of diversity of order 0.5 is obtained with less than 1000 sampled trees (around 2 hectares of inventory). When q increases, Grassberger bias correction is more efficient: for $q=1.5$ and over, very small samples allow a very good evaluation. Both corrections are equivalent around $q=1.2$ (not shown).

Examples

Simple, theoretical example

We first propose a very simple example to visualize the decomposition of entropy. A meta-community containing 4 species is made of 3 communities C1, C2 and C3 with weights 0.5, 0.25 and 0.25. The number of individuals of each species in communities are respectively (25, 25, 40, 10), (70, 20, 10, 0), (70, 10, 0, 20). The resulting meta-community species frequencies is (0.475, 0.2, 0.225, 0.1). Note that community weights do not follow the number of individuals (100 in each community). No bias correction is necessary since the sample coverage is 1 in all cases. Entropy decomposition is plotted in Figure 3. For $q=0$, α and γ entropy equal the number of species minus 1. The meta-community's γ entropy is 3, including α entropy equal to 2.5

(the average number of species minus 1). β entropy is 0.5, equal to the averaged sum of communities contributions. C2's β entropy is negative (the total β entropy is always positive, but communities contributions can be negative).

Considering Shannon entropy, C1 is still the most diverse community (4 species versus 3 in C2 and C3, and a more equitable distribution: it has the greatest α entropy equal to 1.29). C2 and C3 have the same α entropy (their frequency distributions are identical) equal to 0.8. C3's species distribution is more different from the meta-community's than the others: it has the greatest β entropy equal to 0.34. Entropies can be transformed into diversities to be interpreted: the α diversity of communities is 3.6, 2.2 and 2.2 effective species, the total α diversity equals 2.8 effective species. The meta-community's γ diversity is 3.5 effective species (quite close to its maximum value 4 if all species were equally distributed) and β diversity is 1.2 effective communities: the same β diversity could be obtained with 1.2 theoretical, equally weighted communities with no species in common.

Real data application

We now want to compare diversity between Paracou and BCI, the two forests introduced in the previous section.

Diversity profiles are a powerful way to represent diversity of communities advocated recently by [36], as a function of the importance given to rare species which decreases with q . Comparing diversity among communities requires plotting their diversity profiles rather than comparing a single index since profiles may cross (examples from the literature are gathered in [36], Figure 2). Yet, estimation bias depends on the composition of communities, questioning the robustness of comparisons: a consistent bias correction over orders of entropy is required.

Entropy is converted to diversity and plotted against q in Figure 4 for our two forests: plots are given equal weight since they have the same size and gamma diversity is calculated for each meta-community. Paracou is more diverse, whatever the order of diversity. Bias correction allows comparing very unequally sampled forests (2 ha in Paracou versus 50 ha in BCI, sample coverage equal to 92% versus 99.99%).

β diversity profile is calculated between the two plots of Paracou. To compare it with BCI which contains 50 1-ha plots, we calculated α and β entropies between all couples of BCI plots, averaged them and converted them into β diversity (α and β entropies are required to calculate β diversity). We also calculated the 95% confidence envelope of β diversity between two 1-ha plots of BCI by eliminating the upper and lower 2.5% of the distribution of all plot couples β diversity. We chose to use Chao and Shen's correction up to $q=1.2$ and Grassberger's correction for greater q to obtain comparable results in the 1225 pairs of BCI plots. Figure 5 shows Paracou's β diversity is greater than BCI's, especially when rare species are given less importance: for $q=2$ (Simpson diversity), two plots in BCI are as different from each other as 1.2 plots with no species in common, while Paracou's equivalent number of plots is 1.7. In other words, dominant species are very different in Paracou plots, while they are quite similar on average between two BCI plots.

The shape of β diversity profiles is more complex than that of γ diversity. At $q=0$, β diversity equals the ratio between the total number of species and the average number of species in each community [7]. At $q=1$, it is the exponential of the average Kullback-Leibler divergence between communities and the meta-community. A minimum is reached between both. Over $q=1$, β diversity increases to asymptotically reach its maximum value equal to ${}^\infty D_\gamma$, i.e. the inverse of the probability of the most frequent species of the meta-community, divided by ${}^\infty D_x$, i.e. the

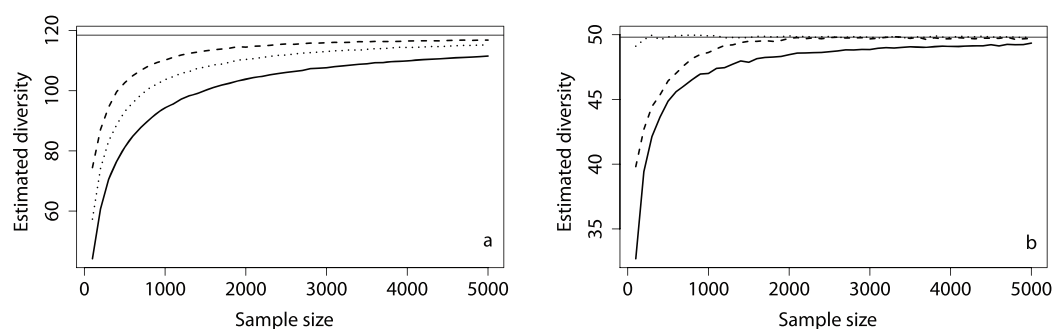


Figure 2. Efficiency of bias correction. Estimation of diversity of the BCI tropical forest plot for two values of the order of diversity q (a: 0.5, b: 1.5). The horizontal line is the actual value calculated from the whole data (around 25000 trees, species frequencies are close to a log-normal distribution). Estimated values are plotted against the sample size (100 to 5000 trees). Solid line: naive estimator with no correction; dotted line: Grassberger correction; dashed line: Chao and Shen's correction. For $q=0.5$, Chao and Shen perform best. For $q=1.5$, Grassberger's correction is very efficient even with very small samples. doi:10.1371/journal.pone.0090289.g002

inverse of the probability of the most frequent species in each community.

Discussion

Diversity can be decomposed in several ways, multiplicatively, additively or non-additively if we focus on entropy. A well-known additive decomposition of Simpson entropy is as a variance (that of Nei [41] among others). It is derived in Appendix S2. It is not a particular case of our generalization: the total variance between communities actually equals β entropy but the relative contribution of each community is different. Among these several decompositions, only the multiplicative partitioning of equally-weighted communities (17) and the non-additive partitioning of entropy (19) allow independent α and β components (except for the special case of $q=1$), but unequal weights are often necessary and ecologists may not want to restrict their studies to Shannon diversity.

We clarify here the differences between non-additive partitioning and our additive decomposition and we address the question of unequally-weighted communities.

Additive versus non-additive decomposition

Jost [3] focused on independence of the β component of the partitioning. He showed (appendix 1 of [3]) that if communities are not equally weighted the only definition of ${}^qH_\alpha$ allowing independence between α and β components is ${}^qH_\alpha = \sum_i (w_i^q / \sum_i w_i^q) {}^qH_{\alpha_i}$. The drawback of this definition is that α may be greater than γ entropy if $q \neq 1$ and community weights are not equal. Each component of entropy partitioning can be transformed into diversity as a Hill number.

We have another point of view. We rely on Patil and Taillie's concept of diversity of a mixture (section 8.3 of [14]), which implies Routledge's definition of α entropy. It does not allow independence between α and β components of the decomposition except for the special case of Shannon entropy, but it ensures that

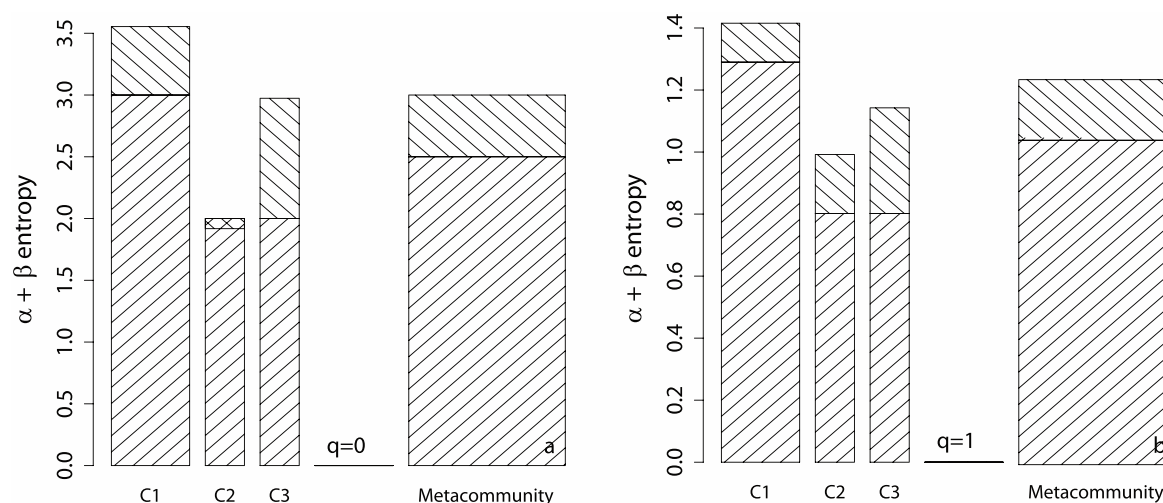


Figure 3. Decomposition of a meta-community entropy. The meta-community is made of three communities named C1, C2 and C3 (described in the text). Their α entropy ${}^qH_\alpha$ (bottom part of the bars) and their contribution to β entropy ${}^qH_\beta$ (top part of the bars) are plotted for $q=0$ (a) and $q=1$ (b). The width of bars is each community's weight. α and β entropies of the meta-community are the weighted sums of those of communities, so the area of the rectangles representing community entropies sum to the area of the meta-community's (width equal to 1). γ entropy of the meta-community is α plus β entropy. doi:10.1371/journal.pone.0090289.g003

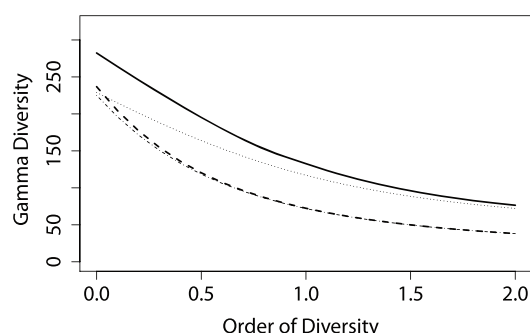


Figure 4. Paracou and BCI γ diversity. Diversity of the forest stations is compared. Solid line: Paracou with bias correction; dotted line: Paracou without bias correction; dashed line: BCI with bias correction; dotted dashed line: BCI without bias correction. Without bias correction, Paracou and BCI diversities appear to be similar for low values of q . Bias correction shows that Paracou is undersampled compared to BCI (actually around 1000 trees versus 25000). Paracou is much more diverse than BCI.
doi:10.1371/journal.pone.0090289.g004

β entropy is always positive. We believe that independence is not essential when dealing with entropy, as it emerges when converting entropy to diversity, at least when community weights are equal. The β component of the decomposition cannot be transformed into β diversity without the knowledge of α entropy but we have shown that it is an entropy, justifying the additive decomposition of Tsallis entropy.

The value of β entropy cannot be interpreted or compared between meta-communities as shown by [4], but combining α and β entropy allows calculating β diversity (Table 1).

Unequally weighted communities

Routledge's definition of α entropy does not allow independence between α and β diversity when community weights are not equal, and β diversity can exceed the number of communities [7]. We show here that the number of communities must be reconsidered to solve the second issue. We consider the independence question then.

We argue that Routledge's definition always allows to reduce the decomposition to the equal-weight case. Consider the example of Chao *et al.* [7]: two communities are weighted $w_1=0.05$ and $w_2=0.95$, their respective number of species are $S_1=100$ and $S_2=10$, no species are shared, and we focus on $q=0$ for simplicity. ${}^0D_\gamma$ equal 110 species, ${}^0D_\alpha$ is the weighted average of S_1 and S_2 equal to 14.5, so ${}^0D_\beta$ is 7.6 effective communities, which is more than the actual 2 communities. But this example is equivalent to that of a meta-community made of 1 community identical to the first one and 19 communities identical to the second one, all equally weighted. β diversity of this 20-community meta-community is 7.6 effective communities.

A more general presentation is as follows. A community of weight w can be replaced by any set of n identical communities of weights w_1, \dots, w_n provided that the sum of these weights is w , without changing α , β and γ diversity of the meta-community because of the linearity of Routledge's definition of entropy. Any unequally weighted set of community can thus be transformed into an equally weighted one by a simple transformation (strictly speaking, if weights are rational numbers).

Consider a meta-community made of several communities with no species in common, and say the smallest one (its weight is w_{\min}) is the richest (its number of species is S_{\max}). If S_{\max} is large enough,

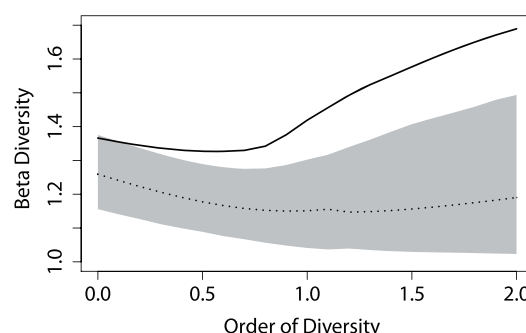


Figure 5. Paracou and BCI β diversity. β diversity profile between Paracou plots (solid line) is compared to that of any two plots of BCI (dotted line with 95% confidence envelope).
doi:10.1371/journal.pone.0090289.g005

the number of species of the meta-community is not much more than it (poor communities can be neglected). γ richness ${}^0D_\gamma$ tends to S_{\max} , ${}^0D_\alpha$ tends to $w_{\min}S_{\max}$, so ${}^0D_\beta$ tends to $1/w_{\min}$. The maximum value β diversity can reach is the inverse of the weight of the smallest community: its contribution to α diversity is proportional to its weight, but its contribution to γ diversity is its richness. Given the weights, the maximum value of β diversity is thus $1/w_{\min}$; it is the number of communities if weights are equal.

Comparing β diversity between meta-communities made of different number of communities is not possible without normalization. Jost [3] suggests normalizing it to the unit interval by dividing it by the number of communities in the equal-weight case. We suggest extending this solution to dividing β diversity by $1/w_{\min}$. When weights are not equal, the number of communities is not the appropriate reference.

Although we could come back to the equally-weighted-community partition case, β diversity is not independent of α diversity because communities are not independent of each other (some are repeated). Chao *et al.* (appendix B1 of [7]) derive the relation between the maximum value of ${}^0D_\beta$ and ${}^0D_\alpha$ for a two-community meta-community: ${}^0D_\beta \leq \frac{1}{w_{\min}} [1 - \frac{w_{\max} - w_{\min}}{{}^0D_\alpha}]$. The last term quantifies the relation between α and β diversity. It vanishes when weights are close to each other, and it decreases quickly with ${}^0D_\alpha$. If α diversity is not too low (say 50 species), the constraint is negligible (${}^0D_\beta$ can be greater than $0.98/w_{\min}$ whatever the weights).

A complete study of the dependence between α and β diversity for all q values and more than two communities is beyond the scope of this paper but these first results show that this dependence is not so serious a problem as that between α and β entropy. As long as weights are not too unequal and diversity is not too small, results can be interpreted clearly.

Very unequal weights imply lower β diversity: the extreme case is when the larger community is the richest. If it is large enough, the meta-community is essentially made of the largest community and ${}^0D_\beta$ tends to 1. This is not an issue of the measure, but a consequence of the sampling design.

Conclusion

The additive framework we proposed here has the advantage of generalizing the widely-accepted decomposition of Shannon entropy, providing a self-contained definition of β entropy and some ways to correct for estimation biases. Deformed logarithms allow a formal parallelism between HCDT and Shannon entropy

(equations (15) and (16) and Table 1). Of course, diversity can be calculated directly, but no estimation-bias correction is available then. The additive decomposition of HCDT entropy can be considered empirically as a calculation tool whose results must systematically be converted to diversity for interpretation.

We rely on Routledge's definition of α entropy which allows decomposing unequally-weighted communities and takes place in a well-established theoretical framework following Patil and Taillie. The price to pay is some dependence between α and β diversity when weights are not equal. It appears to be acceptable since it is unlikely to lead to erroneous conclusions. Still, a rigorous quantifying of it shall be the object of future research.

We only considered communities where individuals were identified and counted, such as forest inventories. Entropy decomposition remains valid when frequencies only are available but our bias correction relies entirely on the number of individual: other techniques will have to be developed for these communities if unobserved species cannot be neglected. Bias correction is still an open question. We proposed a first and rough solution. More research is needed to combine the available approaches rather than using each of them in turn.

References

- Tuomisto H (2010) A diversity of beta diversities: straightening up a concept gone awry. part 1. defining beta diversity as a function of alpha and gamma diversity. *Ecography* 33: 2–22.
- Jost L (2006) Entropy and diversity. *Oikos* 113: 363–375.
- Jost L (2007) Partitioning diversity into independent alpha and beta components. *Ecology* 88: 2427–2439.
- Jost L (2008) Gst and its relatives do not measure differentiation. *Molecular Ecology* 17: 4015–4026.
- Jost L, DeVries P, Walla T, Greeney H, Chao A, et al. (2010) Partitioning diversity for conservation analyses. *Diversity and Distributions* 16: 65–76.
- Ellison AM (2010) Partitioning diversity. *Ecology* 91: 1962–1963.
- Chao A, Chiu CH, Hsieh TC (2012) Proposing a resolution to debates on diversity partitioning. *Ecology* 93: 2037–2051.
- Havrdá J, Charvát F (1967) Quantification method of classification processes. concept of structural α -entropy. *Kybernetika* 3: 30–35.
- Tsallis C (1988) Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics* 52: 479–487.
- Lande R (1996) Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos* 76: 5–13.
- Hill MO (1973) Diversity and evenness: A unifying notation and its consequences. *Ecology* 54: 427–432.
- Marcon E, Hérault B, Baraloto C, Lang G (2012) The decomposition of shannon's entropy and a confidence interval for beta diversity. *Oikos* 121: 516–522.
- Borland L, Plastino AR, Tsallis C (1998) Information gain within nonextensive thermostatics. *Journal of Mathematical Physics* 39: 6490–6501.
- Patil GP, Taillie C (1982) Diversity as a concept and its measurement. *Journal of the American Statistical Association* 77: 548–561.
- Shannon CE (1948) A mathematical theory of communication. *The Bell System Technical Journal* 27: 379–423, 623–656.
- Simpson EH (1949) Measurement of diversity. *Nature* 163: 688.
- Kullback S, Leibler RA (1951) On information and sufficiency. *The Annals of Mathematical Statistics* 22: 79–86.
- Rao C, Nayak T (1985) Cross entropy, dissimilarity measures, and characterizations of quadratic entropy. *Information Theory, IEEE Transactions on* 31: 589–593.
- Routledge R (1979) Diversity indices: Which ones are admissible? *Journal of Theoretical Biology* 76: 503–515.
- Tsallis C, Mendes RS, Plastino AR (1998) The role of constraints within generalized nonextensive statistics. *Physica A* 261: 534–554.
- Tsallis C (1994) What are the numbers that experiments provide? *Química Nova* 17: 468–471.
- MacArthur RH (1965) Patterns of species diversity. *Biological Reviews* 40: 510–533.
- Lin J (1991) Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory* 37: 145–151.
- Lamberti PW, Majtey AP (2003) Non-logarithmic jensen-shannon divergence. *Physica A: Statistical Mechanics and its Applications* 329: 81–90.
- Furuichi S, Yanagi K, Kuriyama K (2004) Fundamental properties of tsallis relative entropy. *Journal of Mathematical Physics* 45: 4868–4877.
- Dauby G, Hardy OJ (2012) Sampled-based estimation of diversity sensu stricto by transforming hurllbert diversities into effective number of species. *Ecography* 35: 661–672.
- Chao A, Shen TJ (2003) Nonparametric estimation of shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics* 10: 429–443.
- Horvitz D, Thompson D (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47: 663–685.
- Good IJ (1953) On the population frequency of species and the estimation of population parameters. *Biometrika* 40: 237–264.
- Bonachela JA, Hinrichsen H, Muñoz MA (2008) Entropy estimates of small data sets. *Journal of Physics A: Mathematical and Theoretical* 41: 1–9.
- Grassberger P (1988) Finite sample corrections to entropy and dimension estimates. *Physics Letters A* 128: 369–373.
- Fisher RA, Corbet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology* 12: 42–58.
- Holste D, Große I, Herzel H (1998) Bayes' estimators of generalized entropies. *Journal of Physics A: Mathematical and General* 31: 2551–2566.
- Hou Y, Wang B, Song D, Cao X, Li W (2012) Quadratic tsallis entropy bias and generalized maximum entropy models. *Computational Intelligence*.
- Chao A, Wang YT, Jost L (2013) Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species. *Methods in Ecology and Evolution* 4: 1091–1100.
- Leinster T, Cobbold C (2011) Measuring diversity: the importance of species similarity. *Ecology* 93: 477–489.
- Hubbell SP, Condit R, Foster RB (2005) Barro colorado forest census plot data. Available: <https://ctfs.arnarb.harvard.edu/webatlas/datasets/bci>.
- Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, et al. *vegan: Community ecology package*. Available: <http://CRAN.R-project.org/package=vegan>.
- R Development Core Team (2013) R: A language and environment for statistical computing.
- Chao A (1984) Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* 11: 265–270.
- Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America* 70: 3321–3323.

We provide the necessary code for R to compute the analyses presented in this paper as a supplementary material in Appendix S4 with a short user's guide in Appendix S3.

Supporting Information

Appendix S1 Detailed derivation of the partitioning.

(PDF)

Appendix S2 Decomposition of Simpson index.

(PDF)

Appendix S3 Using the code: short user's guide.

(PDF)

Appendix S4 R code to compute the analyses.

(ZIP)

Author Contributions

Wrote the paper: EM. Developed the core methods: EM. Contributed to methods: IS BH VR GL. Wrote the R code: EM BH.

APPENDIX F

A Statistical Test for Ripley's Function Rejection of Poisson Null Hypothesis

Marcon, E., S. Traissac et G. Lang (2013). « A Statistical Test for Ripley's Function Rejection of Poisson Null Hypothesis ». In : ISRN Ecology 2013. Article ID 753475.

Research Article

A Statistical Test for Ripley's K Function Rejection of Poisson Null Hypothesis

Eric Marcon,¹ Stéphane Traissac,¹ and Gabriel Lang^{2,3}

¹ AgroParisTech, UMR EcoFoG, BP 709, French Guiana, 97310 Kourou, France

² AgroParisTech, UMR 518 Math. Info. Appli., 75005 Paris, France

³ INRA, UMR 518 Math. Info. Appli., 75005 Paris, France

Correspondence should be addressed to Eric Marcon; eric.marcon@ecofog.gf

Received 10 January 2013; Accepted 14 March 2013

Academic Editors: U. M. Azeiteiro and J.-P. Rossi

Copyright © 2013 Eric Marcon et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Ripley's K function is the classical tool to characterize the spatial structure of point patterns. It is widely used in vegetation studies. Testing its values against a null hypothesis usually relies on Monte-Carlo simulations since little is known about its distribution. We introduce a statistical test against complete spatial randomness (CSR). The test returns the P value to reject the null hypothesis of independence between point locations. It is more rigorous and faster than classical Monte-Carlo simulations. We show how to apply it to a tropical forest plot. The necessary R code is provided.

1. Introduction

The commonest tool used to characterize the spatial structure of a point set is Ripley's K statistic [1, 2]. It has been widely used in ecology and other scientific fields and is well referenced in handbooks [3–7]. Classically, an observed set of points is tested against a homogeneous Poisson point process taken as a null model. Since little is known about the distribution of the K function, the test of rejection of the null hypothesis relies on Monte-Carlo simulations. Large point patterns are difficult to deal with because computation time is roughly proportional to the square of the number of points (to calculate the distances between all pairs of points) multiplied by the number of simulations. Moreover, the test is valid for one distance but using it repeatedly for all distances where the K function is calculated dramatically increases the risk to reject the null hypothesis by error [8]. Duranton and Overman [9] provided a heuristic global test followed by Marcon and Puech [10]. Loosmore and Ford proposed a goodness-of-fit test to eliminate this error, but still rely on Monte-Carlo simulations.

We showed in [11] that a global test was able to return a classical P value, that is to say, the probability to erroneously reject the null model, relying on exact values of the biases

and variances of the statistics. We derived its asymptotic properties in a growing square window. We develop it in this paper so that it can be used in a rectangular window, as most applications require. We show that it can be applied to real point patterns, even with a little number of points and discuss in depth the way to employ it, so that it can be used by empirical researchers.

We first present our motivating example: a tropical forest plot where we want to elucidate the spatial structure of two species of trees. We provide the mathematical framework supporting the test. We apply it to the dataset and present the results. In the Discussion, we review the literature of previous tests to show why this one is a significant improvement and we verify that the test actually works. Finally, we provide a user guide to allow researchers to easily apply the test with the provided R [12] code.

2. Materials and Methods

2.1. Data to Analyze. We consider the distribution of two tree species in Paracou field station, French Guiana [13]. All trees over 10 cm DBH have been plotted. We use data from a 400.6 by 522.3 meters rectangle included in the four plots of the

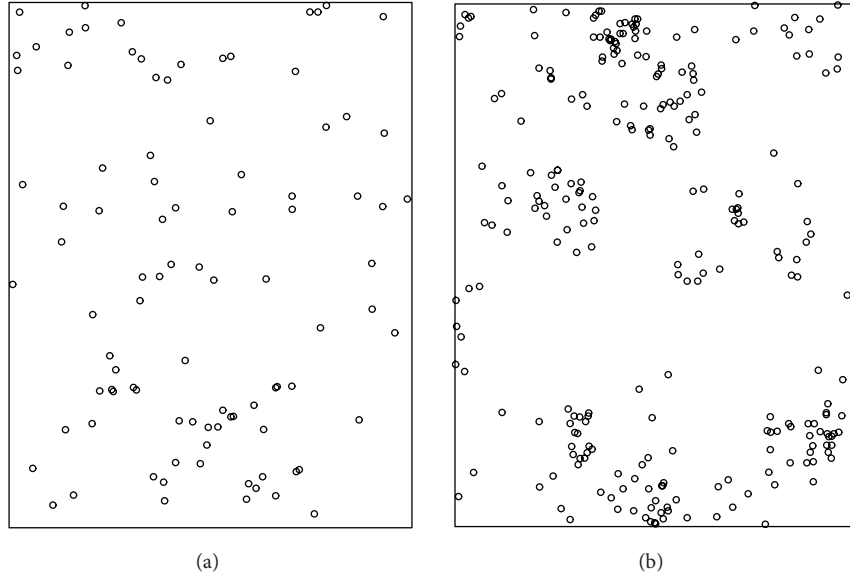


FIGURE 1: Map of *Tachigali melinonii* (94 trees, (a)) and *Dicorynia guianensis* (254 trees, (b)).

southern block of the experimental device. A map of trees is in Figure 1.

Dicorynia guianensis Amsh. is a widely studied species in French Guiana; its spatial structure has been characterized for a long time [14, 15]; as a visual inspection of the map allows to guess, *Dicorynia guianensis* is an aggregated species.

Much less is known about *Tachigali melinonii* (Harms) Barneby. The species has been studied for its special biomechanical behavior [16] or leaf trait plasticity [17]. The spatial structure of its saplings has been reported by Flores et al. [18] but the structure of adult trees is not clear.

2.2. Mathematical Framework. We consider a point pattern in a rectangular window. l_1 and l_2 are the sides of the window (width and length). ρ is the intensity of the underlying point process; estimated by the number of observed points N divided by the area of the window. We denote by \mathbf{r} the vector of distances $(r_1, \dots, r_i, \dots, r_d)$. We omit the subscript for readability when there is no ambiguity; r is one of the distances, and $\widehat{K}(r)$ is the estimator of K at distance r . Points are denoted by ξ_i , and $I\{d(\xi_i, \xi_j) \leq r\}$ is an indicator function equal to 1 when the distance between two points is less than or equal to r , 0 else. Details of the calculation are not provided as we follow exactly Lang and Marcon [11] but its important steps and intuitions are presented here.

Ripley's K function is estimated from the data for each distance r , without correction for edge effects:

$$\widehat{K}(r) = \frac{l_1 l_2}{N(N-1)} \sum_{\xi_i \neq \xi_j} I\{d(\xi_i, \xi_j) \leq r\}. \quad (1)$$

Assumptions are those of Ripley's K function: we test the independence of locations of an observed point pattern, assumed to be a realization of a homogenous point process. Homogeneity means both stationarity (the process

is unchanged by translation) and isotropy (the process is unchanged by rotation). Thus the null hypothesis of complete spatial randomness (CSR) is that the point process is a homogenous Poisson process.

The expectation of $K(r)$ under CSR is πr^2 . Edge effects introduce a bias in $\widehat{K}(r)$, computed for the null model:

$$B(r) = -\frac{4r^3(l_1 + l_2)}{3l_1 l_2} + \frac{r^4}{2l_1^2 l_2^2}. \quad (2)$$

Estimated $\widehat{K}(r)$ can be corrected for the bias of the null model to test them against it. We get a vector of results of length d :

$$\widehat{\mathbf{K}} = (\widehat{K}(r_1) - B(r_1), \widehat{K}(r_2) - B(r_2), \dots, \widehat{K}(r_d) - B(r_d)). \quad (3)$$

For a homogenous point process the vector $\widehat{\mathbf{K}} - \pi \mathbf{r}^2$ is asymptotically normal, with expectation zero and the explicit variance matrix Σ :

$$\Sigma = \begin{pmatrix} \text{Var}(\widehat{K}(r_1)) & \cdots & \text{cov}(\widehat{K}(r_1), \widehat{K}(r_d)) \\ \vdots & \ddots & \vdots \\ \text{cov}(\widehat{K}(r_1), \widehat{K}(r_d)) & \cdots & \text{Var}(\widehat{K}(r_d)) \end{pmatrix}. \quad (4)$$

Consequently $T^2 = \|\Sigma^{-1/2}(\widehat{\mathbf{K}} - \pi \mathbf{r}^2)\|^2$ follows a χ^2 law with d degrees of freedom. Asymptotic value of the variance is reached with dozens of thousand points, so it is of little use, but normality is acceptable with very few points so the test can be used with real data, as the exact value of Σ can be calculated. The exact calculation of variance and covariance is quite painful because it takes into account the edge effects, resulting in the formulas given in Appendix A.

2.3. Application. The test is applied as follows: (i) compute $\widehat{\mathbf{K}} - \pi \mathbf{r}^2$ from the observed point pattern, following (3); (ii) compute Σ according to the window's size and the number of observed points; (iii) finally compare $T^2 = \|\Sigma^{-1/2}(\widehat{\mathbf{K}} - \pi \mathbf{r}^2)\|^2$ to a χ^2 distribution with d degrees of freedom and return the P value.

We provide a classical plot of $L(r) = \sqrt{K(r)/\pi} - r$ [19] against r , computing every 5 meters up to 250 m (to illustrate the discussion). 1,000 simulations of a binomial process with the same number of points as the real data are run. At each distance r , the 25 lower and greater values are eliminated to build the local 5% confidence interval. The global confidence interval is built iteratively [9, 10]; simulations corresponding to extreme values (maximum or minimum) at any distance are eliminated. This process is repeated until 5% of the simulations are concerned. The extreme remaining values are plotted. Interpolation is used if the last iteration eliminates more simulations than required.

We apply our test to these two point sets, up to 150 meters following Collinet [15] who characterized the spatial structure of many species in Paracou and detected possible dependence at this scale. The vector of distances is $\mathbf{r} = (10, 20, \dots, 150)$ meters; we discuss this choice later. Finally, we apply Loosmore and Ford's test based on the same \mathbf{r} and 1,000 simulations.

3. Results

Aggregation of *Dicorynia guianensis* is obvious on Figure 2(b). Our test applied with the vector of distances (10, 20, ..., 150 meters) returns a P value equal to zero; that is to say, the quantile of the χ^2 distribution with 15 degrees of freedom for T is so low that R returns 0. Loosmore and Ford's test gives the same result with a value of their statistic u_1 much greater than all simulations.

Figure 2(a) shows a less clear structure of *Tachigali melinonii*. The curve leaves the local confidence interval many times, but not the global one. Our test applied with the same distance vector (10 to 150 m) returns a P value equal to 2.5%; aggregation is significant.

Loosmore and Ford's test applied to the same distance range returns a P value equal to $1.7\% \pm 0.8\%$ (u_1 is ranked 984th among the 1,000 simulations).

4. Discussion

4.1. Foundations of the Test. Many methods exist to test a point pattern against CSR [7, page 83:98], among which the relative variance [20, 21] or tests based on the nearest neighbors [22–24]. Tests based on the K function are particularly appealing because $\widehat{\mathbf{K}}$ provides data about the relative position of points at different scales, and the function simultaneously gives useful information about the point process.

4.1.1. The Distribution of $\widehat{K}(r)$ under CSR Was Unknown. The classical, local test consists in comparing each observed $\widehat{K}(r)$ to the confidence interval of $\widehat{K}(r)$ obtained by Monte-Carlo

simulations of the null model (which should be a homogeneous Poisson point process, but is usually approximated by a binomial process for simplicity, artificially reducing variability). The null model is rejected at the chosen significance level, 5%, when the observed $\widehat{K}(r)$ is out of the corresponding confidence interval.

To avoid simulations, approximate confidence intervals for $\widehat{K}(r)$ were proposed by Ripley [25], refuted by Koen [26] whose errors were finally corrected by Chiu [27]. All these confidence intervals were built on simulations.

The variance of $\widehat{K}(r)$ has been investigated early (see [5, page 58]). Asymptotic variance was calculated and asymptotic normality was proved, allowing calculating a confidence interval for $\widehat{K}(r)$, but the exact variance remains far from its asymptotic value for usual point sets [11]. We derived the exact variance in Appendix A.

4.1.2. Testing $\widehat{K}(r)$ along Many Values of r Is Not Correct. In our examples, we have 30 values of $\widehat{K}(r)$. If we draw a point pattern in a Poisson process we can expect 5% of them, that is, 1 or 2 of them, to be out of the confidence interval. As a consequence, the local significance level of the test should be decreased dramatically to have a global significance level of 5%. Actually, $\widehat{K}(r)$ are highly correlated because K is a cumulative function; roughly speaking, most of $\widehat{K}(r)$ values come from that of the previous one (see Figure 3). This reduces the need for a correction but does not eliminate it completely [28]. Since no quantification of the correction is available, the local test is used, keeping in mind that the global significance level of the test is somehow higher than announced (see [7, page 456] for a discussion). Each of the local confidence interval values is correct but testing a curve made of 30 points against local confidence envelope is not [8].

To address this issue, solutions have been proposed. Duranton and Overman [9] proposed a test consisting in eliminating simulated $\widehat{\mathbf{K}}$ vectors globally when one of their values is an extreme one. Global confidence intervals plotted in the figures are heuristic; they do not rely on any mathematical proof. They appear to be too conservative for *Tachigali melinonii*.

4.1.3. Goodness-of-Fit Tests Are a Solution. Goodness-of-fit (GoF) tests measure the discrepancy between the expected curve of K under the null hypothesis and the actual curve. This value can be compared to its quantiles under CSR. Loosmore and Ford's [8] test is a GoF test, already proposed by Diggle [4]. Its quantiles are not known so the test relies on Monte-Carlo simulations. Three GoF tests have been proposed by Heinrich [29] but all of them are asymptotic so none can be used with real data. Our test is very similar to Heinrich's χ^2 test, but as we derived the bias due to edge effects and the exact variance-covariance matrix of $\widehat{\mathbf{K}}$ under CSR, we were able to compute the T^2 statistic, which follows an χ^2 law whose quantiles are well known.

4.1.4. Graphical Interpretation of the Test. Figure 3(a) shows the correlation of values of $\widehat{\mathbf{K}} - \pi \mathbf{r}^2$ for two different values

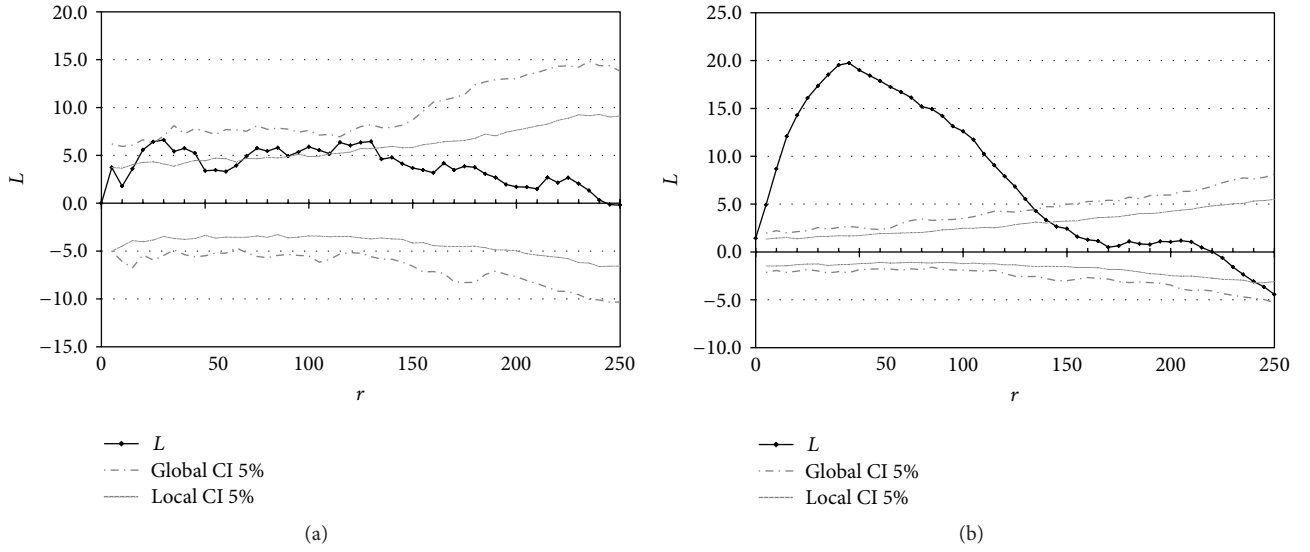


FIGURE 2: L values for *Tachigali melinonii* (a) and *Dicorynia guianensis* (b). Distances are in meters. Confidence intervals are computed for the null hypothesis of complete spatial randomness at the 5% significance level. The local [4] and global [9] confidence intervals are calculated by Monte-Carlo simulations as explained in the text. Test P values are respectively 0 and 2.5%.

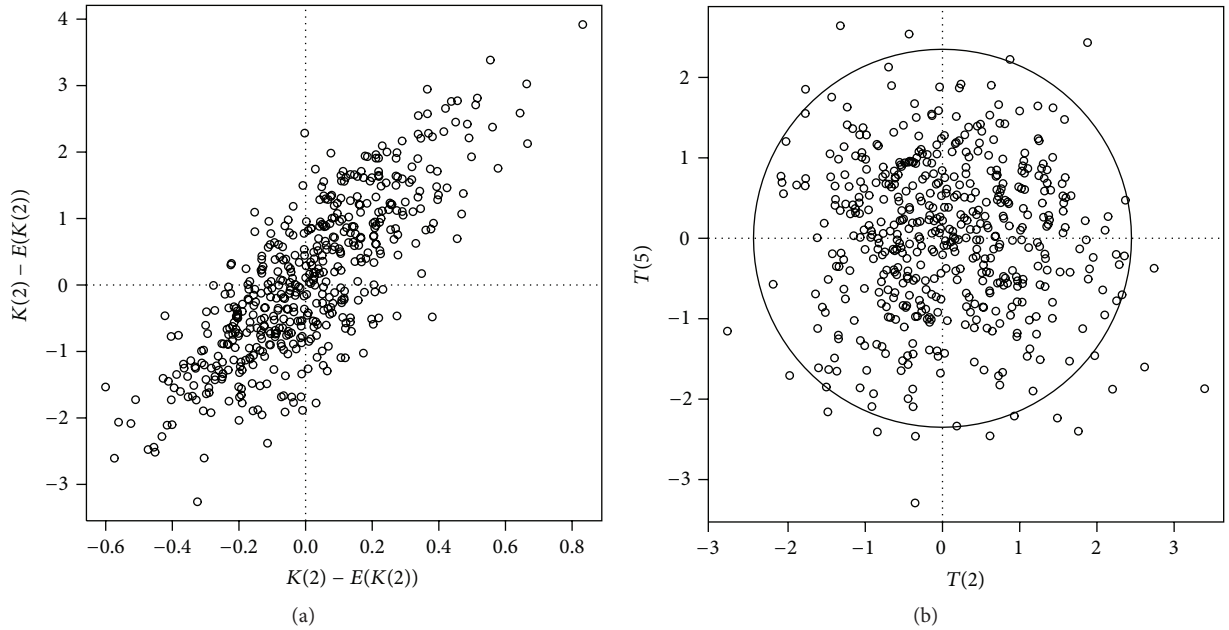


FIGURE 3: (a) Plot of \widehat{K} minus its expectation πr^2 in two dimensions ($r = 2$ and $r = 5$) for 500 simulations of a Poisson process of intensity $\rho = 5$ drawn in a square window of size 10 (500 points on average). (b) Comparison of values of $T(2)$ and $T(5)$, after transformation. Around 25 simulations of the homogenous point process out of 500 lie out of the critical circle corresponding to $T^2 > \chi_{5\%}^2(2)$ so they are rejected by the test.

of r . Each point represents a simulation of a Poisson process. The plot should be imagined in a number of dimensions d equal to the number of r values. As K is a cumulative function, its values are highly autocorrelated. Figure 3(a) presents the results of simulations of a Poisson point process. Some are slightly aggregated (positive values); others are dispersed

(negative values) due to stochasticity. Multiplying by $\Sigma^{-1/2}$ yields values of $\mathbf{T} = \Sigma^{-1/2}(\widehat{\mathbf{K}} - \pi \mathbf{r}^2)$ that are independent, centered, and of variance 1 (Figure 3(b)). We denote by $T(r)$ each element of \mathbf{T} and $T = \|\mathbf{T}\|$.

The circle's radius is the square root of the 5% critical value of an χ^2 distribution with 2 degrees of freedom. Point patterns

TABLE 1: Number of rejections of the null hypothesis (the point process is Poisson) out of 10,000 simulations of a homogenous point process in a rectangular 10 by 15 window. The significance level is 5%, so 500 simulations are expected to be rejected. The intensity varies from 1/3 to 10 so that the expected number of points varies from 50 to 1500, covering the range of usually-studied point patterns.

Expected number of points	50	100	150	300	750	1500
Number of rejections	592	587	543	537	524	543

corresponding to plots outside the circle are rejected. Thus, the test detects significant regularity of points (example not shown) as well as aggregation.

Transforming correlated \hat{K} values into independent T whose squared norm follows an χ^2 distribution relies on the exact, not asymptotic, variance matrix.

4.1.5. Correcting Edge Effects for the Null Model Is a Better Choice. Classically, edge effects are corrected when estimating K . Corrections assume unseen neighbors exist beyond the window's limits and evaluate their number according to observed data inside the window. We prefer to calculate the bias of the null model and compare empirical, uncorrected values of K to their expected, biased values; see (3). We follow Gignoux et al. [24] who showed that testing data against CSR with nearest-neighbor functions was more powerful when ignoring edge effects (i.e., neither the actual point pattern nor Monte-Carlo simulations of the null hypothesis were corrected). Heuristically, correcting edge effects for each point means adding neighbors uniformly, reducing the power of the test against CSR.

4.2. Test of the Test. Although we use the exact variance matrix, we only proved asymptotical normality. This is a common issue of statistical tests; the confidence interval of an average value calculated from 30 repeated measures is usually evaluated assuming normality, only asymptotically proved.

We evaluated the minimum number of points necessary to validate the level of the test. We simulated 10,000 realizations of a Poisson point process in a 10×15 rectangle window and tested them. \hat{K} was calculated at distances 1, 2, 3, 4, and 5. Intensity was chosen between 1/3 and 10, so that the expected number of points ranged between 50 and 1500.

Figure 4 shows the actual levels of rejections of the test (extreme intensities only are shown for readability). When points are few, the number of simulated patterns whose T value is above the critical value of χ^2 at low risk levels (e.g., 1%) is a little more than the risk level. At the 5% risk level, we believe the test results are acceptable for practical purposes even with very few points. We expect 500 simulations to be rejected; Table 1 shows the rejection level is always under 6%.

We also wanted to evaluate the power of the test, that is to say, its ability to reject point patterns that are not completely random but hard to detect. Grabarnik and Chiu [30] proposed to use a mixture of Matérn and Strauss processes as a counterfactual. They tested the ability different statistics including Ripley's K to distinguish them from a Poisson process. We followed them to draw a power test

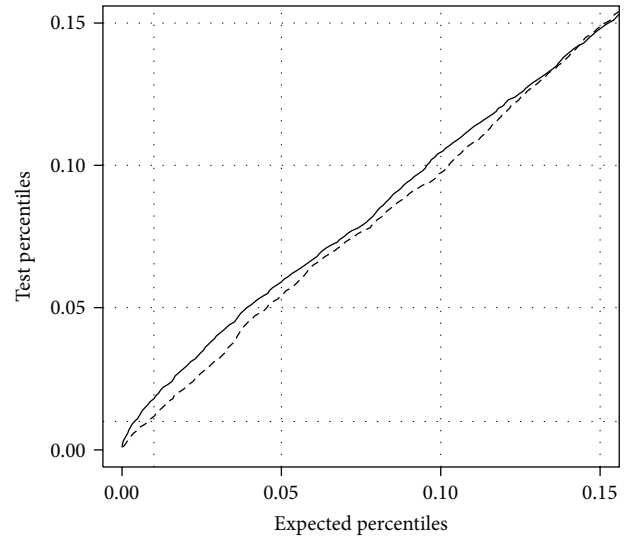


FIGURE 4: Actual rejection levels versus expected ones: ordinates are the quantiles of simulated Poisson point patterns whose P values are below the value in abscissa. Curves are built from 10,000 simulations of homogenous Poisson point processes with different intensities (solid line: 50 points expected, dashed line: 1500 points). More simulations are rejected than should be when the number of points is low, but the discrepancy is less than 1%. With 1500 points, the quantiles of the test are very close to their expected values.

presented in Appendix B. Unsurprisingly, we find that our test's power is very similar to that of K in Grabarnik and Chiu's tests.

4.3. Choosing the Distance Vector. The choice of r values (the vector of distances) is arbitrary. If the point process is actually a homogenous Poisson, results are identical whatever r is. Since it may not be, some rules should be followed; choosing the distances up to the expected range of interactions, with uniform steps, allows an "objective" analysis of the data [29], better than selecting values from the plots.

The T statistic is the sum of contributions of $T(r)$ for all r values. $T(r)$ are made independent by construction. Taking into account distances above the maximum range of interaction between points limits the power of the test since a fraction of the $T(r)$ values are purely stochastic. This is a normal behavior for a goodness-of-fit test. When used on the whole distance range 0–250, Loosmore and Ford's [8] test applied to the *Tachigali* example loses its power and returns a P value equal to 23.5%. In the same conditions, our test returns a P value equal to 6.3%; it appears to lose much less power when noisy data (there is no interaction between points at such distances) is introduced in the analysis. We chose to investigate distances up to 150 meters because we knew this was the possible range of interest.

The last question is the number of distances considered. Too many values increase stochasticity relatively to the number of points (the number of new point pairs at each new value of r gets more variable, if not often zero), while too few values do not allow to detect all scales of the pattern. In

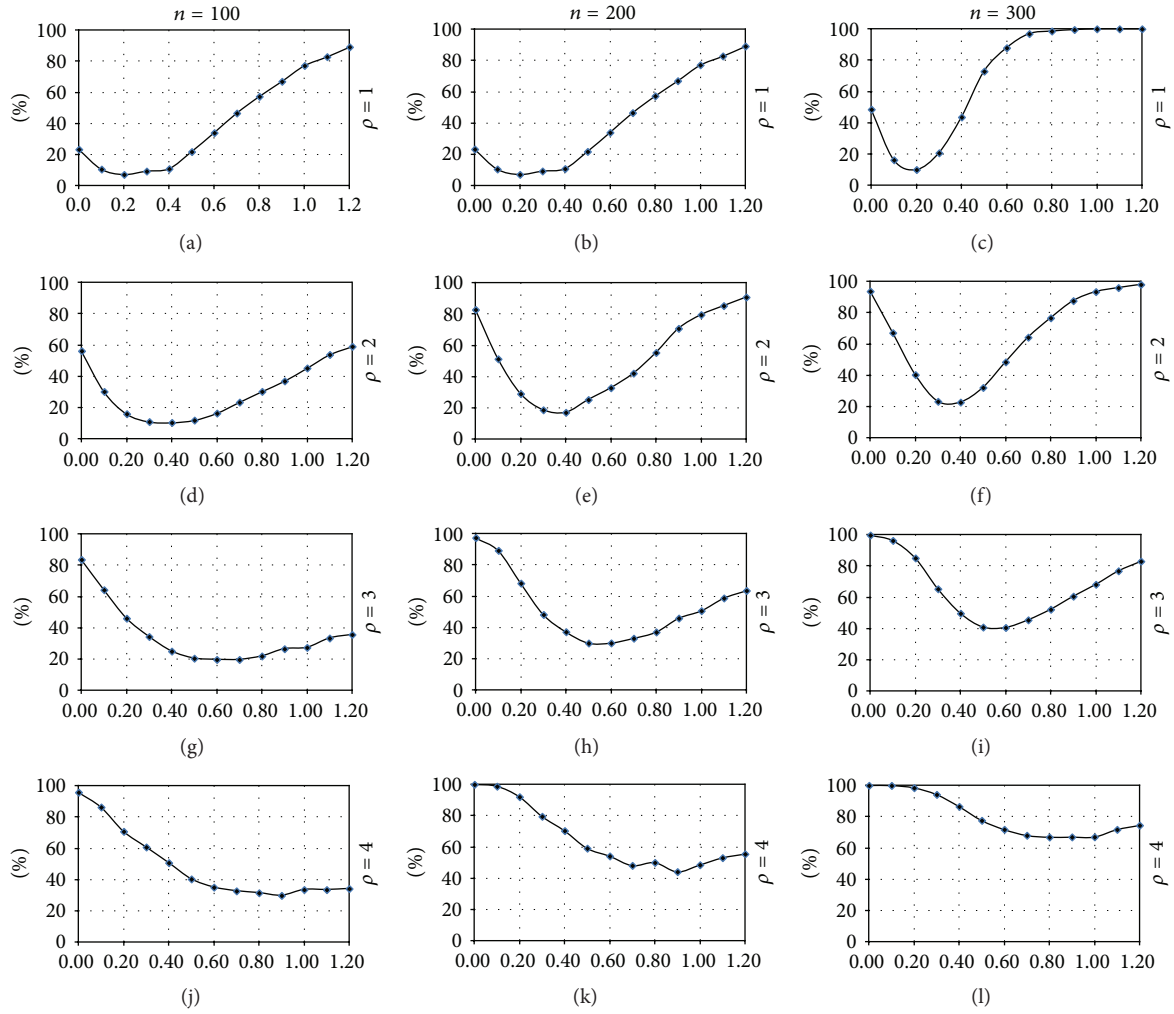


FIGURE 5: Power of the test against a mixture of clustered and repulsive point processes. n is the number of simulated points, ρ a parameter of aggregation. Values are the ratio of simulated point patterns rejected by the test as a function of β , a parameter indicating repulsion strength. See the text for complete explanations.

order to integrate all the information, steps between r values should be coherent with the processes to detect. 15 steps of 10 meters appear to be a correct way to test the structure of *Tachigali*. 5-meter steps are too small considering the density of the pattern (see Figure 1), and 20-meter steps are too large for the process we consider.

4.4. Application of the Test: A User's Guide. We provide a test for a point pattern observed in a rectangular window against CSR. The code to run it with R is available as a supplementary material available online at <http://dx.doi.org/10.1155/2013/753475>. The function *Ktest* accepts a point pattern object as defined in the *spatstat* package [31] and a vector of distances \mathbf{r} . It returns a P value (the risk of error if CSR is rejected).

Distances should be chosen up to the range of possible interactions, with equal intervals small enough to describe correctly the pattern, but too many steps if points are few. The maximum distance $\widehat{K}(r)$ is computed must be less than or

equal to half the width of the rectangle. This is a classical geometrical limitation, already faced by local edge-effect corrections [4] and [32, corrected up to half the length]. Rectangular windows only are supported.

The test does not give information on what values of \mathbf{r} are responsible for its significance. Practically, in order not to lose usual references, a graphical representation of K or, better, its transformed function L , should be provided with local confidence intervals. If the number of points is great, the number of Monte-Carlo simulations can be reduced since these intervals are not used for the test.

5. Conclusion

Characterizing the spatial structure of a dataset representing the location of plants in an experimental plot is a common task for ecologists [33]. We provide a rigorous statistical test to reject the null hypothesis that K values of an observed point pattern in a rectangle window are that of a realization

of a homogenous Poisson point process. This test replaces advantageously the classical Monte-Carlo one. It will rather complete it in practical applications since Monte-Carlo simulations provide useful local information on the point process.

The test is ready to use with the provided R code to be found in the electronic appendices.

Future work includes both supporting more complex shapes, probably triangle assemblies following Pélissier and Goreaud [34], and other point processes as null hypotheses. Although it is still limited to the simplest applications, we believe this test is an important step towards more rigorous spatial statistics, based on analytical results rather than simulations.

Appendices

A. Variance and Covariance

We consider a point pattern in a rectangular window as described in Section 2.2. r and r' are two distances the function is estimated at; r' is larger than r . $I(N > 1)$ is an indicator function equal to 1 when $N > 1$, 0 else. $\mathbb{E}(X)$ is the expectation of the random variable X .

$\widehat{K}(r)$ is the estimator of Ripley's K function at distance r . The formulas of variance of $\widehat{K}(r)$ and covariance of $\widehat{K}(r)$ and $\widehat{K}(r')$ are explicated here.

A.1. *Variance.* One has

$$\begin{aligned} \text{Var}(\widehat{K}(r)) &= 2l_1^2 l_2^2 \mathbb{E}\left(\frac{I(N > 1)}{N(N-1)}\right) (e_{r,l_1,l_2} - e_{r,l_1,l_2}^2) \\ &\quad + 4l_1^2 l_2^2 \mathbb{E}\left(\frac{[I(N > 1)](N-2)}{N(N-1)}\right) V(r, l_1, l_2) \\ &\quad + l_1^2 l_2^2 e^{-\rho l_1 l_2} (1 + \rho l_1 l_2) \\ &\quad \times (1 - e^{-\rho l_1 l_2} - \rho l_1 l_2 e^{-\rho l_1 l_2}) e_{r,l_1,l_2}^2, \end{aligned} \quad (\text{A.1})$$

where

$$e_{r,l_1,l_2} = \frac{\pi r^2}{l_1 l_2} - \frac{4r^3(l_1 + l_2)}{3l_1 l_2} + \frac{r^4}{2l_1^2 l_2^2}, \quad (\text{A.2})$$

$$\begin{aligned} V(r, n_1, n_2) &= \frac{r^5(l_1 + l_2)}{l_1^3 l_2^3} \left(\frac{8}{3}\pi - \frac{256}{45} \right) \\ &\quad + \frac{r^6}{l_1^3 l_2^3} \left(\frac{11}{48}\pi - \frac{8}{9} - \frac{16(l_1 + l_2)^2}{l_1 l_2} \right) \\ &\quad + \frac{4r^7(l_1 + l_2)}{3l_1^4 l_2^4} - \frac{r^8}{4l_1^4 l_2^4}, \end{aligned} \quad (\text{A.3})$$

e_{r,l_1,l_2} is the expectation of $K(r)$ divided by the window's area. The main term is πr^2 divided by $l_1 l_2$ and the other terms correspond to the bias due to edge effects.

ρ in (A.1) is unknown, so it is estimated by $N/(l_1 l_2)$.

A.2. *Covariance.* One has

$$\begin{aligned} \text{cov}(\widehat{K}(r), \widehat{K}(r')) &= 2l_1^2 l_2^2 \mathbb{E}\left(\frac{I(N > 1)}{N(N-1)}\right) (e_{r,l_1,l_2} - e_{r,l_1,l_2} e_{r',l_1,l_2}) \\ &\quad + 4l_1^2 l_2^2 \mathbb{E}\left(\frac{[I(N > 1)](N-2)}{N(N-1)}\right) C(r, r', l_1, l_2) \\ &\quad + l_1^2 l_2^2 e^{-\rho l_1 l_2} (1 + \rho l_1 l_2) \\ &\quad \times (1 - e^{-\rho l_1 l_2} - \rho l_1 l_2 e^{-\rho l_1 l_2}) e_{r,l_1,l_2} e_{r',l_1,l_2}, \end{aligned} \quad (\text{A.4})$$

where

$$\begin{aligned} C(r, r', l_1, l_2) &= (l_1 - 2r')(l_2 - 2r') \frac{r^2 r'^2}{l_1^3 l_2^3} b_{r,l_1,l_2} b_{r',l_1,l_2} \\ &\quad + 2(l_1 + l_2 - 4r') \frac{r^2 r'^3}{l_1^3 l_2^3} b_{r,l_1,l_2} \\ &\quad \times \int_{r/r'}^1 (b_{r',l_1,l_2} - g(x'_1)) dx'_1 \\ &\quad + 2(l_1 + l_2 - 4r') \frac{r^3 r'^2}{l_1^3 l_2^3} b_{r,l_1,l_2} \\ &\quad \times \int_0^1 (b_{r',l_1,l_2} - g\left(\frac{r x_1}{r'}\right)) (b_{r,l_1,l_2} - g(x_1)) dx_1 \\ &\quad + 4 \frac{r^2 r'^4}{l_1^3 l_2^3} \int_0^1 \int_0^1 \left[h_{A1}\left(\frac{r' x'}{r}, r\right) + h_{A2}\left(\frac{r' x'}{r}, r\right) \right. \\ &\quad \left. + h_{A3}\left(\frac{r' x'}{r}, r\right) + h_{A4}\left(\frac{r' x'}{r}, r\right) \right] \\ &\quad \times (h_{A3}(x', r') + h_{A4}(x', r)) dx'_1 dx'_2, \end{aligned} \quad (\text{A.5})$$

$$b_{r,l_1,l_2} = \pi - \frac{l_1 l_2}{r^2} e_{r,l_1,l_2} = -\frac{4r(l_1 + l_2)}{3l_1 l_2} + \frac{r^2}{2l_1 l_2}, \quad (\text{A.6})$$

$$g(x) = I(0 < x < 1) \left(\arccos x + x \sqrt{1 - x^2} \right), \quad (\text{A.7})$$

$$h_{A1}(x, r) = b_{r,l_1,l_2} I(x_1 \geq 1) I(x_2 \geq 1),$$

$$h_{A2}(x, r) = (b_{r,l_1,l_2} - g(x_2)) I(x_1 \geq 1) I(x_2 < 1)$$

$$+ (b_{r,l_1,l_2} - g(x_1)) I(x_2 \geq 1) I(x_1 < 1),$$

$$\begin{aligned}
h_{A3}(x, r) &= (b_{r, l_1, l_2} - g(x_1) - g(x_2)) \\
&\quad \times I(x_1 < 1) I(x_2 < 1) I(x_1^2 + x_2^2 \geq 1), \\
h_{A4}(x, r) &= \left(b_{r, l_1, l_2} - \frac{\pi}{4} + x_1 x_2 - \frac{g(x_1) + g(x_2)}{2} \right) \\
&\quad \times I(x_1^2 + x_2^2 \leq 1).
\end{aligned} \tag{A.8}$$

$\mathbb{E}(I(N > 1)/N(N - 1))$ and $\mathbb{E}([I(N > 1)](N - 2)/N(N - 1))$ are estimated by $1/N(N - 1)$ and $(N - 2)/N(N - 1)$ as N follows a Poisson law [11].

B. Power Test

Grabarnik and Chiu [30] proposed a new statistic (Q^2) to test data against complete spatial randomness. Q^2 is of little use in practical situations because it suffers edge effects with no correction for them. We are interested here in the power test proposed by the authors. From a mathematical point of view, mixtures of a clumped and a repulsive point process are an interesting challenge for a test built to reject Poisson processes whose K values are intermediate. From an ecological point of view, these patterns make sense if we think of the processes responsible for, say, tree locations; aggregation is expected in regeneration processes and repulsion in competition processes. For example, Aldrich et al. [35] study the relative importance of the two processes along 60 years of the life of a forest.

We drew the same simulations as a power test. n (100, 200, or 300) is the number of points of the simulated point pattern in a window of area $n/200$. Half of them are drawn in a Matérn [36] process whose centers are drawn uniformly in the window (centers are not included in the point pattern) and offsprings are drawn less than 0.06 apart from centers. The number of offsprings around each center is drawn in a Poisson law of expectation ρ (1, 2, 3, or 4). Centers are added until the number of offsprings reaches $n/2$. The other half of points is drawn in a Strauss [37] process with interaction radius 0.06 and interaction parameter β . Actually, Grabarnik and Chiu used a Gibbs process with a fixed number of points [7] where β is the pair potential function (from 0 for no interaction to 1.2 for strong repulsion). The interaction parameter in usual presentations of the Strauss process (see [38, page 85]) is $e^{-\beta}$. For each parameter set (n, ρ, β), 10,000 point sets are drawn and tested against CSR. Results are summarized in Figure 5.

Increasing the number of points (going right in the figure) improves the power of the test. Clustering increases with ρ (going down the figure) while repulsion increases with β (going right along curves). The result is the ratio of rejected simulations at a 5% risk level. It should be above 95% if the test was perfect but the point pattern is designed to be difficult to test, especially when parameters are all small (few clusters,

no repulsion; the process is close to Poisson) or intermediate (clustering of some points compensates repulsion of others).

Grabarnik and Chiu rejected CSR when the maximum departure of K from πr^2 exceeded that of 95% of a simulated binomial process of n points. A comparison between Figure 5 and their Figure 2 shows that powers are similar, even though their test is too optimistic according to Loosmore and Ford [8].

In summary, we can see here that the tendencies observed for the K function's power with Monte-Carlo simulations remain valid. K is not very efficient to disentangle mixed clustered and repulsive point processes with both high ρ and high β (Grabarnik and Chiu showed that Diggle's G is better for that purpose) but rather powerful to detect clustering.

Acknowledgment

This work has benefited from an "Investissement d'Avenir" Grant managed by Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-0025).

References

- [1] B. D. Ripley, "The second-order analysis of stationary point processes," *Journal of Applied Probability*, vol. 13, no. 2, pp. 255–266, 1976.
- [2] B. D. Ripley, "Modelling spatial patterns," *Journal of the Royal Statistical Society B*, vol. 39, no. 2, pp. 172–212, 1977.
- [3] B. D. Ripley, *Spatial Statistics*, John Wiley & Sons, New York, NY, USA, 1981.
- [4] P. J. Diggle, *Statistical Analysis of Spatial Point Patterns*, Academic Press, London, UK, 1983.
- [5] D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*, John Wiley & Sons, New York, NY, USA, 1987.
- [6] N. A. Cressie, *Statistics for Spatial Data*, John Wiley & Sons, New York, NY, USA, 1993.
- [7] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan, *Statistical Analysis and Modelling of Spatial Point Patterns*, Wiley-Interscience, Chichester, UK, 2008.
- [8] N. B. Loosmore and E. D. Ford, "Statistical inference using the G or K point pattern spatial statistics," *Ecology*, vol. 87, no. 8, pp. 1925–1931, 2006.
- [9] G. Duranton and H. G. Overman, "Testing for localization using micro-geographic data," *Review of Economic Studies*, vol. 72, no. 4, pp. 1077–1106, 2005.
- [10] E. Marcon and F. Puech, "Measures of the geographic concentration of industries: improving distance-based methods," *Journal of Economic Geography*, vol. 10, no. 5, pp. 745–762, 2010.
- [11] G. Lang and E. Marcon, "Testing randomness of spatial point patterns with the Ripley statistic," *ESAIM: Probability and Statistics*, 2013.
- [12] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013.
- [13] S. Gourlet-Fleury, J. M. Guehl, and O. Laroussinie, Eds., *Ecology & Management of a Neotropical Rainforest. Lessons Drawn from Paracou, a Long-Term Experimental Research Site in French Guiana*, Elsevier, Paris, France, 2004.
- [14] F. Goreaud, B. Courbaud, and F. Collinet, "Spatial structure analysis applied to modelling of forest dynamics: a few

- examples,” in *Proceedings of the IUFRO Workshop: Empirical and Process-Based Models for Forest Tree and Stand Growth Simulation*, A. Amaro and M. Tomé, Eds., pp. 155–172, Novas Tecnologias, Oeiras, Portugal, 1997.
- [15] F. Collinet, *Essai de regroupement des principales espèces structurantes d'une forêt dense humide d'après leur répartition spatiale (forêt de Paracou, Guyane) [Ph.D. thesis]*, Université Claude Bernard-Lyon I, Lyon, France, 1997.
- [16] G. Jaouen, M. Fournier, and T. Almeras, “Thigmomorphogenesis versus light in biomechanical growth strategies of saplings of two tropical rain forest tree species,” *Annals of Forest Science*, vol. 67, no. 2, p. 211, 2010.
- [17] S. Coste, J. C. Roggy, L. Garraud, P. Heuret, E. Nicolini, and E. Dreyer, “Does ontogeny modulate irradiance-elicited plasticity of leaf traits in saplings of rain-forest tree species? A test with *Dicorynia guianensis* and *Tachigali melinonii* (Fabaceae, Caesalpinioideae),” *Annals of Forest Science*, vol. 66, no. 7, p. 709, 2009.
- [18] O. Flores, S. Gourlet-Fleury, and N. Picard, “Local disturbance, forest structure and dispersal effects on sapling distribution of light-demanding and shade-tolerant species in a French Guianian forest,” *Acta Oecologica*, vol. 29, no. 2, pp. 141–154, 2006.
- [19] J. E. Besag, “Comments on Ripley’s paper,” *Journal of the Royal Statistical Society B*, vol. 39, no. 2, pp. 193–195, 1977.
- [20] A. R. Clapham, “Over-dispersion in grassland communities and the use of statistical methods in plant ecology,” *Journal of Ecology*, vol. 24, no. 1, pp. 232–251, 1936.
- [21] P. G. Hoel, “On indices of dispersion,” *The Annals of Mathematical Statistics*, vol. 14, no. 2, pp. 155–162, 1943.
- [22] P. J. Diggle, “On parameter estimation and goodness-of-fit testing for spatial point patterns,” *Biometrics*, vol. 35, no. 1, pp. 87–101, 1979.
- [23] M. N. M. van Lieshout and A. J. Baddeley, “A nonparametric measure of spatial interaction in point patterns,” *Statistica Neerlandica*, vol. 50, no. 3, pp. 344–361, 1996.
- [24] J. Gignoux, C. Duby, and S. Barot, “Comparing the performances of Diggle’s tests of spatial randomness for small samples with and without edge-effect correction: application to ecological data,” *Biometrics*, vol. 55, no. 1, pp. 156–164, 1999.
- [25] B. D. Ripley, “Tests of ‘randomness’ for spatial point patterns,” *Journal of the Royal Statistical Society B*, vol. 41, no. 3, pp. 368–374, 1979.
- [26] C. Koen, “Approximate confidence bounds for Ripley’s statistic for random points in a square,” *Biometrical Journal*, vol. 33, pp. 173–177, 1991.
- [27] S. N. Chiu, “Correction to Koen’s critical values in testing spatial randomness,” *Journal of Statistical Computation and Simulation*, vol. 77, no. 11–12, pp. 1001–1004, 2007.
- [28] E. Marcon and F. Puech, “Evaluating the geographic concentration of industries using distance-based methods,” *Journal of Economic Geography*, vol. 3, no. 4, pp. 409–428, 2003.
- [29] L. Heinrich, “Goodness-of-fit tests for the second moment function of a stationary multidimensional poisson process,” *Statistics*, vol. 22, no. 2, pp. 245–268, 1991.
- [30] P. Grabarnik and S. N. Chiu, “Goodness-of-fit test for complete spatial randomness against mixtures of regular and clustered spatial point processes,” *Biometrika*, vol. 89, no. 2, pp. 411–421, 2002.
- [31] A. Baddeley and R. Turner, “Spatstat: an R package for analyzing spatial point patterns,” *Journal of Statistical Software*, vol. 12, no. 6, pp. 1–42, 2005.
- [32] F. Goreaud and R. Pélissier, “On explicit formulas of edge effect correction for Ripley’s K-function,” *Journal of Vegetation Science*, vol. 10, no. 3, pp. 433–438, 1999.
- [33] R. Law, J. Illian, D. F. R. P. Burslem, G. Gratzner, C. V. S. Gunatilleke, and I. A. U. N. Gunatilleke, “Ecological information from satial patterns of plants: insights from point process theory,” *Journal of Ecology*, vol. 97, no. 4, pp. 616–628, 2009.
- [34] R. Pélissier and F. Goreaud, “A practical approach to the study of spatial structure in simple cases of heterogeneous vegetation,” *Journal of Vegetation Science*, vol. 12, no. 1, pp. 99–108, 2001.
- [35] P. R. Aldrich, G. R. Parker, J. S. Ward, and C. H. Michler, “Spatial dispersion of trees in an old-growth temperate hardwood forest over 60 years of succession,” *Forest Ecology and Management*, vol. 180, no. 1–3, pp. 475–491, 2003.
- [36] B. Matérn, “Spatial variation,” *Meddelanden från Statens Skogsforskningsinstitut*, vol. 49, no. 5, pp. 1–144, 1960.
- [37] D. J. Strauss, “A model for clustering,” *Biometrika*, vol. 62, no. 2, pp. 467–475, 1975.
- [38] J. Møller and R. P. Waagepetersen, *Statistical Inference and Simulation for Spatial Point Processes*, vol. 100, Chapman and Hall, Boca Raton, Fla, USA, 2004.

APPENDIX G

Testing randomness of spatial point patterns with the Ripley statistic

Lang, G. et E. Marcon (2013). « Testing randomness of spatial point patterns with the Ripley statistic ». In : ESAIM : Probability and Statistics 17, p. 767–788.

TESTING RANDOMNESS OF SPATIAL POINT PATTERNS WITH THE RIPLEY STATISTIC

GABRIEL LANG¹ AND ERIC MARCON²

Abstract. Aggregation patterns are often visually detected in sets of location data. These clusters may be the result of interesting dynamics or the effect of pure randomness. We build an asymptotically Gaussian test for the hypothesis of randomness corresponding to a homogeneous Poisson point process. We first compute the exact first and second moment of the Ripley K -statistic under the homogeneous Poisson point process model. Then we prove the asymptotic normality of a vector of such statistics for different scales and compute its covariance matrix. From these results, we derive a test statistic that is chi-square distributed. By a Monte-Carlo study, we check that the test is numerically tractable even for large data sets and also correct when only a hundred of points are observed.

Mathematics Subject Classification. 60G55, 60F05, 62F03.

Received June 22, 2011. Revised November 5, 2012.

INTRODUCTION

Analysis of point patterns is relevant in many sciences: cell biology, ecology or spatial economics. The observation of clusters in point locations is considered as a hint for non observable dynamics. For example the clustering of tree locations in a forest may come from better soil conditions or from spreading of seeds of a same mature individual; but clusters are also observed in random distribution as a Poisson point process sample. It is therefore essential to distinguish between clusters resulting from relevant interactions or from complete randomness. Ripley function [20, 21] is a widely used tool to quantify the structure of point patterns, especially in ecology, and is well referenced in handbooks [7, 8, 15, 18, 23, 25]. Up to a renormalization by the intensity of the process, this statistic denoted here $\hat{K}(r)$ estimates the expectation $K(r)$ of the number of neighbors at distance less than r of a point in the sample. The observed $\hat{K}(r)$ is compared to the value of $K(r)$ for a homogeneous Poisson point process chosen as a null hypothesis: the Poisson point process is characterized by an independence of point locations, modelling an absence of interactions between individuals in ecosystems. In this case $K(r)$ is simply the mean number of points in a ball of radius r divided by the intensity, that is πr^2 . If $\hat{K}(r)$ is significantly larger than πr^2 (respectively smaller), the process is considered as aggregated (respectively over-dispersed) at distance r .

Keywords and phrases. Central limit theorem, goodness-of-fit test, Hoeffding decomposition, K -function, point pattern, Poisson process, U -statistic.

¹ AgroParisTech, UMR 518 Mathématique et Informatique Appliquées, 19 avenue du Maine, 75732 Paris Cedex 15, France. gabriel.lang@agroparistech.fr

² AgroParisTech, UMR 745 Ecologie des Forêts de Guyane, Campus agronomique BP 316, 97379 Kourou Cedex, France. eric.marcon@agroparistech.fr

In order to decide if the difference is statistically significant, we build a test of the Poisson process hypothesis; we need information on the distribution of $\hat{K}(r)$ for this process. But even the variance is not known and statistical methods generally rely on Monte-Carlo simulations. Ripley [22] used them to get confidence intervals. Starting from previous results [24], he also gave critical values for the L function, a normalized version of K introduced by [4]. These critical values are valid asymptotically, for a large number of points but low intensity, so that both edge effects and point-pair dependence can be neglected. Further computations of confidence interval bands based on simulation have been proposed in [16] and corrected in [5]. But the simulation is a practical issue for large point patterns, because computation time is roughly proportional to the square of the number of points (one has to calculate the distances between all pairs of points) multiplied by the number of simulations.

We propose here to compute the exact variance of the Ripley statistic. Ward and Ferrandino [30] studied this variance. But they ignored that point pairs are not independent even though points are (Eq. A8, p. 235), thus their derivation of the variance of $\hat{K}(r)$ was erroneous. A rigorous computation of the variance has been carried out in [27] for a independent sample of uniform variables on the unit square, that is for the Poisson process conditioned by a fixed number of points; for the Poisson process, we compute the exact covariance, considering the Ripley statistic as a U -statistic as remarked in [22] and using the Hoeffding decomposition. As the variance is not enough to build a test, we study the distribution of the statistic. We prove its asymptotic normality as the size of the observation window grows. It is then easy to build an asymptotically Gaussian test.

Another concern is to test simultaneously the aggregation/dispersion at different scales. This is rarely correctly achieved in practical computations with Monte-Carlo simulations. The confidence bands or test rejection zone are often determined without taking the dependence between the numbers of neighbors at different scales into account. Heinrich [14] proposed the first multiscale goodness-of-fit tests based on the Ripley function for Poisson processes. He considered a set of scales (r_1, \dots, r_d) , computed the covariance matrix of the estimates $\hat{K}(r_i)$ and proved the asymptotic normality for the vector $(\hat{K}(r_1), \dots, \hat{K}(r_d))$. He derived Kolmogorov–Smirnov, Cramer von Mises and chi-square goodness-of-fit tests from these results. Grabarnik and Chiu [10] proposed a similar test based on the k first neighbors of a point, that is more difficult to use in practice because the number of neighbours is an additional parameter to tune. The test that we propose is very similar to the chi-square test of Heinrich; the only difference lies in the correction of the bias due to edge effects. Our method of correction allows us to compute the exact value of the covariance matrix and not only its asymptotical value, as for the Heinrich test. This is a major improvement in practice because the level of the test is very sensitive to approximations in the computation of the covariance matrix. A similar exact computation of the variance matrix is untractable for the Heinrich test: only an estimation method based on subsampling of the data may be proposed as done in [12] for the inhomogeneous case.

The paper is built as follows: Section 1 introduces the precise definition of $K(r)$ and the current definition of $\hat{K}(r)$. In Section 2, after the definition of our statistics (no edge effects correction, known or unknown intensity), we list the main results of the paper: exact bias due to the edge effects and exact variance of $\hat{K}(r)$ for a homogeneous Poisson process with known or unknown intensity; covariance between $\hat{K}(r)$ and $\hat{K}(r')$ for two different distances r and r' . The main theorem contains the convergence of the vector $(\hat{K}(r_1), \dots, \hat{K}(r_d))$ to a Gaussian distribution with explicit covariance in the following asymptotic framework: data from the same process are collected on growing squares of observation. These results allow a simple, multiscale and efficient test procedure of the Poisson process assumption. Section 3 provides a Monte-Carlo comparison of the tests and Section 4 gives our conclusions. The last section contains the proofs.

1. DEFINITION OF THE RIPLEY K -FUNCTION

We recall the characterizations of the dependence of the locations for a general point process X over \mathbb{R}^2 . We refer to the presentation of [18].

1.1. Definitions

For a point process X , define the point process $X^{(2)}$ on $\mathbb{R}^2 \times \mathbb{R}^2$ of all the couples of two different points of the original process. The intensity of this new process gives information on the simultaneous presence of points in the original process. Denote $\rho^{(2)}(x, y)$ its density (called the second-order product density). The Poisson process of density $\rho(x)$ is such that $\rho^{(2)}(x, y) = \rho(x)\rho(y)$.

The Ripley statistic is a way to estimate the density $\rho^{(2)}(x, y)$. Precisely it is an estimate of the integral on test sets of the ratio $\mathbf{g}(x, y) = \rho^{(2)}(x, y)/\rho(x)\rho(y)$. The function $\mathbf{g}(x, y)$ characterizes the fact that the points x and y appear simultaneously in the samples of X . If $\mathbf{g}(x, y) = 1$, the points appear independently. If $\mathbf{g}(x, y) < 1$, they tend to exclude each other; if $\mathbf{g}(x, y) > 1$, they appear more frequently together.

We assume the translation invariance of the point process: $\mathbf{g}(x, y) = \mathbf{g}(x - y)$. In order to estimate the function \mathbf{g} , we define its integral as the set function \mathcal{K} . Let A be a Borel set:

$$\mathcal{K}(A) = \int_A \mathbf{g}(x) dx.$$

If we also assume that the point process is isotropic, we define the Ripley K -function as

$$K(r) = \mathcal{K}(B(x, r)),$$

where $B(x, r)$ is the closed ball with center x and radius r . The translation invariance implies that $\mathcal{K}(B(x, r))$ does not depend on x . For example, if the process is a Poisson process then $\mathbf{g}(x) = 1$ and $K(r) = \pi r^2$. We define the Ripley statistic that estimates the K -function. Let A be a bounded Borel set of the plane \mathbb{R}^2 , m the Lebesgue measure, $\hat{\rho}(x)$ an estimator of the local intensity of the process and $\mathbf{I}\{\cdot\}$ denotes the indicator function of a set; for a realization S of the point process X , $S = \{X_1, \dots, X_N\}$, a general form of the Ripley statistic is

$$\hat{K}_A(r) = \frac{1}{m(A)} \sum_{X_i \neq X_j \in S} \frac{\mathbf{I}\{d(X_i, X_j) \leq r\}}{\hat{\rho}(X_i) \hat{\rho}(X_j)}.$$

Note that estimator $\hat{K}_A(r)$ refers to a preexistent estimator of the local intensity $\hat{\rho}(x)$, to make it unsensible to the inhomogeneity of the intensity. In practice, $\hat{\rho}(x)$ is a local kernel estimator, that uses the only available local information contained in the locations of neighbors in a fixed ball around the considered point of the sample. This estimator is then very much dependent of the indicator function in the numerator, because they are based on the same information. It cannot be considered as a constant close to the true value of the local intensity. This is why we do not manage to compute the exact value of the two first moments of this statistic. We only address the problem of testing homogeneous Poisson processes and the Ripley statistic has simplified expressions given below.

2. MAIN RESULTS

This section presents the theoretical results on the Ripley statistic and the resulting test.

2.1. Definitions and assumptions

Throughout the paper, we refer to the expectation $e_{r,n}$, the centered indicator function h and its conditional expectation h_1 . We gather here these definitions.

Let n be an integer; A_n denotes the square $[0, n]^2$; U is a random location in A_n with an uniform random distribution; its density is $1/n^2$ with respect to the Lebesgue measure $d\xi_1 d\xi_2$ over A_n . V is an independent copy of U . We denote $d(x, y)$ the Euclidean distance between x and y in the plane. We define $e_{r,n} = \mathbb{E}(\mathbf{I}\{d(U, V) \leq r\})$, $h(x, y, r) = \mathbf{I}\{d(x, y) \leq r\} - e_{r,n}$ and $h_1(x, r) = \mathbb{E}(h(U, V, r) | V = x)$.

We assume that X is a homogeneous Poisson process on \mathbb{R}^2 with intensity ρ . We consider that the data are available on the square A_n . The setting of the asymptotics was suggested by practitioners in ecological modeling

and forestry: the accumulation of tree location data comes from measuring wider and wider sets of land and the inter-tree distances r do not vary with n .

Let N denote the random number of observed points and $S = \{X_1, \dots, X_N\}$ denote the sample of observed points. We consider two cases:

1. If the intensity ρ is known, the Ripley statistic is expressed as

$$\hat{K}_{1,n}(r) = \frac{1}{n^2 \rho^2} \sum_{X_i \neq X_j \in S} \mathbb{I}\{d(X_i, X_j) \leq r\}.$$

2. If the intensity ρ is unknown, we use the unbiased estimator $\hat{\rho}^2 = N(N-1)/n^4$ (see [26]) and define

$$\hat{K}_{2,n}(r) = \frac{n^2}{N(N-1)} \sum_{X_i \neq X_j \in S} \mathbb{I}\{d(X_i, X_j) \leq r\}.$$

2.2. Bias

It is known that a large number of neighbors of the points located near the edges of A_n may lie outside A_n causing a bias in the estimation. We compute the bias due to this edge effect.

Proposition 2.1. *Assume that $r/n < 1/2$.*

$$\begin{aligned} \mathbb{E}\hat{K}_{1,n}(r) - K(r) &= r^2 \left(-\frac{8r}{3n} + \frac{r^2}{2n^2} \right). \\ \mathbb{E}\hat{K}_{2,n}(r) - K(r) &= r^2 \left(-\frac{8r}{3n} + \frac{r^2}{2n^2} \right) - r^2 (1 + \rho n^2) e^{-\rho n^2} \left(\pi - \frac{8r}{3n} + \frac{r^2}{2n^2} \right). \end{aligned}$$

Notes.

- The assumption $r/n < 1/2$ means that at least some balls of radius r are included in the square A_n .
- The additional term for $K_{2,n}$ corresponds to the probability to draw a sample with zero or one point in the square. This term gives a zero contribution as soon as the mean number of points ρn^2 is larger than 20.
- The proof may be adapted for a convex polygon of perimeter Ln to compute the first order term of the bias; for $u = 1$ or 2:

$$\mathbb{E}\hat{K}_{u,n}(r) - K(r) = -\frac{2Lr^2}{3} \frac{r}{n} + O\left(\frac{r^2}{n^2}\right).$$

2.3. Variance

We compute the covariance matrix of $\hat{K}_{u,n}(r)$ for $u = 1$ or 2. We get an exact computation for the variance, that can be used for any value of n .

Proposition 2.2. *For $0 < r < r'$,*

$$\begin{aligned} \text{var}(\hat{K}_{1,n}(r)) &= \frac{2e_{r,n}}{\rho^2} + \frac{4n^2 e_{r,n}^2}{\rho} + \frac{4n^2}{\rho} \mathbb{E}h_1^2(U, r), \\ \text{cov}(\hat{K}_{1,n}(r), \hat{K}_{1,n}(r')) &= \frac{2e_{r,n}}{\rho^2} + \frac{4n^2 e_{r',n} e_{r,n}}{\rho} + \frac{4n^2}{\rho} \text{cov}(h_1(U, r'), h_1(U, r)), \end{aligned}$$

$$\begin{aligned}
\text{var}(\widehat{K}_{2,n}(r)) &= 2n^4 \mathbb{E} \left(\frac{I\{N > 1\}}{N(N-1)} \right) (e_{r,n} - e_{r,n}^2) = 4n^4 \mathbb{E} \left(\frac{I\{N > 1\}(N-2)}{N(N-1)} \right) \mathbb{E} h_1^2(U, r) \\
&\quad + n^4 e^{-\rho n^2} (1 + \rho n^2) \left(1 - e^{-\rho n^2} - \rho n^2 e^{-\rho n^2} \right) e_{r,n}^2, \\
\text{cov}(\widehat{K}_{2,n}(r), \widehat{K}_{2,n}(r')) &= 2n^4 \mathbb{E} \left(\frac{I\{N > 1\}}{N(N-1)} \right) (e_{r,n} - e_{r',n} e_{r,n}) \\
&\quad + 4n^4 \mathbb{E} \left(\frac{I\{N > 1\}(N-2)}{N(N-1)} \right) \text{cov}(h_1(U, r'), h_1(U, r)) \\
&\quad + n^4 e^{-\rho n^2} (1 + \rho n^2) \left(1 - e^{-\rho n^2} - \rho n^2 e^{-\rho n^2} \right) e_{r',n} e_{r,n},
\end{aligned}$$

$$\text{where } e_{r,n} = \frac{\pi r^2}{n^2} - \frac{8r^3}{3n^3} + \frac{r^4}{2n^4} \text{ and } \mathbb{E} h_1^2(U, r) = \frac{r^5}{n^5} \left(\frac{8}{3} \pi - \frac{256}{45} \right) + \frac{r^6}{n^6} \left(\frac{11}{48} \pi - \frac{56}{9} \right) + \frac{8}{3} \frac{r^7}{n^7} - \frac{1}{4} \frac{r^8}{n^8}$$

Notes.

- The variances of both estimators are exact and can be computed with any precision, as inverse moments of the Poisson variable correspond to fast converging series.
- The covariances are not explicit because the terms $\text{cov}(h_1^2(U, r'), h_1^2(U, r))$ involve parts that have to be numerically integrated.
- The leading terms of the variances of $K_{1,n}(r)$ and $K_{2,n}(r)$ as n tends to infinity are $2\pi r^2/n^2 \rho^2 + 4\pi r^4/n^2 \rho$ and $2\pi r^2/n^2 \rho^2$.

2.4. Central Limit Theorem

We show that a normalized vector of Ripley statistics for different r converges in distribution to a normal vector. Let $\mathcal{N}(0, \Sigma)$ denote the Gaussian multivariate centred distribution with covariance matrix Σ .

Theorem 2.3. *Let d be an integer, $0 < r_1 < \dots < r_d$ a set of reals and for $u = 1$ or 2 , define $\mathcal{K}_{u,n} = (\widehat{K}_{u,n}(r_1), \dots, \widehat{K}_{u,n}(r_d))$. Then $n\sqrt{\rho}(\mathcal{K}_{u,n} - \pi(r_1^2, \dots, r_d^2))$ converges in distribution to $\mathcal{N}(0, \Sigma)$ as n tends to infinity, where for s and t in $\{1, \dots, d\}$*

- if $u = 1$, $\Sigma_{s,t} = \frac{2\pi(r_s^2 \wedge r_t^2)}{\rho} + 4\pi^2 r_s^2 r_t^2$.
- if $u = 2$, $\Sigma_{s,t} = \frac{2\pi(r_s^2 \wedge r_t^2)}{\rho}$.

Note. The first term of the variance corresponds to a case where the couples of points are independent from each others; this was used as an approximation without proof in [30]; our work proves that the actual variance and limit process are different in the first case and that the approximation holds only in the second case.

2.5. Applications to test statistics

From Theorem 2.3, we deduce that $T_u = \Sigma^{-1/2} \mathcal{K}_{u,n}$ is asymptotically $\mathcal{N}(0, I_d)$ distributed. For the hypothesis

$$H_0: X \text{ is a homogeneous Poisson process of intensity } \rho$$

we use $T^2 = \|T_u\|_2^2$ as a test statistic with rejection zone for the level α :

$$T^2 > \chi_\alpha^2(d).$$

where $\chi_\alpha^2(d)$ is the $(1 - \alpha)$ -quantile of the $\chi^2(d)$ distribution.

Note. the covariance matrix Σ depends on the intensity parameter ρ , so that in the case of the unknown parameter we have to use an estimate of ρ in the formula defining Σ .

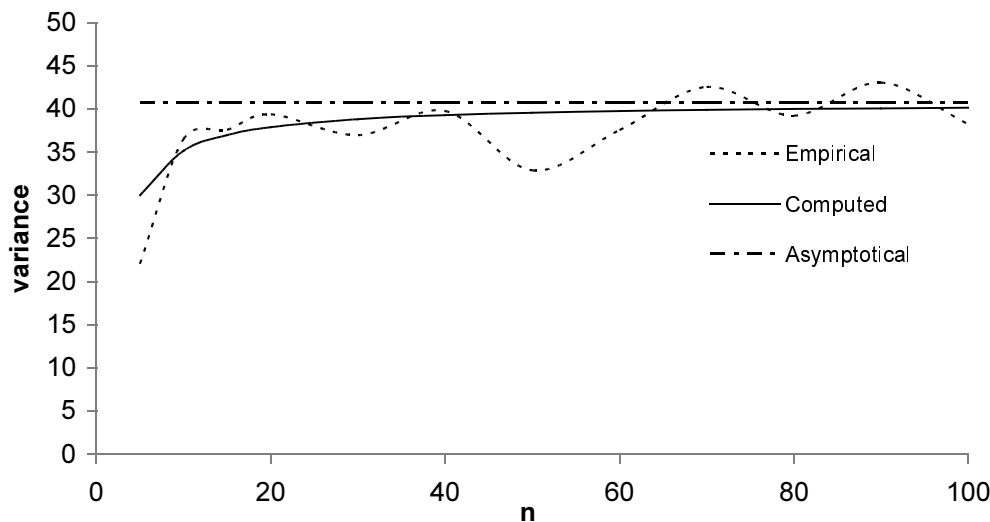


FIGURE 1. Comparison of normalized variances for $K_1(1)$, $\rho = 5$.

3. SIMULATIONS

We study the empirical variance of the proposed statistics by a Monte-Carlo simulation. Then we apply the test procedure to simulated data sets, observe the number of rejections and compare it to the level of the test.

3.1. Variance

We simulate a sample of 1000 repetitions with $\rho = 5$ and compare (after renormalization by $n\sqrt{\rho}$) the empirical variance and the exact computed variance with the limit variance for different value of n (Fig. 1). With 1000 repetitions, the oscillations of the empirical variance are still large; we will use a larger number of repetitions in the following study of the test.

The convergence of the computed variance to the limit value is not so fast and for applications with hundreds of points (corresponding in Fig. 1 to $n < 15$) the distance between the variances is still large. A preliminary study, not presented here, showed that the test procedure is perturbed by any small error in the covariance matrix, as we tried simplified versions of the covariance by ignoring the contribution of points in the corner of the observation window. It is crucial to use an accurate computation of the covariance matrix to have a correct approximation of the square root inverse matrix $\Sigma^{-1/2}$. Therefore we will use the exact variance formula instead of the asymptotic formula in the test procedure.

3.2. Test level

In the known parameter case, the computation of the test statistic T_1 is straightforward. In the unknown parameter case, the computation of the test statistic T_2 is done by replacing the unknown parameter ρ by the estimator N/n^2 . We also choose to replace the expectation $\mathbb{E}(\mathbb{I}\{N > 1\}/(N(N-1)))$ by the observed value $1/(N(N-1))$ and $\mathbb{E}(\mathbb{I}\{N > 1\}(N-2)/(N(N-1)))$ by $(N-2)/(N(N-1))$, because the dispersion of a Poisson variable is low with respect to the expectation when its parameter is large. For comparison, a chi-square test T_3 based on the unbiased Ripley estimator $\hat{K}_{3,n}$ is given using the asymptotic variance as proposed in [14]. The correction of the bias consists in dividing the indicator function not by the constant area of the

TABLE 1. Percentile of rejection over 10 000 repetitions of the test with level $\alpha = 0.05$.

Poisson			T_1	T_2	T_3
$n = 30$	$\rho = 1$	$r = (0.2, 0.5, 1)$	5.14	5.17	5.78*
$n = 10$	$\rho = 5$	$r = (0.2, 0.5, 1)$	4.66	4.74	12.31*
$n = 10$	$\rho = 5$	$r = (0.1, 0.2, \dots, 1)$	5.37	5.10	10.78*
$n = 10$	$\rho = 1$	$r = (1, 2, 5)$	5.62*	5.09	56.30*
$n = 10$	$\rho = .2$	$r = (1, 1.5, 2)$	6.74*	5.27	9.22*
$n = 10$	$\rho = .2$	$r = (0.2, 0.5, 1)$	6.47*	6.59*	7.73*

square $m(A_n) = n^2$, but by the area of the intersection of the translated squares $A_n + X_i$ and $A_n + X_j$. The corresponding unbiased estimator (see [14]) is:

$$\hat{K}_{3,n}(r) = \frac{n^4}{N(N-1)} \sum_{X_i \neq X_j \in S} \frac{\mathbb{I}\{d(X_i, X_j) \leq r\}}{m((A_n + X_i) \cap (A_n + X_j))}.$$

Concerning the choice for the range for distances r , there are two situations. From the theoretical point of view, all the scales are of the same interest. One may plot the statistic K and choose the range where the empirical values depart from expected and investigate if the difference is significative. From the practical point of view, when observing real data, practitioners often know in advance the scale they are interested in: range from 2 to 50 meters for tree locations for example, or ten meters to one kilometer for locations of shops in a city; ... Concerning the number d of different distances in a fixed range, it is theoretically not very useful to compute K for a lot of them, because K is a step function so that there is a limit to the information one gathers by refining the distances.

The test output is a Bernoulli random variable with parameter α . With a sufficient index of repetition m , the mean number of rejection is close to a normal variable with expectation α and variance $\alpha(1-\alpha)/m$. We consider that the test works correctly when the observed frequency of rejection is in the 95% Gaussian confidence interval $[\alpha - 1.96\sqrt{\alpha(1-\alpha)/m}, \alpha + 1.96\sqrt{\alpha(1-\alpha)/m}]$. With $m = 10\,000$ and $\alpha = 0.05$, the interval is $[0.0457; 0.0543]$. Percentiles of rejection in Table 1 should lie in $[4.57; 5.43]$. Stars indicate values outside this confidence interval. The performances of T_1 (known parameter ρ) are good except when the number of points is less than 100. The test T_2 (unknown parameter ρ) performs better than T_1 for small data sets. The comparison of line 5 and 6 in Table 1 shows that T_2 has a bad level if the distances are so small that the corresponding balls have a large probability to be void. The test T_3 is systematically affected by edge effects. This is due to the use of an asymptotic formula for the variance that is not sufficiently accurate even for samples with 500 points.

3.3. Test power against dependence

We investigate the power of the test T_2 against the alternative of dependent point processes. In Table 2, we simulate six Thomas cluster processes [28] and two Hardcore Strauss processes. A Thomas process is a clustered Neyman-Scott process; the germs of the clusters are drawn as a sample of a homogeneous Poisson process of intensity κ . For each germ, an inhomogeneous Poisson process is drawn with intensity measure μf , where f is the density of the Gaussian two-dimensional vector centered on the germ and with independent coordinates of standard error σ . The Thomas process results from the superimposition of these inhomogeneous Poisson processes. The germs are not conserved. Note that this process is homogeneous, with resulting intensity $\rho = \kappa\mu$. Figure 2 presents a sample of these Thomas processes compared to a sample of a Poisson process of the same intensity. It shows that the visual inspection is not sufficient to distinguish between the processes, especially when the number of points is more than 250. The Hardcore Strauss process is an over-dispersed Markov process defined by a density with respect to a homogeneous Poisson process. The density of a point set is equal to zero when two points are at distance less than a constant radius R and constant for other point sets.

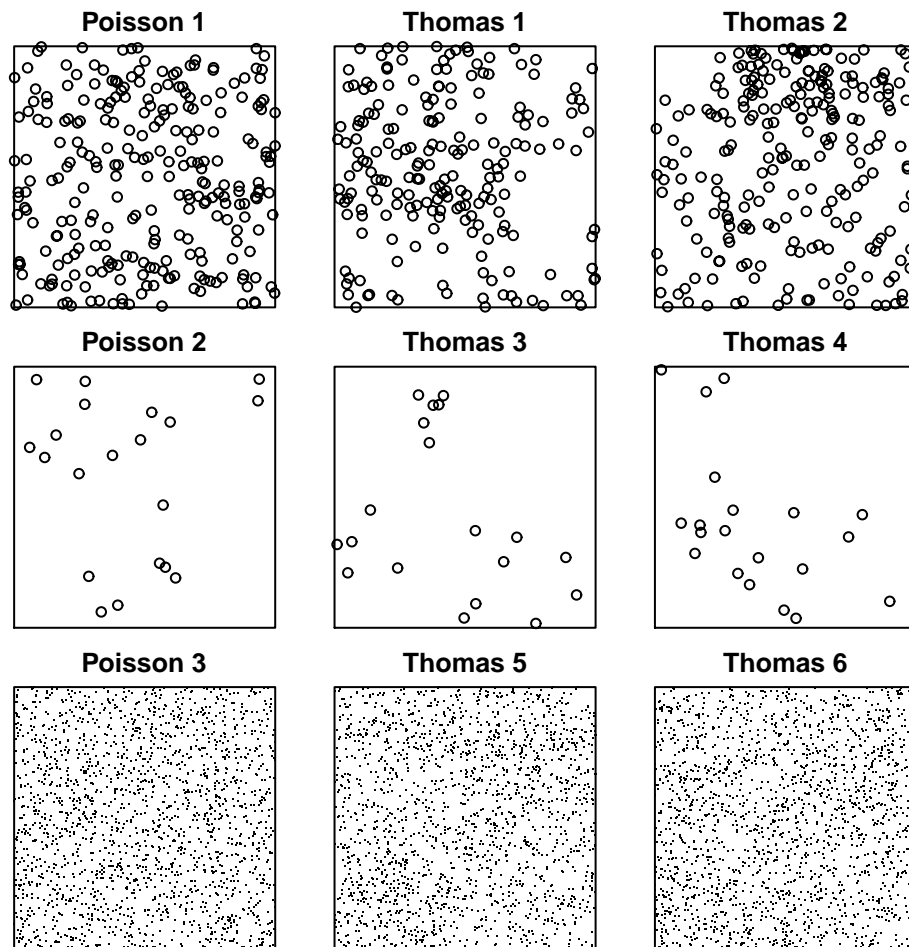


FIGURE 2. Samples of processes of table 2. Expected number of points: first row 250 , second row 20, third row 2000. For the processes Thomas 1 to 6 , the expected size of the clusters are respectively 10, 5, 4, 2, 10 and 4.

We compare test T_2 with Heinrich test T_3 and with the Monte Carlo test Lm based on the uncorrected Besag function $L(r)$ estimated in the band $r \in [0 \ 2.5]$ (see [11]); the test statistic is $Lm = \max |\hat{L}(r) - \mathbb{E}(\hat{L}(r))|$ computed on this band. The value of $\mathbb{E}(L(r))$ is estimated by a first Monte-Carlo sampling of size 10 000 and the distribution of Lm is estimated by a second Monte-Carlo sampling of the same size. The rejection zone corresponds to the largest values corresponding to 5% of the Monte-Carlo sample.

Edge effects. The first row of Table 2 shows that test T_2 rejects a bit less than Heinrich test T_3 ; this is mainly due to the incorrect level of the test T_3 as shown in the third line (28.9% rejection instead of 5% expected for the reference Poisson process). This means that edge effects are too strong for T_3 when the ratio $\max(r/n)$ is equal to 0.2. The test T_2 has a correct level, detects perfectly the large clusters of model Thomas 1 and quite well (67%) the clusters in model Thomas 2. It performs better than the Lm test. The second row displays tests with lesser edge effects ($\max(r/n)$ is equal to 0.1), for a sample with very few points (20 points expected). The third line shows that the level of T_2 and T_3 are acceptable even for small samples. The power of the two tests

TABLE 2. Percentile of rejection over 10 000 repetitions of the test with level $\alpha = 0.05$ for dependent processes compared to Poisson processes.

$n = 10$				T_2	T_3	Lm
250 points	$r = (0.5, 1, 2)$	Thomas 1	$(\kappa, \mu, \sigma) = (0.25, 10, 1)$	94.8	97.6	89.7
		Thomas 2	$(\kappa, \mu, \sigma) = (0.5, 5, 1)$	67.2	83.3	56.8
		Poisson 1	$\rho = 2.5$	4.9	28.9	5.1
20 points	$r = (0.2, 0.5, 1)$	Thomas 3	$(\kappa, \mu, \sigma) = (0.05, 4, 1)$	60.8	62.5	52.0
		Thomas 4	$(\kappa, \mu, \sigma) = (0.1, 2, 1)$	32.0	33.8	23.0
		Poisson 2	$\rho = 0.2$	6.7	7.5	5.4
2000 points	$r = (0.2, 0.5, 1)$	Thomas 5	$(\kappa, \mu, \sigma) = (2, 10, 1)$	72.8	87.5	24
		Thomas 6	$(\kappa, \mu, \sigma) = (5, 4, 1)$	32.0	63.7	16.9
		Poisson 3	$\rho = 20$	5.0	30.3	5.4
2000 points	$r = (0.5, 1, 3)$	Thomas 6	$(\kappa, \mu, \sigma) = (5, 4, 1)$	42.6		16.9
		Poisson 3	$\rho = 20$	5.3		5.4
2000 points	$r = (1, 2, 5)$	Thomas 6	$(\kappa, \mu, \sigma) = (5, 4, 1)$	46.2		16.9
		Poisson 3	$\rho = 20$	5.1		5.4
2000 points	$r = (1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5)$	Thomas 6	$(\kappa, \mu, \sigma) = (5, 4, 1)$	25.6		16.9
		Poisson 3	$\rho = 20$	5.5		5.4
10 000 points	$r = (1, 2, 5)$	Thomas 7	$(\kappa, \mu, \sigma) = (10, 10, 1)$	69.6		47.5
		Thomas 8	$(\kappa, \mu, \sigma) = (25, 4, 1)$	30.9		24.5
		Poisson 4	$\rho = 100$	5.3		5.1
69 points	$r = (0.1, 0.5, 2)$	Strauss 1	$R = 0.4$	44.2	53.2	100
		Poisson 5	$\rho = 0.69$	6.2	12.5	4.9
74 points	$r = (0.1, 0.5, 2)$	Strauss 2	$R = 0.35$	21.2	29.9	100
		Poisson 6	$\rho = 0.74$	5.4	11.9	4.8

are similar and quite good for a sample of Thomas 3 (clusters of expected size 4); they still detect around 30% of the samples of Thomas 4, that is much closer to a Poisson process (clusters of expected size 2).

Comparison with the Lm estimator. In [11], the authors claim that estimators based on maximum absolute deviation as Lm work better for small samples when edge effects are not corrected by the band correction, that discards the data in the band of width r from the edges of the square of observation. We see in row 1 and 2 that the test T_2 performs better than the Monte-Carlo test Lm , even for small sample. This is not contradictory with the conclusions of [11], as our edge effects correction does not discard data. The simulation is also different because we do not fix the number of points as it was done in [11]. For small samples the variation of the number of points is significative and may explain the different conclusions. We think that for very small sample, test Lm is advantageous because it is very easy to compute with a perfect level by construction. But for medium size sample of 250 points, T_2 is easier to compute, with a perfect level and better power.

Effect of the number of points. In row 3, we study samples with a larger number of points (2000 points expected) in the same space and with the same range of distance. First notice that is not easier to distinguish between Poisson and cluster models when the number of points increases because the clusters are forced to overlap. The performance of T_2 is a bit lower than for 250 points but still acceptable. T_3 has an incorrect level and Lm has a weak power.

Effect of the range of distance. In row 4 and 5, we change the range of distance in T_2 for the same process Thomas 6. We do not compute T_3 because its level is worse than in row 1. We see that the performance of T_2 increases when the range of distance is larger with the best result for the value $r = 5$ corresponding to the maximal ratio $r/n = 1/2$. In row 3, we see that T_2 outperforms the Lm estimator even if they are computed on a similar range of distance.

Effect of the number of distances. In row 6, we keep the same process Thomas 6 and the same range of distance but we increase the number of distances. The level of the test is correct, but the power is lower. Notice that increasing the number of distances does not necessarily increase the information (only the jumps in the function \hat{K} are informative), but it is still surprising that the performances decrease so fast. This could be a consequence of instabilities in the computation of the inverse square root of the covariance matrix as its dimension increases.

Larger set of points. In row 7, we study a small sample (100 repetitions) of processes with a larger number of points (10 000 points expected) with the largest range of distance. Comparing the three processes with cluster size equal to 4, (Thomas 3, 6 and 8), we confirm that the power decreases when the number of clusters increases. Clusters of size 10 of Thomas 7 are still well detected. Lm has a low power.

Test of over-dispersion. In row 8 and 9, we study a sample (10 000 repetitions) of hardcore Strauss processes with intensity $\beta = 1$ and hardcore radius equal to 0.4 and 0.35. The Kolmogorov Smirnov test Lm is considered as very powerful in this context [15]. We observe that Lm rejects perfectly the sample where T_2 and T_3 have poor performances. Here again, T_3 seems to have a better power, but its level is not correct as can be seen with Poisson simulation with the corresponding mean number of points.

3.4. Test power against inhomogeneity

In [13], Ho and Chiu propose to test inhomogeneity versus homogeneity with goodness-of-fit tests for the uniform distribution. A general study of those tests is [6]. The advantage is that the distribution is free of edge effects. As our test is also free of edge effects, we investigate here its power against inhomogeneous Poisson processes. Function K being the same for homogeneous and heterogeneous Poisson process, how could the method work for testing homogeneous Poisson versus heterogeneous? Simply because our definition of \hat{K} was adapted to the homogeneous case and has a different distribution under inhomogeneous Poisson assumptions.

We derive our models from those of [9]. In this paper, the authors consider five types of intensity functions $s_i(x)$ for inhomogeneous Poisson processes on the segment $[0, 1]$. We simulate inhomogeneous Poisson processes on the square $[0, 1] \times [0, 1]$ with intensity $100s_i(x)s_i(y)$. The functions $s_i(x)$ are:

$$\begin{aligned} s_1(x) &= (1 + \varepsilon) \mathbf{I}\{0 \leq x < 0.125\} + (1 - \varepsilon) \mathbf{I}\{0.125 \leq x < 0.25\} + \mathbf{I}\{0.25 \leq x \leq 1\} \\ s_2(x) &= \frac{1}{1 + 1.27\varepsilon} \left(1 + \varepsilon \sum_{j=1}^{11} h_j \mathbf{I}\{x < p_j\} \right) \\ s_3(x) &= (1 - \varepsilon) \mathbf{I}\{0 \leq x \leq 1\} + \frac{\varepsilon}{0.284} \left(\sum_{j=1}^{11} g_j \left(1 + \frac{|x - p_j|}{w_j} \right)^{-4} \right) \\ s_4(x) &= (1 - \varepsilon) \mathbf{I}\{0 \leq x < 0.75\} + (1 + 3\varepsilon) \mathbf{I}\{0.75 \leq x \leq 1\} \\ s_5(x) &= (1 - \varepsilon) + \varepsilon \beta x^{\beta-1} \end{aligned}$$

Parameters p_j , h_j , g_j and w_j are constant parameters defining the different functions (their values are the same than in [9]). Parameter ε corresponds to the strength of heterogeneity within a model. Parameter β modifies the shape of function s_5 . Functions s_1 , s_2 and s_4 are step functions, function s_3 shows steep pikes and function s_5 is smooth. All functions have integral equal to 1, so that the expected number of points in the samples is 100. The last column corresponds to the level, that is the homogeneous Poisson process with intensity 100.

The powers of the test T_2 are comparable to those of the tests proposed in [9], for the same expected number of points. They are better for models s_2 and s_5 , the same for model s_4 and worse for models s_1 and s_3 . Notice that the comparison can not be made rigorous, as the Poisson processes in [9] were defined on the line. The range of distance r has been lowered for Heinrich test T_3 to keep the level acceptable (for $r = (0.1, 0.2, 0.5)$ the observed level is 55%). Even then T_3 performs worse than T_2 . The Lm test performs better than T_2 for step functions. This may come from the fact that it uses a finer grid of distances r .

TABLE 3. Percentile of rejection over 10 000 repetitions of the test with level $\alpha = 0.05$ for 12 inhomogeneous Poisson processes and the reference homogeneous Poisson process.

model	s_1	s_1	s_1	s_2	s_2	s_3	s_3	s_3	s_4	s_4	s_5	s_5	ρ
ε	0.5	0.8	1	0.5	2	0.2	0.4	0.6	0.2	0.4	1	0.6	
β											1.5	2	
$T_2, r = (0.1, 0.2, 0.5)$	18.3	55.8	84.3	86.5	100	20.9	61.7	90.1	67.6	100	85.9	73.0	5.5
$T_3, r = (0.03, 0.05, 0.1)$	13.7	39.3	71.1	76.4	99.5	20.3	67.4	94.1	45.6	99.8	70.7	61	6.8
Lm	28.9	72.6	95.8	88.7	100	18.3	67.1	97.2	29.7	99.7	88.0	73.6	4.6

4. CONCLUSION

We provide an efficient test of the null hypothesis of a homogeneous Poisson process for point patterns in a square domain, by proposing a new correction of edge effects. Sample correction (for each point of the data) has rarely been questioned since Ripley's original paper, except by authors claiming test statistics with no correction as more powerful (see [1, 11]). Instead of correcting on each sample to reduce or cancel the bias, we compute the exact bias, so that we avoid to increase of the variance by discarding some of the observed points. The resulting test is efficient on samples with a few dozens of points as encountered in actual data sets.

This is a theoretical and practical improvement on Monte-Carlo methods as it is quicker and often more powerful. Monte-Carlo simulation of the distribution is a good method for small samples but becomes tedious when the number of points increases. Marcon and Puech [17] computed K for a 36,000-point data set (the largest ever published as far as we know), but had to limit the number of simulations to 20. With a personal computer, calculating the distribution of Lm with 10 000 simulations of a 10 000-points set is 2 days long. But it takes approximatively 3 minutes to compute T_2 for three distances with optimized C++ code and 5 minutes with a R routine [19].

Our work should be extended in two directions: to other domain shapes that are of interest for the practitioners and to 3-dimensional data for high resolution medical imaging. A further study of the asymptotics of the distribution of $\hat{K}(r)$ for dependent point process models such as Markov or Cox processes should also be achieved to inform on the power of the test.

5. PROOFS

5.1. Proof of Proposition 2.1

Let U and V be two independent uniform variables on A_n . The expectations of the Ripley statistics are

$$\begin{aligned}\mathbb{E}\hat{K}_{1,n}(r) &= \frac{1}{n^2\rho^2}\mathbb{E}\left(\sum_{X_i \neq X_j \in S} \mathbb{I}\{d(X_i, X_j) \leq r\}\right) = \frac{\mathbb{E}(N(N-1))}{n^2\rho^2}\mathbb{E}(\mathbb{I}\{d(U, V) \leq r\}) = n^2e_{r,n}.\end{aligned}$$

$$\begin{aligned}\mathbb{E}\hat{K}_{2,n}(r) &= n^2\mathbb{E}\left(\frac{1}{N(N-1)}\sum_{X_i \neq X_j \in S} \mathbb{I}\{d(X_i, X_j) \leq r\}\right) = n^2\mathbb{P}(N > 1)\mathbb{E}(\mathbb{I}\{d(U, V) \leq r\}) \\ &= n^2\left(1 - (1 + \rho n^2)e^{-\rho n^2}\right)e_{r,n}.\end{aligned}$$

The following lemma allows to conclude:

Lemma 5.1.

$$e_{r,n} = \frac{\pi r^2}{n^2} - \frac{8r^3}{3n^3} + \frac{r^4}{2n^4}.$$

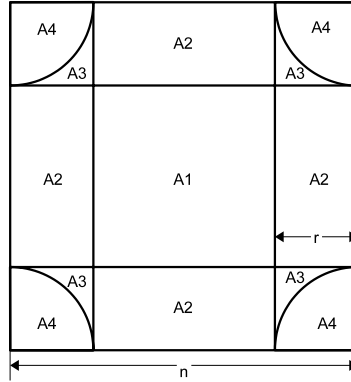


FIGURE 3. Zones in the square.

Proof. We split A_n into four parts to compute $e_{r,n}$:

$$e_{r,n} = \int_{\xi \in A_n^1} \int_{\eta \in A_n} \mathbf{I}\{d(\xi, \eta) \leq r\} \frac{1}{n^4} d\xi d\eta \quad (5.1)$$

$$+ \int_{\xi \in A_n^2} \int_{\eta \in A_n} \mathbf{I}\{d(\xi, \eta) \leq r\} \frac{1}{n^4} d\xi d\eta \quad (5.2)$$

$$+ \int_{\xi \in A_n^3} \int_{\eta \in A_n} \mathbf{I}\{d(\xi, \eta) \leq r\} \frac{1}{n^4} d\xi d\eta \quad (5.3)$$

$$+ \int_{\xi \in A_n^4} \int_{\eta \in A_n} \mathbf{I}\{d(\xi, \eta) \leq r\} \frac{1}{n^4} d\xi d\eta \quad (5.4)$$

where (see Fig. 2)

- (interior) $A_n^1 = \{\xi, \xi \text{ is at distance larger than } r \text{ from the boundary}\}$
- (one edge) $A_n^2 = \{\xi, \xi \text{ is at distance less than } r \text{ from an edge, larger than } r \text{ from the others}\}$
- (two edges) $A_n^3 = \{\xi, \xi \text{ is at distance less than } r \text{ from two edges and larger than } r \text{ from the corner}\}$
- (corner) $A_n^4 = \{\xi, \xi \text{ is at distance less than } r \text{ from the corner}\}$

Note that A_n^2 , A_n^3 and A_n^4 are composed of four parts that contribute identically. We establish formulas only for one of these parts.

Lemma 5.2. Define function $g(x) = \arccos(x) - x\sqrt{1-x^2}$.

If $\xi \in A_n^1$,

$$\int_{\eta \in A_n} \mathbf{I}\{d(\xi, \eta) \leq r\} d\eta = \pi r^2.$$

If $\xi \in A_n^2$, with $n-r < \xi_1 < n$, $x_1 = \frac{1}{r}(n - \xi_1)$,

$$\int_{\eta \in A_n} \mathbf{I}\{d(\xi, \eta) \leq r\} d\eta = r^2(\pi - g(x_1))$$

If $\xi \in A_n^3$, with $n-r < \xi_1 < n$, $n-r < \xi_2 < n$ and $(x_1, x_2) = \frac{1}{r}(n - \xi_1, n - \xi_2)$,

$$\int_{\eta \in A_n} \mathbf{I}\{d(\xi, \eta) \leq r\} d\eta = r^2(\pi - g(x_1) - g(x_2)).$$

If $\xi \in A_n^4$, with $n - r < \xi_1 < n$, $n - r < \xi_2 < n$ and $(x_1, x_2) = \frac{1}{r}(n - \xi_1, n - \xi_2)$,

$$\int_{\eta \in A_n} I\{d(\xi, \eta) \leq r\} d\eta = r^2 \left(\frac{3\pi}{4} + x_1 x_2 - \frac{g(x_1) + g(x_2)}{2} \right).$$

Note. For $0 \leq x \leq 1$, function $g(x)$ is the area of the intersection of a ball of radius 1 with a half plane, when the center of the ball lies outside the half plane at a distance x from its boundary.

Proof. Let $B(\xi, r)$ denote the ball of center ξ and radius r . For the interior points $\xi \in A_n^1$, $B(\xi, r) \subset A_n$.

Let $\xi \in A_n^2$. We compute the area of $B(\xi, r) \cap A_n$.

$$\int_{\eta \in A_n} I\{d(\xi, \eta) \leq r\} d\eta = \frac{\pi r^2}{2} + 2r^2 \int_0^{x_1} \sqrt{1-t^2} dt = r^2 \left(\pi - \arccos(x_1) + x_1 \sqrt{1-x_1^2} \right) = r^2 (\pi - g(x_1)).$$

Note that $r^2 g(x)$ is the part of the ball that lies out of the square A_n if the center is at distance xr from the edge of the square.

Let $\xi \in A_n^3$. Here the ball intersects two edges of the square and the area of $B(\xi, r) \cap A_n$ is

$$\int_{\eta \in A_n} I\{d(\xi, \eta) \leq r\} d\eta = r^2 (\pi - g(x_1) - g(x_2)).$$

Let $\xi \in A_n^4$. Divide the ball into four quarters along axes parallel to the coordinate axes. One of the quarter is inside the square, two intersect the edges, leaving outside an area equal to $(g(x_1) + g(x_2))/2$. The area of the intersection of the last quarter with the square is $x_1 x_2$ so that the area of $B(\xi, r) \cap A_n$ is

$$\int_{\eta \in A_n} I\{d(\xi, \eta) \leq r\} d\eta = r^2 \left(\frac{3\pi}{4} + x_1 x_2 - \frac{g(x_1) + g(x_2)}{2} \right). \quad \square$$

Proof of Lemma 5.1(continued). The left-hand side of (5.1) is $m(A_n^1) \pi r^2 = \pi(n-2r)^2 r^2$. Recall that A_n^2 is composed of four parts that contribute identically. Using the integration formula of the arccos function, we get the integral of g :

$$G(x) = \int_0^x g(u) du = x \arccos(x) - \sqrt{1-x^2} + \frac{1}{3}(1-x^2)^{3/2} + \frac{2}{3}.$$

Then the contribution (5.2) is equal to

$$4r \int_r^{n-r} d\xi_2 \int_0^1 r^2 (\pi - g(x)) dx = 4r^3 (n-2r) (\pi - G(1)) = \left(4\pi - \frac{8}{3}\right) r^3 (n-2r).$$

We consider A_n^3 ; the domain of integration is symmetric in (x_1, x_2) so that the contribution (5.3) is equal to

$$4r^4 \int_0^1 dx_1 \int_{\sqrt{1-x_1^2}}^1 (\pi - 2g(x_1)) dx_2 = r^4 \left(4\pi \left(1 - \frac{\pi}{4}\right) - 8G(1) + 8 \int_0^1 g(x_1) \sqrt{1-x_1^2} dx_1 \right).$$

But $\int_0^1 g(x_1) \sqrt{1-x_1^2} dx_1 = \frac{\pi^2}{16}$, so that contribution (5.3) is equal to $r^4 \left(4\pi - \frac{\pi^2}{2} - \frac{16}{3}\right)$.

We consider A_n^4 ; the contribution (5.4) is equal to

$$4r^4 \int_0^1 dx_1 \int_0^{\sqrt{1-x_1^2}} \left(\frac{3\pi}{4} + x_1 x_2 - g(x_1) \right) dx_2 = r^4 \left(\frac{3\pi^2}{4} + \frac{1}{2} - 4 \int_0^1 g(x_1) \sqrt{1-x_1^2} dx_1 \right) = r^4 \left(\frac{\pi^2}{2} + \frac{1}{2} \right).$$

Gathering the four contributions, we get

$$e_{r,n} = \frac{r^2}{n^2} \left(\pi \left(1 - \frac{2r}{n}\right)^2 + \left(4\pi - \frac{8}{3}\right) \frac{r}{n} \left(1 - \frac{2r}{n}\right) + \left(4\pi - \frac{29}{6}\right) \frac{r^2}{n^2} \right) = \frac{r^2}{n^2} \left(\pi - \frac{8}{3} \frac{r}{n} + \frac{1}{2} \frac{r^2}{n^2} \right). \quad \square$$

5.2. Proof of Proposition 2.2

For $u = 1$ or 2 , we decompose the variance of $K_{u,A_n}(r)$ by conditioning the variable with respect to the number N of points in the sample. Conditionally to N , $K_{u,A_n}(r)$ has the form of a U -statistic. Then we apply the Höfdding decomposition to this U -statistic. We use the relation

$$\text{var}(\hat{K}_{u,A_n}(r)) = \text{var} \mathbb{E}(\hat{K}_{u,A_n}(r)|N) + \mathbb{E} \text{var}(\hat{K}_{u,A_n}(r)|N).$$

We first consider the conditional expectation of $\hat{K}_{u,A_n}(r)$.

$$\begin{aligned} \mathbb{E}(\hat{K}_{1,n}(r)|N) &= \frac{1}{n^2 \rho^2} \left(\sum_{i \neq j=1}^N \mathbb{E} \mathbf{I}\{d(X_i, X_j) \leq r\} \right) = \frac{N(N-1)e_{r,n}}{n^2 \rho^2}, \\ \mathbb{E}(\hat{K}_{2,n}(r)|N) &= \frac{n^2}{N(N-1)} \sum_{i \neq j=1}^N \mathbb{E} \mathbf{I}\{d(U_i, U_j) \leq r\} = n^2 e_{r,n} \mathbf{I}\{N > 1\}. \end{aligned}$$

Because N is a Poisson variable with intensity ρn^2

$$\begin{aligned} \mathbb{E} N^2 (N-1)^2 &= \rho^4 n^8 + 4\rho^3 n^6 + 2\rho^2 n^4. \\ \text{var } N(N-1) &= 4\rho^3 n^6 + 2\rho^2 n^4. \end{aligned} \quad (5.5)$$

Then

$$\text{var} \mathbb{E}(\hat{K}_{1,n}(r)|N) = \frac{(4\rho n^2 + 2)e_{r,n}^2}{\rho^2}. \quad (5.6)$$

$$\text{var} \mathbb{E}(\hat{K}_{2,n}(r)|N) = n^4 \mathbb{P}\{N > 1\}(1 - \mathbb{P}\{N > 1\})e_{r,n}^2 = n^4 e^{-\rho n^2}(1 + \rho n^2) \left(1 - e^{-\rho n^2}(1 + \rho n^2)\right) e_{r,n}^2. \quad (5.7)$$

We compute the conditional variances.

$$\begin{aligned} \text{var}(\hat{K}_{1,n}(r)|N) &= \frac{1}{n^4 \rho^4} \text{var} \left(\sum_{i \neq j=1}^N h(X_i, X_j, r) \right), \\ \text{var}(\hat{K}_{2,n}(r)|N) &= \frac{n^4}{N^2(N-1)^2} \text{var} \left(\sum_{i \neq j=1}^N h(X_i, X_j, r) \right). \end{aligned}$$

Conditionally to N , the locations of the points are independent and uniformly distributed variables U_i over A_n . We introduce the Höfdding decomposition of the U -statistic kernel h :

$$h(x, y, r) = h_1(x, r) + h_1(y, r) + h_2(x, y, r),$$

where $h_1(x) = \mathbb{E}(h(U, V, r)|V = x)$, (U, V) being two independent uniform random variables on A_n .

Then $\mathbb{E}h_1(U, r) = 0$ and $\mathbb{E}(h_2(U, V, r)|U) = \mathbb{E}(h_2(U, V, r)|V) = 0$, so that

$$\text{var } h(U, V, r) = \text{var } h_1(U, r) + \text{var } h_1(V, r) + \text{var } h_2(U, V, r) = 2\mathbb{E}h_1^2(U, r) + \text{var } h_2(U, V, r).$$

From

$$\sum_{i \neq j=1}^N h(U_i, U_j, r) = 2(N-1) \sum_{i=1}^N h_1(U_i, r) + \sum_{i \neq j=1}^N h_2(U_i, U_j, r).$$

we get

$$\begin{aligned}
 \text{var}(\widehat{K}_{1,n}(r)|N) &= \frac{4(N-1)^2}{n^4\rho^4} \text{var}\left(\sum_{i=1}^N h_1(U_i, r)\right) + \frac{1}{n^4\rho^4} \text{var}\left(\sum_{i \neq j=1}^N h_2(U_i, U_j, r)\right) \\
 &= \frac{4N(N-1)^2}{n^4\rho^4} \mathbb{E}h_1^2(U, r) + \frac{2}{n^4\rho^4} \sum_{i \neq j=1}^N \text{var} h_2(U_i, U_j, r) \\
 &= \frac{4N(N-1)^2}{n^4\rho^4} \mathbb{E}h_1^2(U, r) + \frac{2N(N-1)}{n^4\rho^4} (\text{var} h(U, V, r) - 2\mathbb{E}h_1^2(U, r)) \\
 &= \frac{4N(N-1)(N-2)}{n^4\rho^4} \mathbb{E}h_1^2(U, r) + \frac{2N(N-1)}{n^4\rho^4} \text{var} h(U, V, r),
 \end{aligned}$$

Note that the factor 2 in the second line may be surprising in the variance of a sum of independent variables, but each variance term appears four times in the expansion of the variance of the sum over $i \neq j$. Now $\text{var} h(U, V, r) = e_{r,n} - e_{r,n}^2$ and using factorial moments of the Poisson distribution

$$\mathbb{E} \text{var}(\widehat{K}_{1,n}(r)|N) = \frac{4n^2}{\rho} \mathbb{E}h_1^2(U, r) + \frac{2}{\rho^2} (e_{r,n} - e_{r,n}^2). \quad (5.8)$$

Lemma 5.3 gives the exact value of $\mathbb{E}h_1^2(U, r)$. Its proof is postponed at the end of the paper.

Lemma 5.3.

$$\mathbb{E}h_1^2(U, r) = \frac{r^5}{n^5} \left(\frac{8}{3} \pi - \frac{256}{45} \right) + \frac{r^6}{n^6} \left(\frac{11}{48} \pi - \frac{56}{9} \right) + \frac{8}{3} \frac{r^7}{n^7} - \frac{1}{4} \frac{r^8}{n^8}.$$

With relations (5.6) and (5.8), we get

$$\begin{aligned}
 \text{var}(\widehat{K}_{1,n}(r)) &= \frac{2e_{r,n}}{\rho^2} + \frac{4n^2 e_{r,n}^2}{\rho} + \frac{4n^2}{\rho} \mathbb{E}h_1^2(U_j, r) \\
 &= \frac{1}{n^2} \left(\frac{2\pi r^2}{\rho^2} + \frac{4\pi^2 r^4}{\rho} \right) - \frac{1}{n^3} \left(\frac{16}{3} \frac{r^3}{\rho^2} + \left(\frac{32\pi}{3} + \frac{1024}{45} \right) \frac{r^5}{\rho} \right) + \frac{1}{n^4} \left(\frac{r^4}{\rho^2} + \left(\frac{59\pi}{12} + \frac{32}{9} \right) \frac{r^6}{\rho} \right).
 \end{aligned}$$

Similarly

$$\begin{aligned}
 \text{var}(\widehat{K}_{2,n}(r)|N) &= \frac{4n^4 \mathbb{I}\{N > 1\}(N-2)}{N(N-1)} \mathbb{E}h_1^2(U, r) + \frac{2n^4 \mathbb{I}\{N > 1\}}{N(N-1)} \text{var} h(U, V, r), \\
 \mathbb{E} \text{var}(\widehat{K}_{2,n}(r)|N) &= 4n^4 \mathbb{E} \left(\frac{\mathbb{I}\{N > 1\}(N-2)}{N(N-1)} \right) \mathbb{E}h_1^2(U, r) + 2n^4 \mathbb{E} \left(\frac{\mathbb{I}\{N > 1\}}{N(N-1)} \right) (e_{r,n} - e_{r,n}^2).
 \end{aligned}$$

From this and relation (5.7), we get

$$\begin{aligned}
 \text{var}(\widehat{K}_{2,n}(r)) &= 2n^4 \mathbb{E} \left(\frac{\mathbb{I}\{N > 1\}}{N(N-1)} \right) (e_{r,n} - e_{r,n}^2) + 4n^4 \mathbb{E} \left(\frac{\mathbb{I}\{N > 1\}(N-2)}{N(N-1)} \right) \mathbb{E}h_1^2(U_j, r) \\
 &\quad + n^4 e^{-\rho n^2} (1 + \rho n^2) \left(1 - e^{-\rho n^2} - \rho n^2 e^{-\rho n^2} \right) e_{r,n}^2.
 \end{aligned}$$

We now apply the same decomposition to $\text{cov}(\widehat{K}_{1,n}(r), \widehat{K}_{1,n}(r'))$,

$$\text{cov}(\mathbb{E}(\widehat{K}_{1,n}(r')|N), \mathbb{E}(\widehat{K}_{1,n}(r)|N)) = \frac{(4\rho n^2 + 2)e_{r',n}e_{r,n}}{\rho^2}. \quad (5.9)$$

$$\begin{aligned}
\text{cov}(\widehat{K}_{1,n}(r'), \widehat{K}_{1,n}(r)|N) &= \frac{4(N-1)^2}{n^4 \rho^4} \text{cov} \left(\sum_{i=1}^N h_1(U_i, r'), \sum_{i=1}^N h_1(U_i, r) \right) \\
&\quad + \frac{1}{n^4 \rho^4} \text{cov} \left(\sum_{i \neq j=1}^N h_2(U_i, U_j, r'), \sum_{i \neq j=1}^N h_2(U_i, U_j, r) \right) \\
&= \frac{4N(N-1)(N-2)}{n^4 \rho^4} \text{cov}(h_1(U, r'), h_1(U, r)) \\
&\quad + \frac{2N(N-1)}{n^4 \rho^4} \text{cov}(h(U, V, r'), h(U, V, r)).
\end{aligned}$$

$$\mathbb{E} \text{cov}(\widehat{K}_{1,n}(r'), \widehat{K}_{1,n}(r)|N) = \frac{4n^2}{\rho} \text{cov}(h_1(U, r'), h_1(U, r)) + \frac{2}{\rho^2} (e_{r,n} - e_{r',n} e_{r,n})$$

To compute $\text{cov}(h_1(U, r'), h_1(U, r))$, the square A_n should now be split into 16 different zones according to the 4 zones of the preceding section with respect to r and the 4 zones with respect to r' . Because of inclusions, the actual number of zones to consider reduces to 9. The corresponding computation is easy in the center zone, but can not be achieved in a close form in the edge bands and in the corner. We consider the following zones:

- (interior) $A_n^{1,1} = \{\xi, \xi \text{ is at distance larger than } r' \text{ from the boundary}\}$,
- (interior-edge) $A_n^{1,2} = \{\xi, \xi \text{ is at distance between } r \text{ and } r' \text{ from an edge, larger than } r' \text{ from the others}\}$,
- (edge) $A_n^{2,2} = \{\xi, \xi \text{ is at distance less than } r \text{ from an edge, larger than } r' \text{ from the others}\}$,
- (corner) $A_n^{3,3} = \{\xi, \xi \text{ is at distance less than } r' \text{ from two edges}\}$.

Denoting $x_1 = \frac{1}{r}(n - \xi_1)$ and $x'_1 = \frac{1}{r'}(n - \xi_1)$ we get

$$\begin{aligned}
h_1(X_j, r') h_1(X_j, r) &= \left(\frac{\pi r'^2}{n^2} - e_{r',n} \right) \left(\frac{\pi r^2}{n^2} - e_{r,n} \right) \text{ on } A_n^{1,1}, \\
&= \left(\frac{\pi r'^2}{n^2} - e_{r',n} - \frac{r'^2}{n^2} g(x'_1) \right) \left(\frac{\pi r^2}{n^2} - e_{r,n} \right) \text{ on } A_n^{1,2}, \\
&= \left(\frac{\pi r'^2}{n^2} - e_{r',n} - \frac{r'^2}{n^2} g(x'_1) \right) \left(\frac{\pi r^2}{n^2} - e_{r,n} - \frac{r^2}{n^2} g(x_1) \right) \text{ on } A_n^{2,2}.
\end{aligned}$$

Denote $b_{r,n} = \left(\pi - \frac{n^2}{r^2} e_{r,n} \right) = \frac{8r}{3n} - \frac{r^2}{2n^2}$.

$$\text{cov}(h_1(X_j, r'), h_1(X_j, r)) = C(A_n^{1,1}) + C(A_n^{1,2}) + C(A_n^{2,2}) + C(A_n^{3,3})$$

$$\begin{aligned}
C(A_n^{1,1}) &= \frac{r'^2 r^2}{n^4} \left(1 - \frac{2r'}{n} \right)^2 b_{r',n} b_{r,n} \\
C(A_n^{1,2}) &= 4 \left(1 - \frac{2r'}{n} \right) \frac{r'^3 r^2}{n^5} b_{r,n} \int_{r/r'}^1 (b_{r',n} - g(x'_1)) dx'_1 \\
C(A_n^{2,2}) &= 4 \left(1 - \frac{2r'}{n} \right) \frac{r^3 r'^2}{n^5} \int_0^1 (b_{r',n} - g(r x_1 / r')) (b_{r,n} - g(x_1)) dx_1.
\end{aligned}$$

The first integral may be expressed in terms of function G , the second integral is elliptic and has to be numerically evaluated; as the integrand is bounded and very smooth this can be achieved without difficulties. To compute

the term $C(A_n^{3,3})$, we rewrite the different values of function h_1 with the help of indicator functions:

$$\begin{aligned} h_{A1}(x, r) &= b_{r,n} \mathbf{I}\{x_1 \geq 1; x_2 \geq 1\} \\ h_{A2}(x, r) &= (b_{r,n} - g(x_2)) \mathbf{I}\{x_1 \geq 1; x_2 < 1\} + (b_{r,n} - g(x_1)) \mathbf{I}\{x_2 \geq 1; x_1 < 1\} \\ h_{A3}(x, r) &= (b_{r,n} - g(x_1) - g(x_2)) \mathbf{I}\{x_1 < 1; x_2 < 1; x_1^2 + x_2^2 \geq 1\} \\ h_{A4}(x, r) &= (b_{r,n} - \pi/4 + x_1 x_2 - (g(x_1) + g(x_2))/2) \mathbf{I}\{x_1^2 + x_2^2 < 1\} \end{aligned}$$

$$\text{For } x' = \frac{1}{r'}(n - \xi_1, n - \xi_2), \quad C(A_n^{3,3}) = 4 \frac{r^2 r'^4}{n^6} \int_0^1 \int_0^1 \sum_{i=1}^4 h_{Ai}(r' x' / r, r) \times \sum_{i=3}^4 h_{Ai}(x', r') dx'_1 dx'_2$$

and this integral also can be numerically evaluated.

Note. The whole computation of this term of the covariance could be numerically achieved, but we gain some useful precision with an exact computation whenever it is possible.

The case of the covariance of $K_{2,n}(r)$ is analogous:

$$\begin{aligned} \text{cov}(\mathbb{E}(\widehat{K}_{2,n}(r')|N), \mathbb{E}(\widehat{K}_{2,n}(r)|N)) &= n^4 e^{-\rho n^2} (1 + \rho n^2) (1 - e^{-\rho n^2} (1 + \rho n^2)) e_{r',n} e_{r,n}. \\ \mathbb{E} \text{ cov}(\widehat{K}_{2,n}(r'), \widehat{K}_{2,n}(r)|N) &= 4n^4 \mathbb{E} \left(\frac{\mathbf{I}\{N > 1\}(N-2)}{N(N-1)} \right) \text{cov}(h_1(U, r'), h_1(U, r)) \\ &\quad + 2n^4 \mathbb{E} \left(\frac{\mathbf{I}\{N > 1\}}{N(N-1)} \right) (e_{r,n} - e_{r',n} e_{r,n}). \end{aligned}$$

5.3. Proof of Theorem 2.3

We show that any linear combination of the $K_{1,n}(r_t)$ is asymptotically normal. Let $\Lambda = (\lambda_1, \dots, \lambda_d)$ be a vector of real coefficients. Define $Z_1 = \sum_{t=1}^d \lambda_t K_{1,n}(r_t)$. We use the Bernstein blocks technique [3]: we divide the square A_n into squares of side p with $p = o(n)$. These squares are separated by gaps of width $2r_d$ so that the sums over couples of points in each square are independent. The couples of points with at least one point in the gaps give a negligible contribution, so that the statistic Z_1 is equivalent to a sum of independent variables and asymptotically normal.

Set $p = n^{1/4}$. Assume that the Euclidean division of n by $(p + 2r_d)$ gives a quotient a and a remainder q . For $l = 0, \dots, a$, we define the segment $I_l = [(p + 2r_d)l, (p + 2r_d)l + p - 1]$. We order the set $\{0, \dots, a\}^2$ by the lexicographic order. To any integer i such that $1 \leq i \leq k = (a + 1)^2$, corresponds an element (j_1, j_2) of this set; we define the block $P_{i,n} = I_{j_1} \times I_{j_2}$ and $Q = A_n \setminus \cup_i P_{i,n}$ the set of points that are in none of the $P_{i,n}$'s. For each block $P_{i,n}$ and Q , we define the partial sums:

$$\begin{aligned} u_{i,n} &= \frac{1}{n\rho^{3/2}} \sum_{X_l \neq X_m \in P_{i,n}} \sum_{t=1}^d \lambda_t \mathbf{I}\{d(X_l, X_m) \leq r_t\}, \\ v_{i,n} &= \frac{1}{n\rho^{3/2}} \sum_{X_l \in P_{i,n}, X_m \in Q} \sum_{t=1}^d \lambda_t \mathbf{I}\{d(X_l, X_m) \leq r_t\} \\ w_n &= \frac{1}{n\rho^{3/2}} \sum_{X_l \neq X_m \in Q} \sum_{t=1}^d \lambda_t \mathbf{I}\{d(X_l, X_m) \leq r_t\}. \end{aligned}$$

then

$$n\sqrt{\rho}(Z_1 - \mathbb{E}Z_1) = \sum_{i=1}^k (u_{i,n} - \mathbb{E}u_{i,n}) + \sum_{i=1}^k (v_{i,n} - \mathbb{E}v_{i,n}) + w_n - \mathbb{E}w_n,$$

We show that the sum of the $u_{i,n}$ converges in distribution to a Gaussian variable and that the other term are negligible in L^2 . We check the conditions of the following CLT adapted from [2].

Theorem 5.4. Let $(z_{i,n})_{0 \leq i \leq k(n)}$ be an array of random variables satisfying

1. There exists $\delta > 0$ such that $\sum_{i=0}^{k(n)} \mathbb{E}|z_{i,n}|^{2+\delta}$ tends to 0 as n tends to infinity,
2. $\sum_{i=0}^{k(n)} \text{var } z_{i,n}$ tends to σ^2 as n tends to infinity,

then $\sum_{i=0}^{k(n)} z_{i,n}$ tends in distribution to $\mathcal{N}(0, \sigma^2)$ as n tends to infinity.

To check Condition 1, we compute the fourth order moment of $u_{i,n} - \mathbb{E}u_{i,n}$. Let N_i be the number of points of S that fall in $P_{i,n}$. Denote $f(x, y) = \sum_{t=1}^d \lambda_t (\mathbb{I}\{d(x, y) \leq r_t\} - e_{r,p}) = \sum_{t=1}^d \lambda_t h(x, y, r_t)$, then

$$\mathbb{E}((u_{i,n} - \mathbb{E}u_{i,n})^4 | N_i) = \frac{1}{n^4 \rho^6} \mathbb{E} \left(\sum_{l \neq m=1}^{N_i} f(U_l, U_m) \right)^4$$

Denote f_1 and f_2 the decomposing functions of f : $\mathbb{E}(f_1(U_l)) = 0$, $\mathbb{E}(f_1(U_l)f_2(U_l, U_m)) = \mathbb{E}(f_1(U_m)f_2(U_l, U_m)) = 0$, for U_l and U_m two independent uniform variables on $P_{i,n}$.

$$\sum_{l \neq m=1}^{N_i} f(U_l, U_m) = 2(N_i - 1) \sum_{l=1}^{N_i} f_1(U_l) + \sum_{l \neq m=1}^{N_i} f_2(U_l, U_m).$$

Lemma 5.5. f_1 is bounded by Cp^{-2} . f_2 is bounded by a constant and $f_2(x, y) \leq Cp^{-2}$ as soon as $\|x - y\| > r$.

Proof. All the quantities computed in Lemma 5.2 for the four different cases are bounded by a constant so that $\mathbb{E}(\mathbb{I}\{d(U_1, U_2) \leq r\} | U_1) = O(p^{-2})$. As $e_{r,p} = O(p^{-2})$, this is also true for $h_1(x, r)$ for any r and then for f_1 . Because $f_2(x, y) = f(x, y) - f_1(x) - f_1(y)$, it is bounded by a constant and $f_2(x, y) = O(p^{-2})$ as soon as the indicator function vanishes.

Define $M_1 = \mathbb{E} \left(\sum_{l=1}^{N_i} f_1(U_l) \right)^4$. Then $M_1 = N_i E(f_1^4(U)) + 6N_i(N_i - 1)E(f_1^2(U))^2$ and $\mathbb{E}(N_i - 1)^4 M_1 = O(p^2)$.

Define $M_2 = \mathbb{E} \left(\sum_{l \neq m=1}^{N_i} f_2(U_l, U_m) \right)^4$. Because f_2 is zero mean with respect to one coordinate, only the products where variables appear at least two times contribute.

$$\begin{aligned} M_2 &= 8 \sum_{l \neq m=1}^{N_i} \mathbb{E} f_2^4(U_l, U_m) + 48 \sum_{l \neq m \neq u=1}^{N_i} \mathbb{E} f_2^2(U_l, U_u) f_2^2(U_u, U_m) \\ &\quad + 96 \sum_{l \neq m \neq u=1}^{N_i} \mathbb{E} f_2^2(U_l, U_m) f_2(U_m, U_u) f_2(U_u, U_l) \\ &\quad + 12 \sum_{l \neq m \neq u \neq v=1}^{N_i} \mathbb{E} f_2^2(U_l, U_m) f_2^2(U_u, U_v) \\ &\quad + 48 \sum_{l \neq m \neq u \neq v=1}^{N_i} \mathbb{E} f_2(U_l, U_m) f_2(U_m, U_u) f_2(U_u, U_v) f_2(U_v, U_l). \end{aligned}$$

Consider the first sum

$$\sum_{l \neq m=1}^{N_i} \mathbb{E} f_2^4(U_l, U_m) \leq \sum_{l \neq m=1}^{N_i} \mathbb{P}\{d(U_l, U_m) \leq r\} + Cp \mathbb{P}\{d(U_l, U_m) > r\} p^{-8} \leq CN_i(N_i - 1)p^{-2}.$$

In all the sums, the main term comes from sets of points with all interdistance less than r and the resulting magnitude of the expectation is $O(p^2)$, so that

$$\sum_{i=0}^k \mathbb{E}(u_{i,n} - \mathbb{E}u_{i,n})^4 = O(n^{-2}).$$

Thus condition 1 is realised. To check condition 2, note that the vector $(K_{1,P_i}(r_1), \dots, K_{1,P_i}(r_d))$ has a covariance matrix Σ_p defined by Proposition 2.2 by substituting p to n in the expressions. The $u_{i,n} = \frac{p^2\sqrt{p}}{n} \sum_{t=1}^d \lambda_t(K_{1,P_i}(r_t) - \mathbb{E}K_{1,P_i}(r_t))$ are i.i.d variables with variance equal to $\frac{p^4}{n^2} \Lambda^t \Sigma_p \Lambda$. But $p^2 \rho \Sigma_p$ tends to Σ as p tends to infinity and

$$\sum_{i=0}^k \text{var } u_{i,n} = \frac{kp^4\rho}{n^2} \Lambda^t \Sigma_p \Lambda \longrightarrow \Lambda^t \Sigma \Lambda$$

so that $\sum_{i=1}^k u_{i,n}$ tends in distribution to $\mathcal{N}(0, \Lambda^t \Sigma \Lambda)$.

Note that the $v_{i,n}$ are k independent variables. Denote N_{i,r_d} the number of points X_l in the boundary region P_{i,r_d} of $P_{i,n}$ such that the ball $B(X_l, r_d)$ intersects Q and let $D(X_l)$ denote this intersection. Note that

$$\mathbb{E}N_{i,r_d} = \rho m(P_{i,r_d}) \leq Cpr_d.$$

$$\text{var } v_{i,n} \leq \frac{C}{n^2} \mathbb{E} \left(\sum_{l=1}^{N_{i,r_d}} \sum_{m=1}^{N_Q} \mathbb{I}\{X_m \in D(X_l)\} \right)^2 \leq \frac{C}{n^2} (T_1 + T_2),$$

where

$$\begin{aligned} T_1 &= \mathbb{E} \sum_{l=1}^{N_{i,r_d}} \sum_{m=1}^{N_Q} \sum_{u=1}^{N_Q} \mathbb{I}\{X_m \in D(X_l)\} \mathbb{I}\{X_u \in D(X_l)\} \\ T_2 &= \mathbb{E} \sum_{l=1}^{N_{i,r_d}} \sum_{m=1}^{N_{i,r_d}} \sum_{u=1}^{N_Q} \mathbb{I}\{X_u \in D(X_l) \cap D(X_m)\}. \end{aligned}$$

$$T_1 \leq \mathbb{E}N_{i,r_d} \mathbb{E}N_Q^2 \mathbb{P}\{X_m \in D(X_l) | X_m \in Q\} \leq \rho^3 m(P_{i,r_d}) (m^2(Q) + m(Q)) \left(\frac{\pi r_d^2}{2m(Q)} \right)^2 = O(p).$$

$$\begin{aligned} T_2 &= \mathbb{E} \sum_{l=1}^{N_{i,r_d}} \sum_{m=1}^{N_{i,r_d}} \sum_{u=1}^{N_Q} \mathbb{I}\{X_m \in B(X_l, 2r_d)\} \mathbb{I}\{X_u \in D(X_l) \cap D(X_m)\} \\ &\leq \mathbb{E}N_{i,r_d}^2 \mathbb{P}\{X_m \in B(X_l, r_d) | X_m \in P_{i,r_d}\} \mathbb{E}N_Q \mathbb{P}\{X_u \in D(X_l) | X_u \in Q\} \\ &\leq \rho^3 (m^2(P_{i,r_d}) + m(P_{i,r_d})) \left(\frac{\pi r_d^2}{m(P_{i,r_d})} \right) m(Q) \left(\frac{\pi r_d^2}{2m(Q)} \right) = O(p) \end{aligned}$$

and $\text{var} \left(\sum_{i=1}^k v_{i,n} \right) = O(kp/n^2) = O(p^{-1})$, so that this sum is negligible in L^2 . Similarly

$$\text{var}(w_n) \leq \frac{C}{n^2} \mathbb{E} \left(\sum_{l \neq m=1}^{N_Q} \mathbb{I}\{X_m \in B(X_l, r_d)\} \right)^2 \leq \frac{C}{n^2} (T_1 + T_2),$$

where

$$\begin{aligned}
T_1 &= \mathbb{E} \sum_{l=1}^{N_Q} \sum_{m=1}^{N_Q} \mathbb{I}\{X_m \in B(X_l, r_d)\} \\
&\leq \mathbb{E} N_Q (N_Q - 1) \mathbb{P}\{X_m \in B(X_l, r_d) | X_m \in Q\} \leq m^2(Q) \frac{\pi r_d^2}{m(Q)}. \\
T_2 &= \mathbb{E} \sum_{l=1}^{N_Q} \sum_{m=1}^{N_Q} \sum_{u=1}^{N_Q} \mathbb{I}\{X_m \in B(X_l, r_d)\} \mathbb{I}\{X_u \in B(X_l, r_d)\} \\
&\leq \mathbb{E} N_Q^2 (N_Q - 1) \mathbb{P}^2\{X_m \in B(X_l, r_d) | X_m \in Q\} \leq (m^3(Q) + 2m^2(Q)) \left(\frac{\pi r_d^2}{m(Q)} \right)^2.
\end{aligned}$$

Note that $m(Q) = O(\sqrt{kn})$. Then $\text{var}(w_n) = O(m(Q)/n^2) = O(p^{-1})$ and w_n is negligible in L^2 .

Consider now $K_{2,n}(r)$. Define $Z_2 = \sum_{t=1}^d \lambda_t K_{2,n}(r_t) = A_{N,n} Z_1$ where $A_{N,n} = \frac{n^4 \rho^2}{N(N-1)}$. We have $\mathbb{E}(A_{N,n}^{-1}) = 1$ and from (5.5), $\text{var}(A_{N,n}^{-1}) = \frac{4}{n^2 \rho} + \frac{2}{n^4 \rho^2}$. For $\delta > 0$, the Markov inequality gives

$$\mathbb{P}(|A_{N,n}^{-1} - 1| > \delta) \leq \frac{\text{var}(A_{N,n}^{-1})}{\delta^2}.$$

Then, with $\delta = n^{-1/4}$, $\sum_{n=1}^{\infty} \mathbb{P}(|A_{N,n}^{-1} - 1| > n^{-1/4}) < \sum_{n=1}^{\infty} \frac{4}{n^{3/2} \rho} + \frac{2}{n^{7/2} \rho^2} < \infty$. From the Borel–Cantelli lemma, we get that $A_{N,n}^{-1}$ converges a.s. to 1. By the Slutsky lemma, $A_{N,n} Z_1$ converges in distribution to $\mathcal{N}(0, A^t \Sigma A)$. \square

5.4. Proof of Lemma 5.3

This lemma is equivalent to Result 1 of [27], substituting r/n to the parameter h and subtracting $e_{r,n}^2$. From the computation of the bias, denoting $x_i = \frac{1}{r}(n - \xi_i)$, we get

$$\begin{aligned}
h_1(\xi, r) &= \frac{\pi r^2}{n^2} - e_{r,n} \text{ on } A_n^1 \\
&= \frac{r^2}{n^2} (\pi - g(x_1)) - e_{r,n} \text{ on } A_n^2 \\
&= \frac{r^2}{n^2} (\pi - g(x_1) - g(x_2)) - e_{r,n} \text{ on } A_n^3 \\
&= \frac{r^2}{n^2} \left(\frac{3\pi}{4} + x_1 x_2 - \frac{g(x_1) + g(x_2)}{2} \right) - e_{r,n} \text{ on } A_n^4
\end{aligned}$$

Integrating on the four zones, we get

$$\begin{aligned}
\mathbb{E}(h_1(X_j, r))^2 &= \pi^2 \left(1 - \frac{2r}{n} \right)^2 \frac{r^4}{n^4} - e_{r,n}^2 + T_1 + T_2 + T_3 \\
T_1 &= 4 \left(1 - \frac{2r}{n} \right) \frac{r^5}{n^5} \int_0^1 (\pi - g(x_1))^2 dx_1 \\
T_2 &= 4 \frac{r^6}{n^6} \int_0^1 dx_1 \int_{\sqrt{1-x_1^2}}^1 (\pi - g(x_1) - g(x_2))^2 dx_2 \\
T_3 &= 4 \frac{r^6}{n^6} \int_0^1 dx_1 \int_0^{\sqrt{1-x_1^2}} \left(\frac{3\pi}{4} + x_1 x_2 - \frac{g(x_1) + g(x_2)}{2} \right)^2 dx_2.
\end{aligned}$$

To rewrite these three terms with the notations of [27], we denote $\theta = \arccos(x_1)$ and $\phi = \arccos(x_2)$.

$$\begin{aligned} T_1 &= 4 \left(1 - \frac{2r}{n}\right) \frac{r^5}{n^5} \int_0^{\pi/2} (\pi - \theta + \cos(\theta) \sin(\theta))^2 \sin(\theta) d\theta \\ T_2 &= 4 \frac{r^6}{n^6} \int_0^{\pi/2} \sin(\theta) d\theta \int_0^{\pi/2-\theta} (\pi - \theta + \cos(\theta) \sin(\theta) - \phi + \cos(\phi) \sin(\phi))^2 \sin(\phi) d\phi \\ T_3 &= \frac{r^6}{n^6} \int_0^{\pi/2} \sin(\theta) d\theta \int_{\pi/2-\theta}^{\pi/2} (3\pi/2 + 2 \cos(\theta) \cos(\phi) - \theta - \phi + \cos(\theta) \sin(\theta) + \cos(\phi) \sin(\phi))^2 \sin(\phi) d\phi \\ &= \frac{r^6}{n^6} \int_0^{\pi/2} \cos(\theta') d\theta' \int_0^{\theta'} (\pi/2 + 2 \sin(\theta') \sin(\phi') + \theta' + \phi' + \cos(\theta') \sin(\theta') + \cos(\phi') \sin(\phi'))^2 \cos(\phi') d\phi'. \end{aligned}$$

changing variables by $\theta' = \pi/2 - \theta$ and $\phi' = \pi/2 - \phi$. Then formulas (5), (6) and (7) in [27] give respectively

$$T_1 = \left(1 - \frac{2r}{n}\right) \frac{r^5}{n^5} \left(4\pi^2 - \frac{256}{45} - \frac{8\pi}{3}\right) \quad (5.10)$$

$$T_2 = \frac{r^6}{n^6} \left(-\frac{\pi^3}{4} + 4\pi^2 - \frac{9\pi}{2} - \frac{512}{45}\right) \quad (5.11)$$

$$T_3 = \frac{r^6}{n^6} \left(\frac{\pi^3}{4} + \frac{19\pi}{48} + \frac{8}{9}\right). \quad (5.12)$$

Note that the upper bound of the second integral in formula (7) of [27] is a mistyping. Gathering the expression of $e_{r,n}$, (5.10)–(5.12) gives the result.

Acknowledgements. We are thankful to Michel Koskas for his help in accelerating the computation of K with C^{++} . We also wish to thank the anonymous referees for their suggestions to improve the section concerning the power of the test.

REFERENCES

- [1] A.J. Baddeley, M. Kerscher, K. Schladitz and B.T. Scott, Estimating the J function without edge correction. *Research report of the department of mathematics*, University of Western Australia (1997).
- [2] J.-M. Bardet, P. Doukhan, G. Lang and N. Ragache, Dependent Lindeberg central limit theorem and some applications. *ESAIM: PS* **12** (2008) 154–172.
- [3] S. Bernstein, Quelques remarques sur le théorème limite Liapounoff. *C.R. (Dokl.) Acad. Sci. URSS* **24** (1939) 3–8.
- [4] J.E. Besag, Comments on Ripley's paper. *J. Roy. Statist. Soc. Ser. B* **39** (1977) 193–195.
- [5] S.N. Chiu, Correction to Koen's critical values in testing spatial randomness. *J. Stat. Comput. Simul.* **77** (2007) 1001–1004.
- [6] S.N. Chiu and K.I. Liu, Generalized Cramér-von Mises goodness-of-fit tests for multivariate distributions. *Comput. Stat. Data Anal.* **53** (2009) 3817–3834.
- [7] N.A. Cressie, *Statistics for spatial data*. John Wiley and Sons, New York (1993).
- [8] P.J. Diggle, *Statistical analysis of spatial point patterns*. Academic Press, London (1983).
- [9] M. Fromont, B. Laurent and P. Reynaud-Bouret, Adaptive tests of homogeneity for a Poisson process. *Ann. I.H.P. (B)* **47** (2011) 176–213.
- [10] P. Grabarnik and S.N. Chiu, Goodness-of-fit test for complete spatial randomness against mixtures of regular and clustered spatial point processes. *Biometrika* **89** (2002) 411–421.
- [11] J. Gignoux, C. Duby and S. Barot, Comparing the performances of Diggle's tests of spatial randomness for small samples with and without edge effect correction: application to ecological data. *Biometrics* **55** (1999) 156–164.
- [12] Y. Guan, On nonparametric variance estimation for second-order statistics of inhomogeneous spatial point Processes with a known parametric intensity form. *J. Am. Stat. Ass.* **104** (2009) 1482–1491.
- [13] L.P. Ho and S.N. Chiu, Testing Uniformity of a Spatial Point Pattern. *J. Comput. Graph. Stat.* **16** 2 (2007) 378–398.
- [14] L. Heinrich, Goodness-of-fit tests for the second moment function of a stationary multidimensional Poisson process. *Statistics* **22** (1991) 245–268.
- [15] J. Illian, A. Penttinen, H. Stoyan and D. Stoyan, *Statistical analysis and modelling of spatial point patterns*. Wiley-Interscience, Chichester (2008).
- [16] C. Koen, Approximate confidence bounds for Ripley's statistic for random points in a square. *Biom. J.* **33** (1991) 173–177.

- [17] E. Marcon and F. Puech, Evaluating the geographic concentration of industries using distance-based methods. *J. Econom. Geogr.* **3** (2003) 409–428.
- [18] J. Møller and R.P. Waagepetersen, Statistical inference and simulation for spatial point processes, vol. 100 of *Monographs on statistics and applied probability*. Chapman and Hall/CRC, Boca Raton (2004).
- [19] R Development Core Team (2012). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. <http://www.R-project.org>.
- [20] B.D. Ripley, The second-order analysis of stationary point processes. *J. Appl. Probab.* **13** (1976) 255–266.
- [21] B.D. Ripley, Modelling spatial patterns. *J. Roy. Statist. Soc. Ser. B* **39** 2 (1977) 172–212.
- [22] B.D. Ripley, Tests of randomness for spatial point patterns. *J. Roy. Statist. Soc. Ser. B* **41** 3 (1979) 368–374.
- [23] B.D. Ripley, *Spatial statistics*. John Wiley and Sons, New York (1981).
- [24] R. Saunders and G.M. Funk, Poisson limits for a clustering model of Strauss. *J. Appl. Probab.* **14** (1977) 776–784.
- [25] D. Stoyan, W.S. Kendall and J. Mecke, *Stochastic geometry and its applications*. Akademie-Verlag, Berlin (1987).
- [26] D. Stoyan and H. Stoyan, *Fractals, Random Shapes and Point Fields. Methods of Geometrical Statistics*. John Wiley and Sons, New York (1994).
- [27] C.C. Taylor, I.L. Dryden and R. Farnoosh, The K function for nearly regular point processes. *Biometrics* **57** (2000) 224–231.
- [28] M. Thomas, A generalization of Poisson’s binomial limit for use in ecology. *Biometrika* **36** (1949) 18–25.
- [29] E. Thönnies and M.-C. van Lieshout, A comparative study on the power of van Lieshout and Baddeley’s J function. *Biom. J.* **41** (1999) 721–734.
- [30] J.S. Ward and F.J. Ferrandino, New derivation reduces bias and increases power of Ripley’s L index. *Ecological Modelling* **116** (1999) 225–236.

APPENDIX H

The decomposition of Shannon's entropy and a confidence interval for beta diversity

Marcon, E., 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.

The decomposition of Shannon's entropy and a confidence interval for beta diversity

Eric Marcon, Bruno Hérault, Christopher Baraloto and Gabriel Lang

E. Marcon (eric.marcon@ecofog.gf), AgroParisTech, UMR EcoFoG, BP 709, FR-97310 Kourou, French Guiana. – B. Hérault, Univ. des Antilles et de la Guyane, UMR EcoFoG, BP 709, FR-97310 Kourou, French Guiana. – C. Baraloto, INRA, UMR EcoFoG, BP 709, FR-97310 Kourou, French Guiana. – G. Lang, AgroParisTech, UMR 518 Math. Info. Appl., 16 rue Claude Bernard, FR-75005 Paris, France.

Beta diversity is among the most employed theoretical concepts in ecology and biodiversity conservation. Up to date, a self-contained definition of it, with no reference to alpha and gamma diversity, has never been proposed. Using Kullback-Leibler divergence, we present the explicit formula of Shannon's β entropy, a bias correction for its estimator and a confidence interval. We also provide the mathematical framework to decompose Shannon diversity into several hierarchical nested levels. From botanical inventories of tropical forest plots in French Guiana, we estimate Shannon diversity at the plot, forest and regional level. We believe this is a complete and usefulness toolbox for ecologists interested in partitioning biodiversity.

Alpha, beta and gamma diversities are among the most employed theoretical concepts in ecology and biodiversity conservation. For most ecologists, alpha diversity traditionally reflects the within-habitat diversity (MacArthur 1965), whereas beta diversity is the component of 'total diversity' that is produced by differences in species composition among the sampling units, i.e. 'the extent of change of community composition' (Whittaker 1960). The need to partition diversity within and among habitats has both theoretical (e.g. gradient analyses) and applied (e.g. defining protected areas) consequences such that these concepts are widely employed in both ecology (Crist et al. 2003) and conservation biology (Steinitz et al. 2005).

The partitioning of diversity began with ecological niche studies (Allan 1975), but recent interest has focused 1) in partitioning biodiversity measures into independent components (Jost 2007, Pélissier and Couteron 2007, Jost et al. 2009) and 2) in analyzing patterns of diversity sampled from hierarchically scaled studies (Lande 1996, Loreau 2000). This recent interest largely builds on Lande's (1996) explanations for additive partitions of total diversity (gamma) into components within-samples (alpha) and among-samples (beta), following thus the original concepts of alpha, beta and gamma diversity (Whittaker 1972), even though Lande's framework was correct only for Shannon diversity (Jost 2006, 2007). Up to now, diversity partitioning approaches have been used over a wide range of ecosystems, including tropical (Condit et al. 2002) as well as temperate landscapes (Qian et al. 2005). Although diversity partitioning has deep conceptual meanings for ecologists, its usefulness to date has been impeded by 1) the lack of theoretical basis for

interpreting the results (Jurasinski et al. 2009) and 2) the lack of statistical methods for testing null hypotheses (Crist et al. 2003).

As a good starting point, it has been acknowledged that most usual measures of diversity are particular cases of generalized entropy measures (Tsallis 1988). In this framework, the species richness is an entropy-based diversity measure of order 0, the Shannon diversity (Shannon 1948, Shannon and Weaver 1963) of order 1 and the Simpson (1949) of order 2 (Jost 2007). Decreasing the order of the diversity estimates is conceptually equivalent to enhancing the weight of rare species in the final diversity estimates (Keylock 2005). The Shannon estimates (order 1) are ecologically meaningful because all species are strictly weighted by their frequency. Furthermore, Jost (2007) shows that Shannon's estimate is the only common entropy measure that can be decomposed additively, so that the gamma entropy H_γ is the weighted sum of α entropy of partitions, H_α , and a between-partition entropy H_β , even when community weights are unequal. Transforming Shannon entropies into Hill numbers makes possible the derivation of the so-called 'true diversity' indices (Jost 2006, Tuomisto 2010), i.e. the number of equally-abundant elements needed to produce a given value of Shannon entropy. In the remainder, we will write 'entropy' for classical measures, and 'diversity' for their Hill numbers, to keep a consistent terminology.

Finally, Shannon measures have several intuitively expected properties of a diversity measure (Jost 2007): 1) alpha and beta components are mathematically independent, meaning that a high value of alpha does not force the beta component to be high and vice versa, 2) gamma

is completely determined by alpha and beta, and 3) alpha is never greater than gamma. These properties of Shannon measures, 1) and 2) shared by all Rényi's (1961) measures of entropy but not 3) when community weights are unequal, give them a privileged place as estimates of diversity. But, as far as we know, still lacking is an explicit mathematical formulation of H_β , whose value is always obtained by the difference $H_\gamma - H_\alpha$. Recently, Tuomisto (2010) provided an extremely detailed review of literature defining 'beta diversity as a function of alpha and gamma diversity' and, from this review, it emerges that a self-contained definition of H_β , with no reference to H_α and H_γ has never been proposed.

The aim of this paper is to give the explicit mathematical formulation of a dataset measure of biodiversity (γ) into within- (α) and between- (β) partition diversities following Whittaker's concepts (1972).

The paper is organized as follows. First, we derive the decomposition of Shannon's entropy H using Kullback-Leibler divergence, yielding the analytic formulation of H_β . Then, we show how to compute its confidence interval and we derive its bias correction. We use simulated datasets to show properties of H_β when data are samples from the same community. We use a real dataset to show how Shannon's diversity can be decomposed and how biases can be corrected, including hierarchical nested levels of decomposition: after splitting gamma diversity into alpha and beta components, alpha diversity of each community can itself be further considered as a gamma diversity for sub-communities and decomposed.

Methods

Derivation of the decomposition

Kullback-Leibler divergence

A Kullback-Leibler (1951) divergence, now synonymous of 'relative entropy' although Kullback and Leibler did not mention entropy, measures how different two distributions of probabilities are. Given a model distribution \mathbf{p} , actual frequencies \mathbf{q} of a finite number of observations i , the Kullback-Leibler divergence T between \mathbf{p} and \mathbf{q} is:

$$T = \sum_i q_i \ln \frac{q_i}{p_i} \quad (1)$$

Economists know this measure T as Theil's (1967) dissimilarity index. Note that T is a measure of divergence *s.s.*, neither a dissimilarity nor a distance measure, because \mathbf{p} and \mathbf{q} do not play a symmetric role. We also note that, theoretically, the observed values in a given system are only estimates of the actual frequencies \mathbf{q} . Unlike Mori et al. (2005), we interchange for simplicity \mathbf{q} and $\hat{\mathbf{q}}$, without consequences for our purpose here.

Consider an ecological community partitioned into plots. The number of individuals of each species s in each plot i is denoted n_{si} . The number of individuals in plot i is $n_{+i} = \sum_s n_{si}$, the number of individuals of species s is $n_{s+} = \sum_i n_{si}$. The total number of individuals is n_{++} . The corresponding actual frequencies are $q_{si} = n_{si}/n_{++}$, with $\sum_i \sum_s q_{si} = 1$. The expected distribution will be $p_{si} = n_{+i}/n_{++}$ where S is the number of

species. In other words, we expect that all species have the same frequency, and the number of individuals is proportional to the size of the plot.

Grouping rule

In this section, we derive a general Eq. 5 for grouping plots or species. Similar approaches are common in the economic literature (Bickenbach and Bode 2008), so we will follow its terminology.

Data are organized in a table where lines are species, indexed by s and columns are plots indexed by i . Consider any group of cells G : the contribution of the group to the total relative entropy is the sum of each cell's relative entropy. We denote it T_G^α :

$$T_G^\alpha = \sum_{g \in G} q_g \ln \frac{q_g}{p_g} \quad (2)$$

After grouping, a single cell remains. Its relative entropy is the between-group relative entropy, we denote it T_G^γ :

$$T_G^\gamma = q_G \ln \frac{q_G}{p_G} = \left(\sum_{g \in G} q_g \right) \ln \frac{\sum_{g \in G} q_g}{\sum_{g \in G} p_g} \quad (3)$$

Proof: the probability for an individual to belong to the group is the sum of the probabilities that it belongs to any cell of the group.

The within-group relative entropy is:

$$T_G^\beta = \sum_{g \in G} \frac{q_g}{\sum_{g \in G} q_g} \ln \frac{\frac{q_g}{\sum_{g \in G} q_g}}{\frac{p_g}{\sum_{g \in G} p_g}} = \left(\sum_{g \in G} q_g \right)^{-1} \times \left[\sum_{g \in G} q_g \ln \frac{q_g}{p_g} - \left(\sum_{g \in G} q_g \right) \ln \frac{\sum_{g \in G} q_g}{\sum_{g \in G} p_g} \right] \quad (4)$$

Proof: within the group, the sum of probabilities is 1. Within-group probabilities are therefore normalized.

Finally, the total relative entropy of the group equals its between-group plus its within-group relative entropy:

$$T_G^\alpha = T_G^\gamma + \left(\sum_{g \in G} q_g \right) T_G^\beta \quad (5)$$

At this step, alpha, beta and gamma are purely conventional notations. They will be justified later.

Application to Shannon's index

We apply the previous result to Shannon's index of diversity. The expected probability for species s in plot i is $1/S$ (all species are expected to have the same frequency) multiplied by n_{+i}/n , the weight of plot i . The observed frequency is $q_{si} = n_{si}/n_{++}$. We group all the cells of species s . The relative entropy of the dataset for species s is:

$$T_s^\alpha = \sum_i q_{si} \ln \frac{q_{si}}{p_{si}} = \sum_i \frac{n_{si}}{n_{++}} \left(\ln \frac{n_{si}}{n_{++}} + \ln S \right) \quad (6)$$

The gamma relative entropy of species s is:

$$T_s^\gamma = q_{s+} \ln \frac{q_{s+}}{p_{s+}} = \frac{n_{s+}}{n_{++}} \left(\ln \frac{n_{s+}}{n_{++}} + \ln S \right) \quad (7)$$

The between-plot relative entropy of species s is:

$$\begin{aligned} \left(\sum_i q_{si} \right) T_s^\beta &= \left(\sum_i q_{si} \right) \sum_i \frac{q_{si}}{q_{s+}} \ln \frac{\frac{q_{si}}{q_{s+}}}{\frac{p_{si}}{p_{s+}}} \\ &= \frac{n_{s+}}{n_{++}} \sum_i \frac{n_{si}}{n_{s+}} \ln \frac{\frac{n_{si}}{n_{s+}}}{\frac{n_{+i}}{n_{++}}} = \sum_i \frac{n_{si}}{n_{++}} \ln \frac{n_{si}}{n_{+i}} \end{aligned} \quad (8)$$

We know Eq. 5 that $T_s^\alpha = T_s^\gamma + (\sum_i q_{si}) T_s^\beta$. This equality will be summed over all species to introduce diversity measures:

$$\begin{aligned} T_\alpha &= \sum_s T_s^\alpha = \ln S + \sum_i \frac{n_{+i}}{n_{++}} \sum_s \frac{n_{si}}{n_{s+}} \ln \frac{n_{si}}{n_{+i}} \\ &= \ln S - \sum_i \frac{n_{+i}}{n_{++}} H_i^\alpha = \ln S - H_\alpha \end{aligned} \quad (9)$$

H_i^α is the alpha diversity of plot i . It is computed according to local frequencies n_{si}/n_{+i} . H_α is the weighted sum of H_i^α . T_α is the Kullback-Leibler divergence between \mathbf{p} and \mathbf{q} for all plots and all species.

The gamma relative entropy sums to give the Kullback-Leibler divergence for the dataset:

$$T_\gamma = \sum_s T_s^\gamma = \ln S + \sum_s \frac{n_{s+}}{n_{++}} \ln \frac{n_{s+}}{n_{++}} = \ln S - H_\gamma \quad (10)$$

Finally, we sum between-plot relative entropy:

$$\begin{aligned} T_\beta &= \sum_s \left(\sum_i q_{si} \right) T_s^\beta = \sum_i \sum_s \frac{n_{si}}{n_{++}} \ln \frac{\frac{n_{si}}{n_{s+}}}{\frac{n_{+i}}{n_{++}}} \\ &= \sum_i \frac{n_{+i}}{n_{++}} \sum_s \frac{n_{si}}{n_{s+}} \ln \frac{\frac{n_{si}}{n_{s+}}}{\frac{n_{+i}}{n_{++}}} \end{aligned} \quad (11)$$

Combining Eq. 9, 10 and 11 and assuming $H_\gamma = H_\alpha + H_\beta$, we identify diversity:

$$H_\beta = \sum_i \frac{n_{+i}}{n_{++}} H_i^\beta = \sum_i \frac{n_{+i}}{n_{++}} \sum_s \frac{n_{si}}{n_{s+}} \ln \frac{\frac{n_{si}}{n_{s+}}}{\frac{n_{+i}}{n_{++}}} \quad (12)$$

H_β is the weighted sum of contributions of plots i , H_i^β . These contributions are Kullback-Leibler divergences. The expected probabilities are $p_{si} = n_{s+}/n_{++}$. The probability to find an individual of species s in plot i is proportional to the frequency of the species in the dataset. All plots are expected to be identical. Observed frequencies are $q_{si} = n_{si}/n_{++}$; actual frequencies differ from plot to plot. In agreement with intuition, diversity is the divergence between identical plots and real plots.

Hill (1973) numbers are the numbers of equiprobable species yielding the same measure of diversity as the actual data, also called the effective number of species. They allow one to transform non-intuitive values of Shannon diversity into easy-to-understand numbers. The Hill number for β diversity is the number of equally-weighted, completely distinct plots giving the same value of H_β , that is to say the effective number of plots.

Confidence intervals

We want to have a confidence interval of H_β : plot data are samples of wider communities so observed values of H_β may vary due to sampling stochasticity. The confidence interval is computed by Monte-Carlo simulations assuming the species distribution of plots and resampling them. First, we draw each value of n_{si} in a binomial law $B(n_{+i}, n_{si}/n_{+i})$ and we calculate H_β . We then repeat the simulation a large number of times, (e.g. 10 000) and eliminate extreme values according to the chosen risk level α . For $\alpha = 5\%$, the confidence interval of the null hypothesis is between the 251st and the 9750th simulated values of H_β .

Sampling bias

Sampling bias occurs because the community under study contains an unknown number of species denoted \tilde{S} . Only S species have been observed in plots: some rare species have not been sampled, introducing a downward bias of H equal to $(\tilde{S} - 1)/2n$ plus a negligible term, derived by Basharin (1959), that remains intractable. Chao and Shen (2003) built an unbiased estimator for H_α in plot i :

$$H_i^\alpha = \sum_s \frac{C_i \frac{n_{si}}{n_{+i}} \ln C_i \frac{n_{si}}{n_{+i}}}{1 - \left(1 - C_i \frac{n_{si}}{n_{+i}} \right)^n} \quad (13)$$

C_i is the estimator of the sample coverage (Good 1953), that is to say the proportion of observed species in terms of probabilities in plot i . We denote C the sample coverage for the whole dataset. A C equal to 90% means that unobserved species represent 10% of individuals of the community. The sample coverage is estimated by $1 - S_1/n$ where S_1 is the number of species observed only once (we will call them singletons) in the sample. Shannon's alpha or gamma entropy can be estimated easily without bias by simply taking into account singletons and the sample size.

We follow Chao and Shen to derive an unbiased estimator for \tilde{H}_β . The denominator was introduced by Horvitz and

Thompson (1952) to correct for unobserved species: it is equal to the probability to not sample each species s . Probability estimators are observed frequencies multiplied by the sample coverage so that they sum to 1 including unobserved species. These corrections are applied to H_i^β . The Horvitz-Thompson correction is the same as for \tilde{H}_i^α . Frequencies in each plot are multiplied by the plot's coverage, while those of the whole dataset are multiplied by the whole dataset's coverage. We get:

$$\tilde{H}_i^\beta = \sum_s \frac{C_i \frac{n_{si}}{n_{+i}} \ln \frac{C_i \frac{n_{si}}{n_{+i}}}{C \frac{n_{s+}}{n_{++}}}}{1 - \left(1 - C_i \frac{n_{si}}{n_{+i}}\right)^n} \quad (14)$$

In turn, when resampling plots (by drawing binomial laws as we do, or by bootstrapping), some rare species among the S observed are eliminated. Alpha and gamma diversities of simulated plots are thus systematically lower than those of the original ones. Alpha diversity is more biased than gamma diversity due to sample sizes. Beta diversity is thus overestimated. This can be corrected by applying Chao and Shen's bias correction, but remains incomplete because the unobserved number of species samples are drawn from S , not \tilde{S} . Finally, the unbiased simulated values are distributed around the biased actual ones. We will call them semi-unbiased values. No technique is available for a nested bias correction that would correct for the unobserved species among \tilde{S} .

A complete correction of the bias of simulations can be done numerically, for H_β as well as H_α and H_γ . For a given plot, the bias is constant as it only depends on the unobserved number of species and the sample size. As a result, the simulated variance of H is not affected: the biased simulations can be simply re-centered around the actual value of H .

Examples

We first used a simulated dataset to illustrate that the minimum value of H_β depends on both the number of species in the community and the sampling effort. We simulated two frequency distributions of respectively 20 and 40 species in plots. Simulated frequencies follow a uniform law. Then we drew a pair of plots 10 000 times from these communities, with an expectancy of 500 individuals or 5000 individuals. We computed H_β for each pair of plots to construct a frequency histogram of H_β , smoothed as a density function. No bias correction is needed here. H_β is not zero due to stochasticity although plots are samples of the same community.

We also provide a real example to show how deal with actual data. We measured Shannon diversity for tree communities in four 1-ha plots of tropical rain forest at the Nouragues and Paracou field stations in French Guiana. Both sites are seasonal lowland forests receiving about 3 m of annual precipitation, with tree composition dominated by *Fabaceae*, *Chrysobalanaceae*, *Lecythidaceae*, *Sapotaceae* and *Burseraceae* (Bongers et al. 2001, Gourlet-Fleury et al. 2005). The two plots within each site were chosen to represent the

most common contrasting environments found for hilltop terra firme forest at each site. At Paracou, the two example plots occur on migmatite associated with the Bonidoro geological formation, with one plot exhibiting blocked vertical drainage (P06) and the other with strong vertical drainage and incipient podzolization (P18). At Nouragues, the two example plots occur on weathered granite with sandy soils (NH20) and metavolcanic rock of the Paramaca formation with clay-rich laterite soils. In all four plots all trees were sampled in 2008 by professional climbers to obtain herbarium vouchers, each of which was identified to distinct morphospecies at the Cayenne regional herbarium (Baraloto et al. 2010). We assume for simplicity here that an acceptable sample of each forest site is obtained when its two plots are united.

Results

The simulations exemplified how H_β values depend on the number of individuals sampled. H_β does not change if all numbers of individuals are multiplied by 10, while maintaining actual frequencies the same. But frequencies vary randomly and, as a divergence, H_β accumulates these fluctuations. Its expected value is not 0 even when plots are from the same community. When more individuals are drawn, species frequencies converge to their probability due to the law of large numbers, so H_β converges to 0. In a 20-species community, an observed value $H_\beta = 0.005$ shows a significant difference between plots if it is obtained from two 5000-tree plots (Fig. 1, solid curve on the left). If the plots

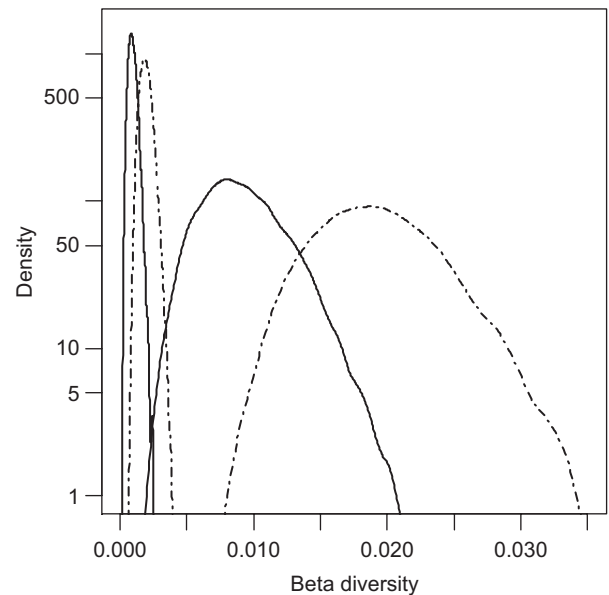


Figure 1. Probability densities of H_β obtained from 10 000 simulations of the model described in details in the text. Two plots are drawn from the same community. H_β is not zero because of stochastic differences between the plots. The first two curves on the left concern plots around 500 individuals, the right ones plots around 5000 individuals. Dotted lines are for 40-species plots, solid lines for 20 species. Everything else equal, expected H_β s decrease with the number of individuals and increase with the number of species.

Table 1. An example of hierarchical decomposition of Shannon entropy (H) for tropical tree communities. Trees with diameter at breast height > 10 cm were inventoried in four 1-ha tropical rain forest plots in French Guiana. The first two plots (NH20, NL11) are from the Nouragues forest station, the last two from the Paracou (P006, P018) forest station. Within forests, the weighted sum of alpha (H_{α}) and beta entropies (H_{β}) equals the within forest gamma entropy (H_{γ}). This within forest gamma entropy can be considered as the alpha entropy at the between-forest level. In this way, adding H_{γ} to the beta entropy between forests (H_{β}) gives the total entropy. Hill numbers are the numbers of equiprobable species or completely different plots or forests yielding the same measure of diversity as the actual data. Beta diversities are given with their 95% confidence interval between square brackets. All values are bias corrected.

	NH20	NL11	P006	P018
No. of trees	558	515	643	481
No. of observed species S	203	182	147	149
Total no. of species \tilde{S} estimated by Jackknife1	321	279	215	223
\tilde{H}_{α}	5.01	4.92	4.40	4.67
Hill no. (true plot alpha diversity)	151	137	82	107
\tilde{H}_{β} with 95% CI	0.33 [0.30;0.36]		0.36 [0.33;0.39]	
Hill no. plots (true beta diversity)	1.39 [1.35;1.43]		1.44 [1.39;1.48]	
\tilde{H}_{γ}	5.31		4.90	
Hill no. (true forest diversity)	201		134	
\tilde{H}_{β} with 95% CI			0.33 [0.31;0.35]	
Hill no. forests (true beta diversity)			1.39 [1.36;1.41]	
\tilde{H}_{total}			5.44	
Hill no. total (true gamma diversity)			230	

contain only 500 trees (Fig. 1, solid curve on the right), the same value is no longer significant. H_{β} also tends to be higher when the number of species increases for the same sample size: less individuals per species mean more stochasticity because relative entropy is calculated per species (Eq. 8) and summed.

The second example illustrates how Shannon diversity can be hierarchically partitioned (Table 1) and how to deal with biases. We first followed Beck and Schwanghart (2010) to validate the possibility to correctly estimate diversity. The total number of species \tilde{S} was estimated according to Brose et al.'s (2003) framework to evaluate sample completeness, that is to say the proportion in numbers of observed species (S/\tilde{S}), which is lower than sample coverage, the proportion in probabilities. Jackknife1 appeared to be the appropriate estimator, and completeness was around 2/3 in all plots. This is enough to validate bias correction of Shannon indices according to Beck and Schwanghart's empirical findings.

The first result is that plots at the Nouragues site are more diverse than those at Paracou. Hill numbers offer an intuitive representation of the level of diversity. For example, the Nouragues NH20 plot is as diverse as one of the same size with 151 equally frequent species, almost twice the value obtained at Paracou P006. Next, plots can be grouped into forests, i.e. Nouragues and Paracou. The value of H_{γ} , that is to say the γ diversity of the forest, is the sum of the weighted H_{α} of plots plus the H_{β} between plots. In turn, H_{γ} can be treated as an α diversity and one can reiterate the same procedure. Successive values of H are given in Table 1.

The confidence interval of H_{β} is shown. For example, concerning the Nouragues plots, H_{β} is estimated at 0.33, corresponding to a Hill number equal to 1.39 plots (95% confidence interval between 1.35 and 1.43). Note that theoretical Hill values for this example of two plots are between 1 (perfect equality of distribution, $H_{\beta} = 0$) and 2 (equal number of trees with no species in common, $H_{\beta} = \ln 2 \approx 0.7$). We can see that diversity within forests is roughly the same as that between forests (all values of Hill Numbers are around 1.4) and all values are highly significant (the

probability to have $H_{\beta} = 0$ is so low that this can be considered as impossible).

We could have chosen to group the four sampled plots directly. In this case, H_{β} between all plots would be 0.67. The corresponding Hill number would be 1.95 (confidence interval between 1.89 and 2.01) meaning that the four plots are almost equivalent to two completely different ones.

Discussion

Previous works

In this paper, we propose a self-contained definition of H_{β} , with no reference to H_{α} and H_{γ} . The form of H_{β} we provide has already been derived by Ricotta and Avena (2003), but they did not relate it with H_{α} and H_{γ} . Also, Ludovisi and Taticchi (2006) decomposed a Kullback-Leibler divergence with a different approach, in order to develop new measures of β diversity.

Interpretation and properties of H_{β}

Kullback-Leibler divergences provide the necessary framework to decompose Shannon diversity. Shannon's α diversity is the difference between the logarithm of the number of species and Theil's relative entropy, that is to say the Kullback-Leibler divergence between a distribution where all species have the same frequency and actual data. Shannon's γ diversity has the same definition after grouping plots. Shannon's β diversity is the Kullback-Leibler divergence between actual plots and identical ones.

In the proposed diversity partitioning framework, H_{β} is very different from H_{α} and H_{γ} because it is a measure of divergence, not of diversity itself. More similar species relative abundances result in higher H_{α} and H_{γ} , but H_{β} increases when plots are less similar. This is in agreement with the original definition of diversity expressed by Whitaker (1960). Converting Shannon's entropy to Hill numbers allows a unified definition of diversity as a number of effective objects

(species or plots), or ‘true diversity’ measures (thoroughly discussed by Tuomisto 2010, p. 8).

The maximum theoretical value of H_β is $\ln S$ when all plots have an α diversity equal to zero (i.e. they contain a single species that is different among plots); and γ diversity also has its maximum value, equal to $\ln S$. This is possible only if the number of samples equals the number of species and the number of individuals in every sample is the same. A more realistic situation is H_β equal to the logarithm of the number of samples. In this case, samples contain the same number of individuals (equal weight) but have no species in common. This is a special case of the maximum value derived by Jost (2007, Eq. 21), equal to the Shannon entropy of the weights of plots when they have no species in common. This is why observed β diversity makes sense only when compared to the weighted number of plots (Jost 2007 proposed to normalize it to the unit interval): β diversity between Nouragues and Paracou in our example is 1.39 for two samples, interpretable as a marked difference between communities. If it were 1.39 for 10 samples, it would mean that they are almost similar.

The minimum value is 0 when all plots are completely identical in species relative abundances. This never happens if they are random samples of the same community (Fig. 1). So the confidence interval cannot be used as a test for community equality because it never contains 0. We do not provide such a test, following Jones and Matloff (1986) for example, because increasing sample size always allows to make it significant as, unlike models of our theoretical examples, real communities are never exactly identical.

The estimator \tilde{H}_β may be negative if bias correction is erroneous. The simplest example is given by two or more exactly identical plots with singletons. Biased H_β is 0 so $\tilde{H}_\beta < 0$. Bias correction assumes that data are random samples so that singletons allow estimating unobserved species, but this artificial example violates these assumptions.

A user's guide

In summary, we propose a complete procedure to analyze data. The R (R Development Core Team 2010) code we wrote and used can be found as a supplementary material for use in further studies to compute unbiased values of H and confidence intervals.

The first step consists in evaluating the completeness of each plot in the sense of Beck and Schwanghart (2010) following Brose et al.'s (2003, Fig. 6) framework. Note that the term ‘coverage’ was replaced by ‘completeness’ by Beck and Schwanghart (2010) to avoid confusion with coverage defined by Good (1953). If completeness is above 50%, bias correction is very efficient such that Shannon's diversity can be estimated as follows:

- Unbiased \tilde{H}_i^α are computed according to Eq. (13) in each plot i . \tilde{H}_α is the weighted sum of plot diversities: $\tilde{H}_\alpha = \sum_i n_{+i}/n_{++} \tilde{H}_i^\alpha$.
- Unbiased \tilde{H}_β is obtained the same way, from Eq. (14).
- The confidence interval of \tilde{H}_β is computed by simulating plots, calculating their semi-unbiased beta diversity and re-centering the distribution around \tilde{H}_β .

- \tilde{H}_γ is calculated according to Eq. (13) after grouping data.
- Finally, all entropy values should be transformed into true diversities (their exponential) to allow interpretation.

Conclusion

In this paper, we provided the explicit formula of Shannon's β diversity and the mathematical framework to justify it. We showed that Shannon's β diversity is the Kullback-Leibler divergence between actual plots and the average plot. We also explained how to calculate the interval confidence of H_β . As real data are almost always incomplete (i.e. some rare species have not been sampled), we provided bias correction for the estimator of H_β . Finally, we showed how to decompose Shannon diversity into several nested levels. All results are interpretable intuitively after transformation into Hill Numbers.

This diversity partitioning is flexible enough to analyze any a priori determinant of species diversity. We believe that the use of explicit diversity partitioning will help both ecologists to understand the factors shaping the spatial and temporal distribution of biodiversity and nature practitioners to design effective strategies for protecting biodiversity (Veech et al. 2002).

Acknowledgements – We wish to sincerely thank Lou Jost whose suggestions and demands made us improve the quality of the paper very significantly.

References

- Allan, J. D. 1975. Components of diversity. – *Oecologia* 18: 359–367.
- Baraloto, C. et al. 2010. Functional trait variation and sampling strategies in species rich plant communities. – *Funct. Ecol.* 24: 208–216.
- Basharin, G. P. 1959. On a statistical estimate for the entropy of a sequence of independent random variables. – *Theor. Probabil. Appl.* 4: 333–336.
- Beck, J. and Schwanghart, W. 2010. Comparing measures of species diversity from incomplete inventories: an update. – *Meth. Ecol. Evol.* 1: 38–44.
- Bickenbach, F. and Bode, E. 2008. Disproportionality measures of concentration, specialization, and localization. – *Int. Regional Sci. Rev.* 31: 359–388.
- Bongers, F. et al. (eds) 2001. Nouragues: dynamics and plant-animal interactions in a neotropical rainforest. – Kluwer.
- Brose, U. et al. 2003. Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. – *Ecology* 84: 2364–2377.
- Chao, A. and Shen, T. J. 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. – *Environ. Ecol. Stat.* 10: 429–443.
- Condit, R. et al. 2002. Beta-diversity in tropical forest trees. – *Science* 295: 666–669.
- Crist, T. O. et al. 2003. Partitioning species diversity across landscapes and regions: a hierarchical analysis of alpha, beta, and gamma diversity. – *Am. Nat.* 162: 734–743.
- Good, I. J. 1953. On the population frequency of species and the estimation of population parameters. – *Biometrika* 40: 237–264.

- Gourlet-Fleury, S. et al. 2005. Using models to predict recovery and assess tree species vulnerability in logged tropical forests: a case study from French Guiana. – *For. Ecol. Manage.* 209: 69–86.
- Hill, M. O. 1973. Diversity and evenness: a unifying notation and its consequences. – *Ecology* 54: 427–432.
- Horvitz, D. G. and Thompson, D. J. 1952. A generalization of sampling without replacement from a finite universe. – *J. Am. Stat. Ass.* 47: 663–685.
- Jones, D. and Matloff, N. 1986. Statistical hypothesis testing in biology: a contradiction in terms. – *J. Econ. Entomol.* 79: 1156–1160.
- Jost, L. 2006. Entropy and diversity. – *Oikos* 113: 363–375.
- Jost, L. 2007. Partitioning diversity into independent alpha and beta components. – *Ecology* 88: 2427–2439.
- Jost, L. et al. 2009. Partitioning diversity for conservation analyses. – *Divers. Distrib.* 16: 65–76.
- Jurasinski, G. et al. 2009. Inventory, differentiation, and proportional diversity: a consistent terminology for quantifying species diversity. – *Oecologia* 159: 15–26.
- Keylock, C. J. 2005. Simpson diversity and the Shannon-Wiener index as special cases of a generalized entropy. – *Oikos* 109: 203–207.
- Kullback, S. and Leibler, R. A. 1951. On information and sufficiency. – *Ann. Math. Stat.* 22: 79–85.
- Lande, R. 1996. Statistics and partitioning of species diversity, and similarity among multiple communities. – *Oikos* 76: 5–13.
- Loreau, M. 2000. Are communities saturated? On the relationship between alpha, beta and gamma diversity. – *Ecol. Lett.* 3: 73–76.
- Ludovisi, A. and Taticchi, M. I. 2006. Investigating beta diversity by Kullback-Leibler information measures. – *Ecol. Modell.* 192: 299–313.
- MacArthur, R. H. 1965. Patterns of species diversity. – *Biol. Rev.* 40: 510–533.
- Mori, T. et al. 2005. A divergence statistic for industrial localization. – *Rev. Econ. Stat.* 87: 635–651.
- Pélissier, R. and Couteron, P. 2007. An operational, additive framework for species diversity partitioning and beta-diversity analysis. – *J. Ecol.* 95: 294–300.
- Qian, H. et al. 2005. Beta diversity of angiosperms in temperate floras of eastern Asia and eastern North America. – *Ecol. Lett.* 8: 15–22.
- Rényi, A. 1961. On measures of entropy and information. – In: Neyman, J. (ed.), 4th Berkeley Symp. Math. Stat. Probabil. Univ. of California Press, pp. 547–561.
- Ricotta, C. and Avena, G. 2003. An information-theoretical measure of β -diversity. – *Plant Biosyst.* 137: 57–61.
- Shannon, C. E. 1948. A mathematical theory of communication. – *Bell System Tech. J.* 27: 379–423, 623–656.
- Shannon, C. E. and Weaver, W. 1963. The mathematical theory of communication. – Univ. of Illinois Press.
- Simpson, E. H. 1949. Measurement of diversity. – *Nature* 163: 688.
- Steinitz, O. et al. 2005. Predicting regional patterns of similarity in species composition for conservation planning. – *Conserv. Biol.* 19: 1978–1988.
- Theil, H. 1967. Economics and information theory. – Rand McNally and Co.
- Tsallis, C. 1988. Possible generalization of Boltzmann-Gibbs statistics. – *J. Stat. Phys.* 52: 479–487.
- Tuomisto, H. 2010. A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. – *Ecography* 33: 2–22.
- Veech, J. A. et al. 2002. The additive partitioning of species diversity: recent revival of an old idea. – *Oikos* 99: 3–9.
- Whittaker, R. H. 1960. Vegetation of the Siskiyou Mountains, Oregon and California. – *Ecol. Monogr.* 30: 279–338.
- Whittaker, R. H. 1972. Evolution and measurement of species diversity. – *Taxon* 21: 213–251.

APPENDIX I

Characterizing the Relative Spatial Structure of Point Patterns

Marcon, E., F. Puech et S. Traissac (2012). « Characterizing the Relative Spatial Structure of Point Patterns ». In : International Journal of Ecology 2012.Article ID 619281.

Research Article

Characterizing the Relative Spatial Structure of Point Patterns

Eric Marcon,¹ Florence Puech,² and Stéphane Traissac¹

¹ AgroParisTech, UMR EcoFoG, BP 709, 97310 Kourou, French Guiana

² LET (Université de Lyon, CNRS, ENTPE), Institut des Sciences de l'Homme, 14 avenue Berthelot, 69363 Lyon Cedex 07, France

Correspondence should be addressed to Eric Marcon, eric.marcon@ecofog.gf

Received 26 May 2012; Revised 23 July 2012; Accepted 21 August 2012

Academic Editor: Cajo J. F. ter Braak

Copyright © 2012 Eric Marcon et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We generalize Ripley's K function to get a new function, M , to characterize the spatial structure of a point pattern relatively to another one. We show that this new approach is pertinent in ecology when space is not homogenous and the size of objects matters. We present how to use the function and test the data against the null hypothesis of independence between points. In a tropical tree data set we detect intraspecific aggregation and interspecific competition.

1. Introduction

Investigating the spatial structure of point patterns has been a long-time challenge for ecologists. Pielou [1] claimed that the information an ecologist wants to have immediately when observing a point set representing a vegetal community is the density of each species and the existence of interactions between plants. Density can be estimated by various methods [2] but interactions have motivated a living literature for more than half a century. In ecology, Ripley's K function [3], or its square-root transformation L [4], is the most used tool to characterize them [5], assuming the pattern is a realization of a homogenous point process; that is to say the probability to find a point is the same everywhere.

However, identifying interactions under the assumption of nonhomogeneity of space is still an open question. Twenty years ago, Cuzick and Edwards [6] followed by Diggle and Chetwynd [7] paved the way by introducing some specific tests. The D function proposed by Diggle and Chetwynd is defined as the difference between the K function for the studied points (called cases) and the K function for others (called controls). It is not completely satisfactory yet because both K functions are computed separately so all the data contained in the relative position of cases and controls is lost. A more recent important advance was proposed by Baddeley et al. [8] who generalized K to inhomogeneous point processes. They developed a complete theoretical framework but practical applications are still difficult because assumptions are

necessary about the relative scales of heterogeneity and interactions, leading to possibly opposite analyses of the results [9]. A recent review of these methods can be found in [5].

Other tools were also developed by economists [10] with a different approach, comparing the distribution of points of interest relatively to that of other points. Brülhart and Traeger [11] call them *relative* measures, as opposed to *topographic* measures such as K which take space (measured by areas) as their benchmark.

In this study, we introduce a relative measure of spatial structure, namely, the M function [12] to extend the ecologists' toolbox and allow a more pertinent approach when the null hypothesis is that points of interest are distributed like others. It bypasses the issue of heterogeneity and allows weighting points. Moreover, its computation is easy.

The paper is organized as follows: first, we derive the M function as a generalization of Ripley's K function; then, we apply it to theoretical examples and real data sets in a tropical forest and in epidemiology; we finally discuss the way it can be usefully applied after clarifying the assumptions it relies on.

2. Methods

2.1. Ripley's K Function

2.1.1. Definition. The theoretical framework is a point process whose realization is observed in a window of area

A. Nonparametric methods such as Ripley's K are used to reject the null hypothesis of independence of points. To use Ripley's K , we will assume that the point process is *stationary* (i.e., its intensity does not vary by translation). All point processes used in this paper are *second-order stationary* (i.e., interactions between points do not vary under translation) and *isotropic* (i.e., they do not depend on direction). The null hypothesis to reject is therefore complete spatial randomness (CSR); that is, the point pattern is a realization of a homogenous Poisson process. More details on K can be found in [5] or [13]. We focus on its estimator here.

Points are denoted x . We call a point x 's neighbors all the points less than r apart from it (all points in a disk of radius r centered on the point x).

Ripley's K function estimator was built by counting neighbors (indexed by n) around reference points (indexed by f), which can belong to the same type (*univariate* K function) or not (*intertype* or *bivariate* K function). N points are found in the window, and we denote $\mathbf{1}(\|x_f - x_n\| \leq r)$ the indicator function equal to 1 if the distance between x_f and x_n is less than or equal to r , 0 else.

An unbiased estimator of univariate K [14] with no edge-effect correction is

$$\hat{K}(r) = \frac{A}{N(N-1)} \sum_{f=1}^N \sum_{n=1, n \neq f}^N \mathbf{1}(\|x_f - x_n\| \leq r). \quad (1)$$

The bivariate version of K (denoted $K_{f,n}$) is very similar. We denote N_f the number of reference points and N_n the number of neighbors. We have

$$\hat{K}_{f,n}(r) = \frac{A}{N_f N_n} \sum_{f=1}^{N_f} \sum_{n=1}^{N_n} \mathbf{1}(\|x_f - x_n\| \leq r). \quad (2)$$

2.1.2. Edge-Effect Correction. Points located close to the window borders are problematic because a part of the circle inside which points are supposed to be counted is outside the window. Various answers have been proposed to correct for this [15, 16]. We prefer Besag's [4] correction. Let us denote A_{fr} the part of the area of the circle of radius r centered on the point x_f located inside the window. We count the number of neighbors inside the circle, and we correct it by the ratio between the circle's area and its inside part. We suppose that the outside part of the circle would have contained the same neighbor density than the inside part. Finally, an unbiased estimator of K with edge-effect correction is

$$\hat{K}(r) = \frac{A}{N(N-1)} \sum_{f=1}^N \sum_{n=1, n \neq f}^N \mathbf{1}(\|x_f - x_n\| \leq r) \frac{\pi r^2}{A_{fr}}. \quad (3)$$

2.1.3. Normalization. Besag [4] proposed to normalize K to obtain a benchmark of r rather than πr^2 . The well-known L function is defined as $L(r) = \sqrt{K(r)/\pi}$. It can be interpreted as a distance [17]: $L(r) = r + l$ means that as many neighbors are found around reference points up to distance r as would be expected at distance $r + l$ under CSR. We believe that $K(r)/\pi r^2$ is a better normalization. Its reference value is 1, and it can be interpreted as the density of neighbors around reference points divided by the density of neighbors anywhere.

2.2. The M Function

2.2.1. Definition of M . Equation (3) can be rearranged:

$$\frac{\hat{K}(r)}{\pi r^2} = \frac{\left[(1/N) \sum_{f=1}^N \left(\sum_{n=1, n \neq f}^N \mathbf{1}(\|x_f - x_n\| \leq r) / A_{fr} \right) \right]}{[(N-1)/A]}. \quad (4)$$

Around each point x_f , $[\sum_{n=1, n \neq f}^N \mathbf{1}(\|x_f - x_n\| \leq r)] / A_{fr}$ is the number of neighbors divided by the area where it is counted. Its average value is compared to what it is expected to be all over the window, $(N-1)/A$.

Topographic measures like K use space as their benchmark; that is, the number of points is divided by an area. The benchmark may also be another point pattern; for example, the number of trees of a species under study may be divided by the total number of neighbor trees, defining *relative* measures.

We transpose K into a relative framework. The ratio is now built comparing a number of neighbors of interest to the total number of neighbors. Weights can be associated to points without changing the construction of the measure. Reference points are indexed by f (x_f is a reference point), neighbor points by n ; all points whatever their type (i.e., the benchmark) by a ; their numbers are N_f , N_n , and N_a . w_i is the weight of point x_i , $W_i = \sum_{i=1}^{N_i} w_i$ is the total weight of this type of points.

The average weighted ratio of neighbor points around reference points is $(1/N_f) \sum_{f=1}^{N_f} (\sum_{n=1, n \neq f}^{N_n} \mathbf{1}(\|x_f - x_n\| \leq r) w_n) / \sum_{a=1, a \neq f}^{N_a} \mathbf{1}(\|x_f - x_a\| \leq r) w_a$.

In the whole window, the same ratio is $(1/N_f) \sum_{f=1}^{N_f} ((W_n - w_f) / (W_a - w_f))$ if neighbor points and reference points belong to the same type, $(1/N_f) \sum_{f=1}^{N_f} (W_n / (W_a - w_f))$ else. If reference and neighbor points belong to the same type, $N_f = N_n$ and $W_f = W_n$.

We define the univariate M function as

$$M(r) = \frac{\sum_{f=1}^{N_f} \left(\sum_{n=1, n \neq f}^{N_n} \mathbf{1}(\|x_f - x_n\| \leq r) w_n \right) / \sum_{a=1, a \neq f}^{N_a} \mathbf{1}(\|x_f - x_a\| \leq r) w_a}{\sum_{f=1}^{N_f} ((W_f - w_f) / (W_a - w_f))}. \quad (5)$$

The bivariate $M_{f,n}$ function is

$$M_{f,n}(r) = \frac{\sum_{f=1}^{N_f} \left(\sum_{n=1}^{N_n} \mathbf{1}(\|x_f - x_n\| \leq r) w_n / \sum_{a=1, x_a \neq x_f}^{N_a} \mathbf{1}(\|x_f - x_a\| \leq r) w_a \right)}{\sum_{f=1}^{N_f} (W_n / (W_a - w_f))}. \quad (6)$$

Equations (5) and (6) are simplified when points are not weighted:

$$M(r) = \frac{N_a - 1}{N_f(N_f - 1)} \sum_{f=1}^{N_f} \frac{\sum_{n=1, x_n \neq x_f}^{N_n} \mathbf{1}(\|x_f - x_n\| \leq r)}{\sum_{a=1, x_a \neq x_f}^{N_a} \mathbf{1}(\|x_f - x_a\| \leq r)},$$

$$M_{f,n}(r) = \frac{N_a - 1}{N_f N_n} \sum_{f=1}^{N_f} \frac{\sum_{n=1}^{N_n} \mathbf{1}(\|x_f - x_n\| \leq r)}{\sum_{a=1, x_a \neq x_f}^{N_a} \mathbf{1}(\|x_f - x_a\| \leq r)}. \quad (7)$$

2.2.2. Case-Control Design. A particular attention must be paid to case-control designs. In practical terms, all points of interest (called *cases*) are carefully referenced, and the

benchmark point set (called *controls*) is just sampled. This approach has been widely used for spatial clustering of diseases [7, 9, 18–20, among others]: sick people are the *cases* and the rest of the population the *controls*. Case-control design is of course not limitative to geographical epidemiology and can easily be applied to ecology questions.

The M function defined previously can be slightly modified to take into account this feature. Since the controls are chosen to be a representative sample of the population at every scale, neighbors of any kind are replaced by controls, indexed by a . Reference and neighbor points are N_c cases; their total weight is W_c . After simplifications, M_{cases} can be written as follows:

$$M_{\text{cases}}(r) = \frac{\sum_{f=1}^{N_c} \left(\sum_{n=1, x_n \neq x_f}^{N_c} \mathbf{1}(\|x_f - x_n\| \leq r) w_n / \sum_{a=1}^{N_a} \mathbf{1}(\|x_f - x_a\| \leq r) w_a \right)}{W_c(N_c - 1)/W_a}. \quad (8)$$

2.2.3. Significance. The first-order property (intensity) of the process must be controlled to allow the detection of the second-order property (nonindependence of points, that is to say interactions between the objects they represent). Thus, a point distribution generated according to the null hypothesis must respect, on the one hand, the local values of the density of the process the point distribution is a realization of and, on the other hand, its points must be distributed independently from each other.

The practical difficulty comes from the lack of knowledge of the point process that gave the point distribution, which is its unique available realization. Its first-order property is consequently widely unknown. We can only assume that the actual set of point locations is a good approximation of it, following Duranton and Overman [10]. Consequently, we generate random data sets for the univariate and case-control M functions by redistributing the actual point set (type and weight couples) on the actual location set (coordinates). The confidence interval of the null hypothesis is then computed by the Monte Carlo technique [21].

The intertype function must support two null hypotheses [22]. The random labeling hypothesis is simulated by permuting the point types, keeping point locations and weights unchanged. The population independence hypothesis is more complex to test. The reference points are kept unchanged, so that the spatial structure of the reference point type is maintained, and all other points are redistributed across the available locations. This allows

testing the independence of populations considering the structure of the reference point type. Then, the reference and neighbor types are interchanged and the test is repeated. If both $M_{f,n}$ and $M_{n,f}$ functions leave their null hypothesis confidence envelope in a range of distances, then population independence is rejected. This test requires that some points do not belong to either the reference or the neighbor point type or there will be nothing to redistribute. More generally, testing the relative spatial structure only makes sense if the tested point types are a small part of the point pattern; see discussion Section 4.1.

The tests based on Monte Carlo simulations are actually not correct because they are repeated at each step of the function (see [23] for an extensive discussion). A global test, without the need of simulations, is available only for K against CSR [14]. We can follow Loosmore and Ford's goodness-of-fit (GoF) test to obtain a correct P -value to reject the null hypothesis. We first need to compute the average value of $M(r)$ on all simulations, more exactly [24]:

$$\overline{M}(r) = \frac{1}{s-1} \sum_{i=1}^s M_i(r), \quad (9)$$

where s is the number of simulations and $M_i(r)$ the value of $M(r)$ in the i th simulation. Then, the statistic u_i is calculated for the i th simulation by summing on all values of r , where δr

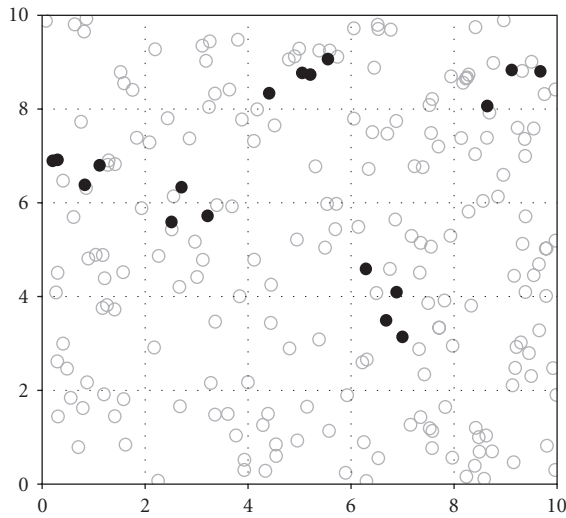


FIGURE 1: Aggregates, Point map. Grey circles are drawn from a homogenous Poisson process. Black disks are generated by a Matérn (radius = 0.5) process.

is the difference between the next value of r and the present one:

$$u_i = \sum_r [M_i(r) - \bar{M}(r)]^2 \delta r. \quad (10)$$

The same statistic for the actual data, denoted u , is compared to the simulated values to get a P -value:

$$P_u \approx \frac{\sum_{i=1}^s \mathbf{1}(u_i > u)}{s}. \quad (11)$$

If u is greater than all simulated values, the P -value to reject the null hypothesis erroneously is around 0. To avoid 0 or 1 P -values, we can assume that another simulation would have given a value of u_i higher or lower than u and write $P_u < 1/s$ or $P_u > 1 - 1/s$.

2.3. Examples. Three theoretical examples are given. Two of them illustrate very simple point patterns on a homogeneous space for a comparison of L and M functions (Sections 3.1.1 and 3.1.2). The third one computes an inhomogeneous Poisson point process to show how the M function controls for the first-order property of point processes (Section 3.1.3). No theoretical example is given with weighted points because they are not so easy to understand visually. Three real point patterns are considered then. They do not allow a classical analysis by the K function because of heterogeneity. Cuzick and Edwards [6] introduced the first formal way to deal with nonhomogeneous point processes: they used a dataset (published with the paper) concerning the location of 62 cases of childhood leukemia between 1974 and 1986 in the North Humberside area, England. A control set of 141 children representing the whole concerned population was chosen from the birth register (all weights are 1). They could conclude that the cases were significantly clumped. We use this data set to go further: we are now able to corroborate

their conclusion and also to precise the size of aggregates. The M function is computed according to the case-control design, (8).

We overall want to provide evidence of the interest of relative spatial structure in ecology. Trees are considered in a 25 ha plot of tropical rainforest in Paracou field station in French Guiana [25]. We investigate the spatial structure of two species, *Vouacapoua americana* Aublet (Caesalpinaceae) and *Qualea rosea* Aublet (Vochysiaceae) in a point set of 11,276 trees above 10 cm diameter at breast height (DBH), excluding flooded zones. All trees above 1 cm diameter have been measured and plotted for a few species, allowing us to study the spatial relations between saplings (up to 10 cm DBH) and possibly reproductive trees (30 cm or more) of *V. americana*. Points are weighted by the basal area of the tree they represent, the reference and neighbor points are mentioned in the results, and all trees of the maps, including references and neighbors, are used as the benchmark.

3. Results

3.1. Theoretical Examples. In what follows, we generate a point pattern (“black points,” represented by closed circles in the figures) to investigate its spatial structure with the L and M univariate functions. Other points (“grey points” represented by grey open circles) are used by M only: grey and black points together constitute the benchmark. Black points may be considered as trees of a species of interest in a forest plot, while grey points are all other trees.

All confidence intervals are computed at 1% risk level generated from 10,000 simulations.

3.1.1. Aggregates. 200 grey points are completely randomly distributed. Black points are generated by a Matérn process [26]: 5 aggregates (radius 0.5) of 5 points. All point weights equal 1. The map is in Figure 1; the curves are in Figure 2.

The M curve shape is similar to L ’s: significant, positive peaks denote concentration. The benchmark points of the M function are distributed almost homogeneously so the number of neighbors around each point is proportional to the area: the relative and the topographic measures are nearly equivalent. Nevertheless, while L peaks approximately correspond to the diameter of aggregates [27], M peaks occur exactly at distances at which the local density is the greatest, that is, approximately the distance between points in the aggregates. The differences are due to the way L is normalized: $M(r)$ peaks occur at the same distance as those of $K(r)/\pi r^2$ (not shown on the figure).

3.1.2. Regularity. 200 grey points are drawn from a homogeneous Poisson process again. 100 black points have a regular distribution around a square, 1-by-1 grid, with a perturbation: each point is randomly moved horizontally and vertically within a 0.4 interval around the grid nodes (Figure 3). All point weights equal 1.

The first part of the univariate M curve (Figure 4) is made of 0 values, showing the absence of neighbors at any distance up to 0.6. Note that the univariate L curve shape is

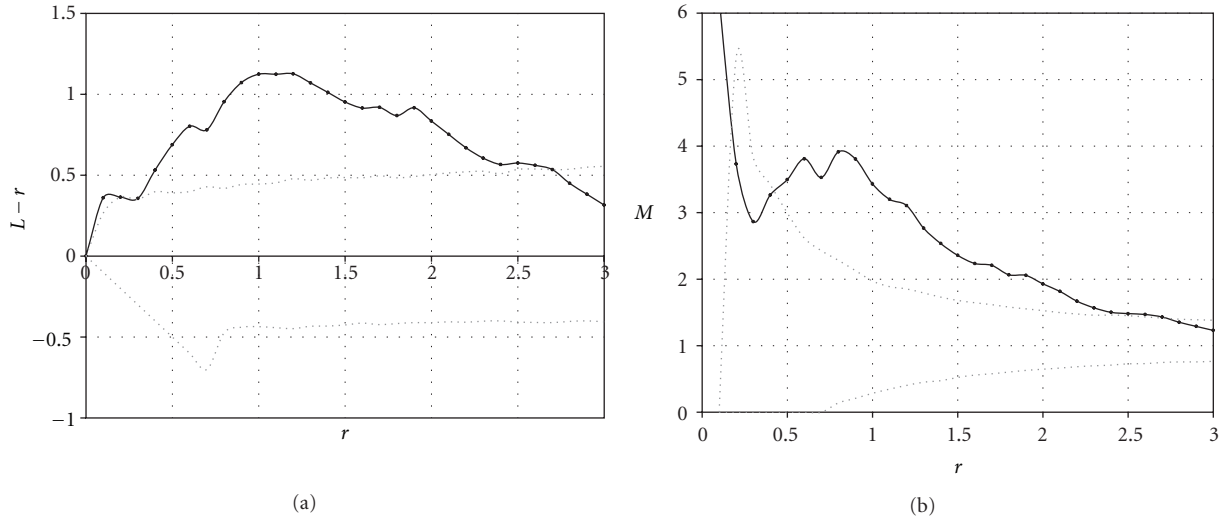


FIGURE 2: Aggregates, univariate L and M functions for the aggregated point set. Solid curves are the L and M function values; dotted lines are the envelope of the confidence interval of the null hypothesis. Both functions detect clumping. $L(r) - r$ is plotted rather than $L(r)$ for convenience.

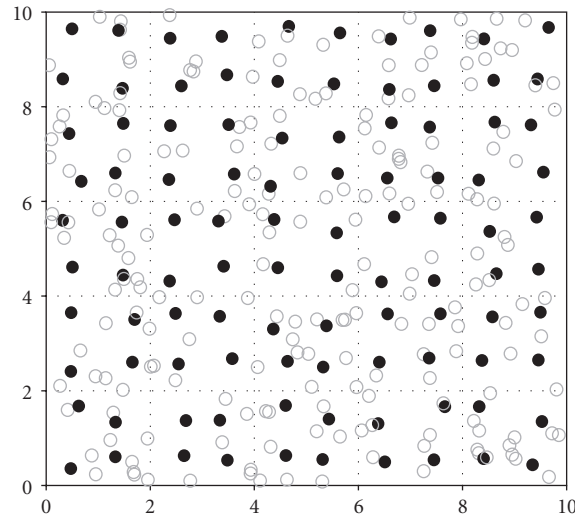


FIGURE 3: Regular point set, Point map. Grey circles are from a Poisson process. Black disks are located close to a 1×1 square grid.

different since its original value is 0 and its minimum slope is -1 by construction.

Negative peaks of both the univariate L and M functions allow detecting the grid size (before $r = 1, 2$, and 3). Maximum values correspond to the diagonal of the grid ($\sqrt{2} \approx 1.41$ is the diagonal length, then $\sqrt{5} \approx 2.24$ and so on).

3.1.3. Inhomogeneous Point Set. We generated two completely random point sets in a 10-by-10 window. Then, we transformed the point coordinates: after having calculated the polar coordinates (r, θ) of each point from the center of the point set, we squared the distance to get (r^2, θ) . The result is a nonhomogeneous Poisson pattern, shown in Figure 5. Both point types have the same random distribution, but the center of the map shows a greater density.

It can be seen (Figure 6) that the L function is not applicable: assuming homogeneity, it interprets the black point distribution as a single big aggregate. This issue is known as “virtual aggregation” [28, 29]. The M function is able to control for density variations: since the pattern of the black points does not differ from that of all points, M values are around 1.

3.2. Cuzick and Edwards Data Set. The case-control M function is used (Figure 7). 0.7 km apart from cases, the average case density is about 70% higher than it would be if the cases followed the control pattern (at this distance, the peak of the M function reaches 1.7).

In the discussion of [6], Diggle (page 101) suggested that the D function, equal to $K_{\text{cases}} - K_{\text{controls}}$, would be appropriate

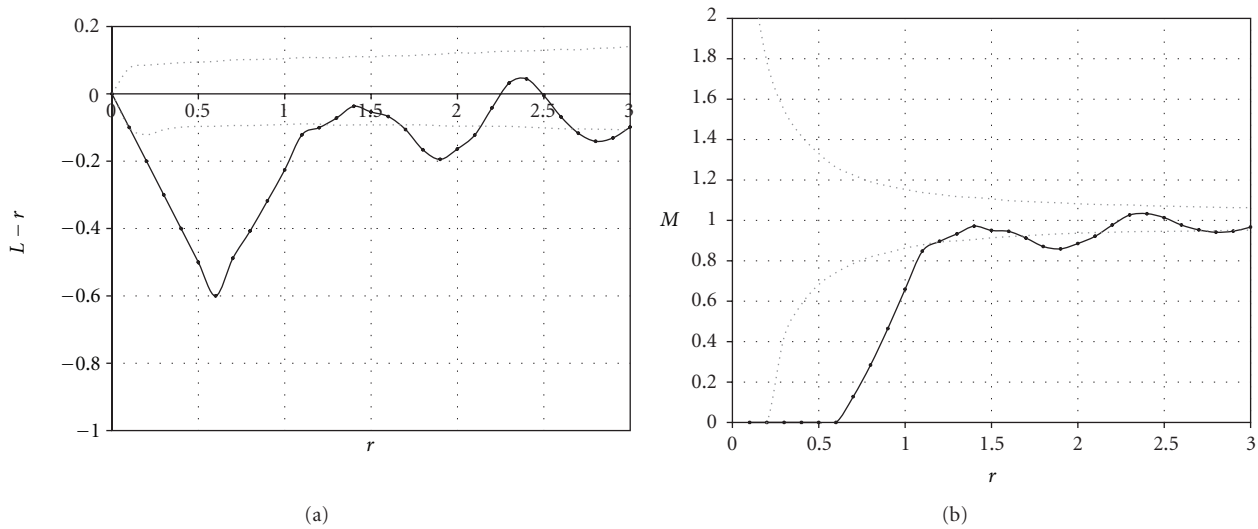


FIGURE 4: Regular point set, univariate L and M functions for the regular point set. Both functions detect dispersion.

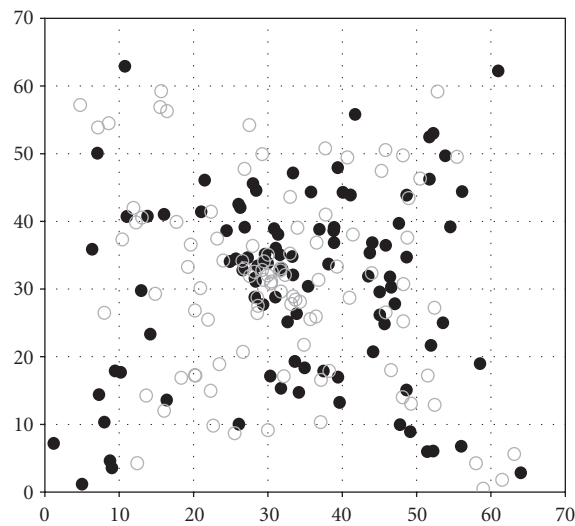


FIGURE 5: Inhomogeneous point set, Point map. All points are drawn from an inhomogeneous Poisson point process.

for this point pattern. The next year, Diggle and Chetwynd [7] published the D function and applied it to the same dataset. We recomputed D considering the rectangle window shown in the figure. This data set has been widely used and gave slightly different results according to the window definition in [7, page 1160] or [30, page 634]. It can be seen that the M and D functions give the same results if points are not weighted. Nevertheless, D values cannot be interpreted easily and cannot be compared across distances.

Both methods suffer here a severe lack of power due to the very little number of controls. The confidence envelopes are computed at 10% levels (from 1000 simulations). The GoF test applied to M returns $P_u = 23\%$. Diggle and Chetwynd obtained a P -value equal to 14% for D . Increasing the number of controls would not have been a real problem if the experimental design had included a distance-based point pattern analysis.

3.3. *V. americana* and *Q. rosea*. The dataset (map in Figure 8) contains 156 *V. americana*, 388 *Q. rosea*, and 10,732 other trees in a 25 ha plot where flooded zones have been excluded, leaving a 20.06 ha, polygonal shape for the study.

Aggregation of both species is detected up from 4–6 meters (Figures 9(a) and 9(d)) by the univariate M function. The species repulse each other (Figures 9(b) and 9(c)). $P_u < 0.1\%$ in all cases, according to the bivariate M function. All trees are used as the benchmark in all analyses.

These results suggest competition if our null hypothesis is correct: we suppose that both species could locate anywhere if the other did not impede it. Of course, it might be wrong so further work is necessary to test alternate hypotheses: the environment may be different and niche preferences may be the reason for segregation, or else the spatial distribution of populations may not be in equilibrium, and we may be observing the contact of two colonization fronts.

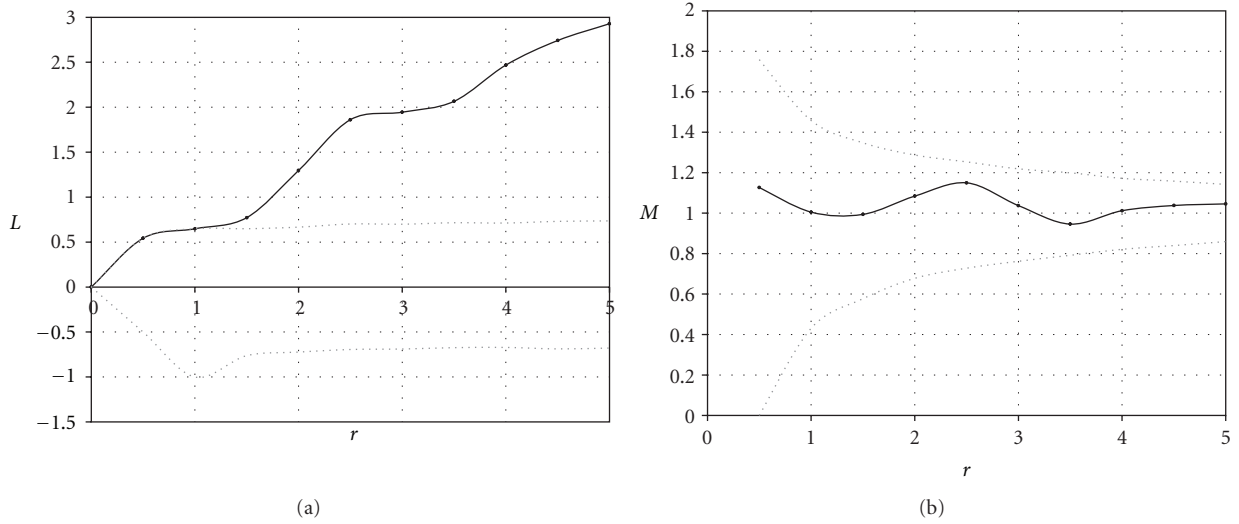


FIGURE 6: Inhomogeneous point set, univariate L and M functions for black points. L is not pertinent contrarily to M that controls for first-order heterogeneity.

3.4. *V. americana* Regeneration. The density of *V. americana* is very variable (Figure 10(a)). The univariate M function is applied to saplings (reference and neighbor points are saplings, the benchmark is all trees; weights are basal areas). Saplings are aggregated (Figure 10(b)) at all distances, $P_u < 0.1\%$. The intertype M function shows that potentially reproducing trees repulse saplings (Figure 10(c)), with significant results up to 15 meters. Actually, $P_u = 6.7\%$ (reference points are potentially reproducing trees, neighbors are saplings, and the benchmark is all trees; weights are basal areas).

Jansen et al. [31] already mentioned the absence of seedlings around adult *V. americana* at short distances (6 meters). Our results show that no sapling can be found less than 3 meters apart from adult trees, but also that the relative abundance of saplings among neighbors is low (these results are significant according to Monte-Carlo simulations), suggesting that *Vouacapoua* regeneration follows the Janzen-Connell hypothesis [32, 33].

4. Discussion

4.1. Using M . The theoretical examples illustrate that $M(r)$ is equivalent to $\hat{K}(r)/\pi r^2$ when applied to a homogenous, nonweighted point pattern. The univariate M function is not affected by virtual aggregation when the point pattern is not homogenous, using all points as its benchmark. This means that the points of interest should be a small enough fraction of the whole point set to allow considering the latter as a valid control set. The case-control approach is meaningful when the points of interest must not be included in the control set. Then, a sufficient number of controls is required, or the test against the null hypothesis of independence of points will not be powerful.

Although the M function requires several point types, it is completely different from a bivariate K or L function. This may be illustrated by the example of Section 3.1.1: the univariate M function characterizes the spatial structure of

black points, exactly like K ; grey points are added to black points to obtain a benchmark for $M(r)$ (the number of points whatever their type less than r apart from each black point), while the disk area πr^2 is the benchmark for $K(r)$. In Section 3.3, the bivariate K function could be computed instead of the bivariate M function: it would only consider the 156 *V. americana* and 388 *Q. rosea* trees (and suppose both are distributed homogeneously), while M also includes the 10,732 other trees to constitute its benchmark.

Figure 9(a) shows a good example of the behavior of the M function. Confidence envelopes are around the expected value equal to 1. At very low distances, they are not defined if no pair of points less than r apart exists, and the confidence interval is very wide then because of stochasticity, amplified by the little number of point pairs. At long distances, all values tend to 1. When point weights are not homogeneously distributed, the envelope is not around 1 (Figure 10). Heuristically, M measures the spatial structure of square centimeters of basal areas of trees: when points are redistributed independently from each other under the null hypothesis, square centimeters are still aggregated. More or less aggregation than under the null hypothesis is detected relatively to its envelope and by the GoF test. As shown by Loosmore and Ford [23], the classical, Monte-Carlo-generated confidence envelope may be too optimistic: while the M curve is clearly out of the 1% envelope, the P -value for *V. americana* regeneration (Figure 10) is only 6.7% (but no power study for the GoF has been conducted as far as we know).

4.2. Relative versus Topographic Measures. Distance-based measures of spatial concentration can be classified into two main categories [34] following Brühlhart and Traeger [11]. Topographic ones compare a number of neighbors to a measure of space (a surface area) while relative ones compare it to another number of neighbors. All indices used in ecology are topographic, except for Diggle and Chetwynd [7] D . On the opposite, economists, who are often interested

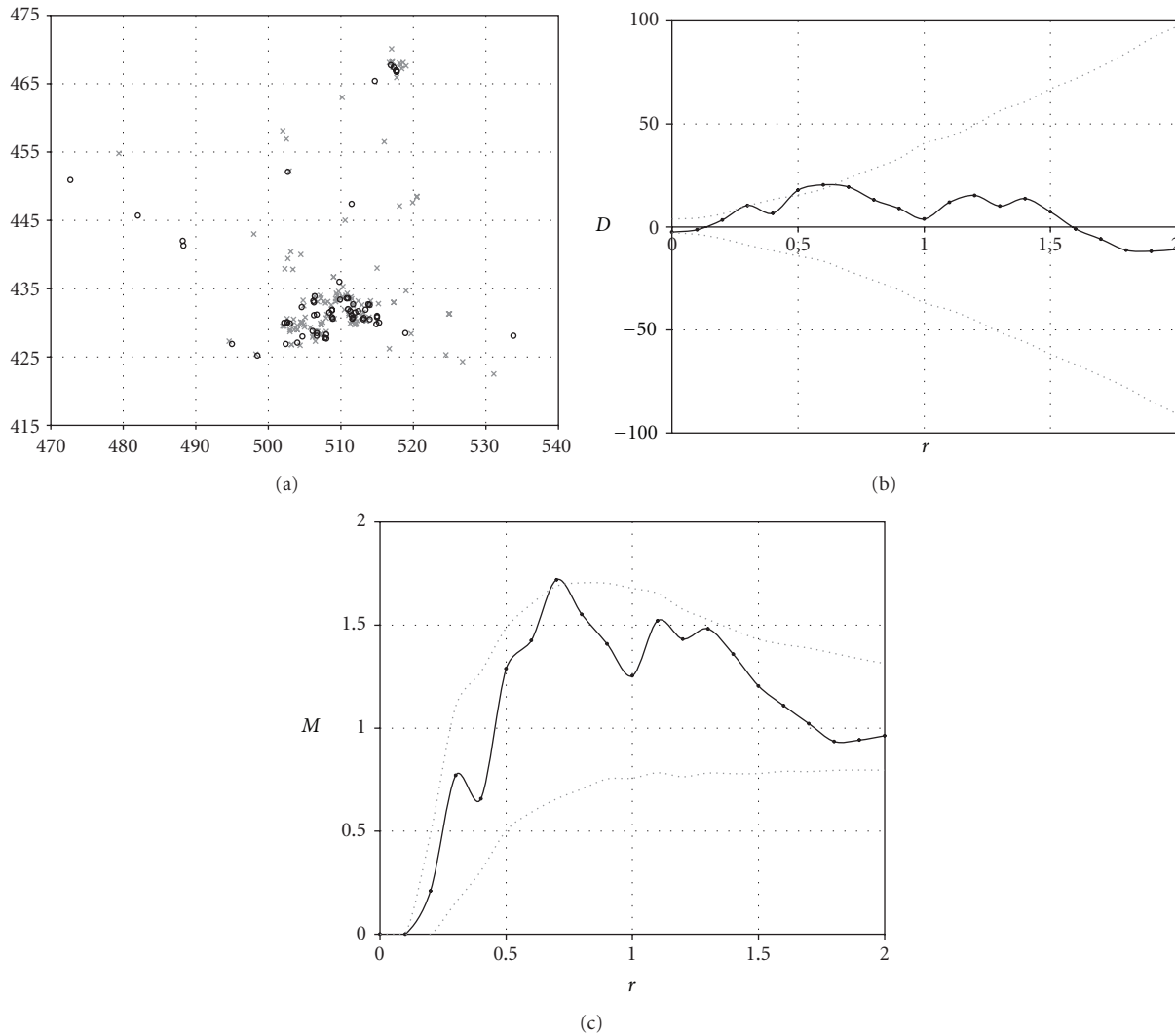


FIGURE 7: Childhood Leukemia epidemiology [6]. Map (a): cases (62 circles) are ill children locations; controls (141 crosses) are a sample of the whole population; distances are in km. Cases are significantly clumped, as shown by both functions D (b) and M (c), drawn as solid lines. M shows that in a 0.7 km radius circle around each case, the case density is 70% higher than expected without aggregation ($M = 1.7$). Confidence intervals for the null hypothesis of independence (dotted lines) are computed by Monte-Carlo simulations at the 10% risk level. The poor significance levels are due to too few controls.

in the spatial distribution of firms on a territory, mostly use relative measures: using the distribution of the whole industry as a benchmark to study the spatial structure of an economic sector is even one of the good properties a measure must respect according to Duranton and Overman [10]. We believe both frameworks can help addressing ecological questions, as they allow different null models to be tested.

The topographic toolbox is already well furnished, with Baddeley et al.'s [8] K_{inhom} and Wiegand and Moloney's [28, 35] O-ring allowing to deal with inhomogeneous point patterns. Diggle et al. [9] separated control points to evaluate intensity and cases to evaluate dependence in K_{inhom} ; that is to say they used K_{inhom} as a relative measure. The M function is designed for this purpose. It is similar to K_{inhom} with a simple box kernel [13] with bandwidth parameter r

used to estimate density around each reference point, but it also allows weighting points. The relative structure of basal areas of trees is more meaningful than that of individuals in many applications (biomass spatial structure, competition for light, etc.): roughly speaking, a big tree is often more similar to many small trees than to a single one.

5. Conclusion

The M function is defined as a generalization of Ripley's K function to allow its application to inhomogeneous point processes and to take into account point weights. From a more theoretical point of view, it is a weighted, relative measure of spatial structure. We believe that relative measures (comparing a point pattern to another) are powerful tools, even though the topographic approach is more used.

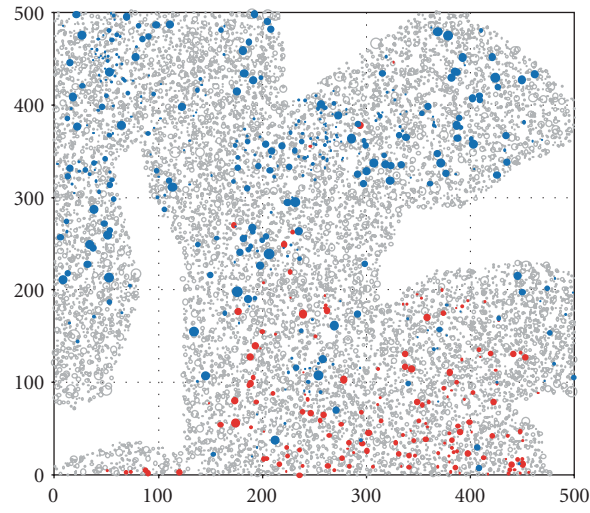


FIGURE 8: Map of trees. *Vouacapoua americana* trees are blue, *Qualea rosea* red and other trees grey. Circle sizes are proportional to those of the trees. Flooded zones are excluded. Distances are in meters.

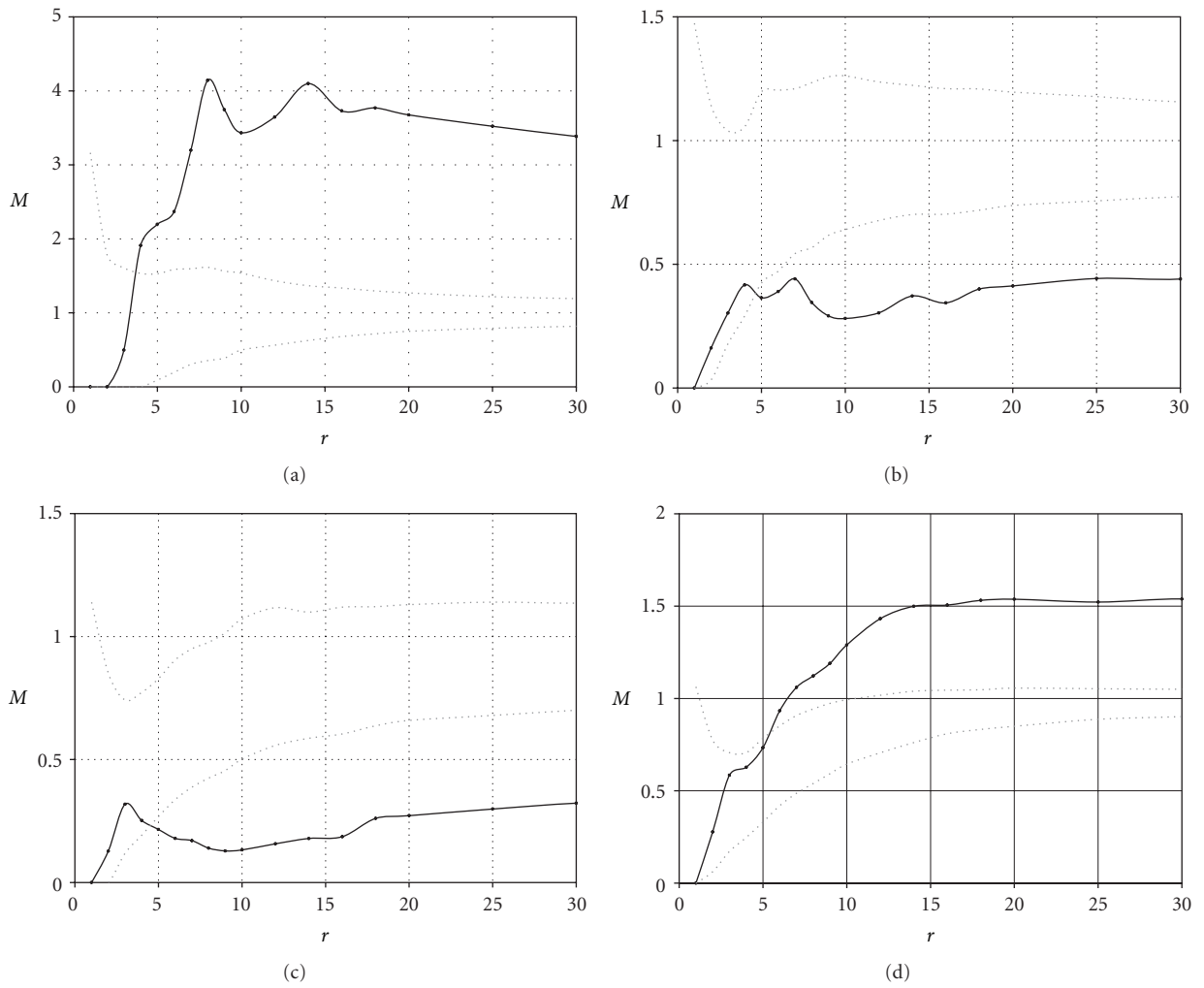


FIGURE 9: Spatial structure of *Vouacapoua americana* ((a) univariate M function) and *Qualea rosea* ((d) univariate M function), and bivariate M functions ((b) $M_{Va, Qr}$ and (c) $M_{Qr, Va}$). Confidence intervals are computed at 1% risk level. Both species are aggregated over around 5 meters, and they significantly repulse each other.

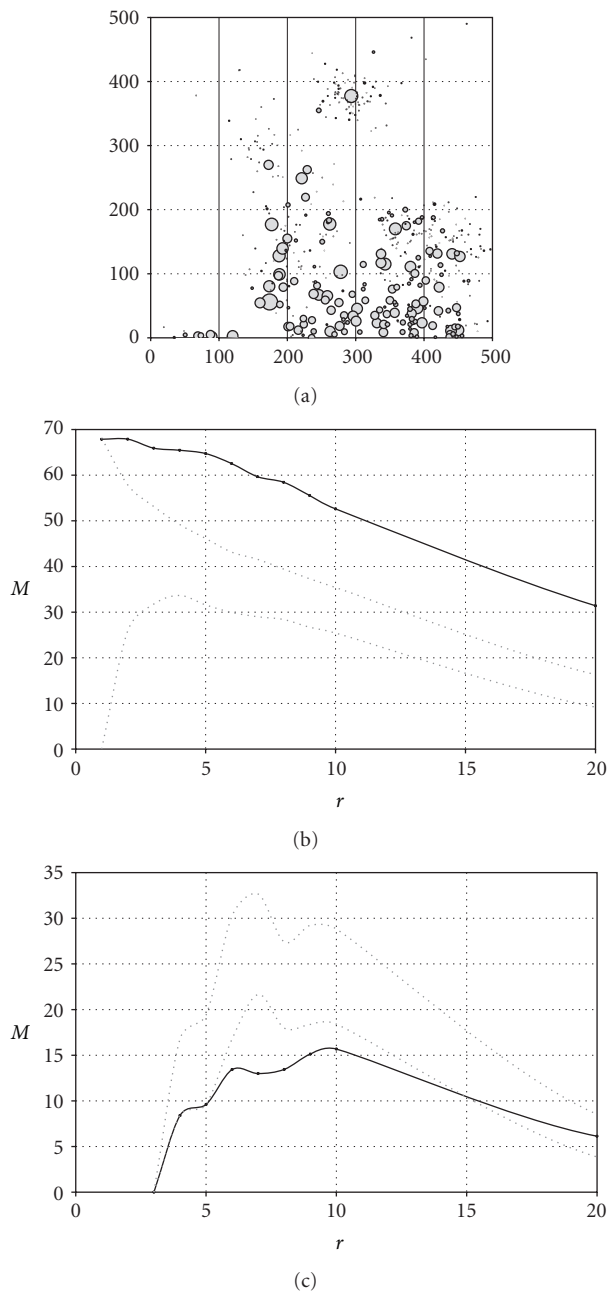


FIGURE 10: Regeneration of *Vouacapoua americana*. (a) The map shows trees with their size. Distances are in meters. (b) Univariate M function applied to saplings: saplings are aggregated at all distances. (c) Bivariate M function applied to saplings around potentially reproducing trees (over 30 cm DBH): a significant lack of saplings is detected between 6 and 9 meters. Confidence intervals are computed at 1% risk level.

To allow the effective use of the M function, we developed the necessary code for R [36], available as a supplementary material available online at doi:10.1155/2012/619281.

Acknowledgments

The authors thank the editor and an anonymous referee for useful suggestions. This work has benefited from an

“Investissement d’Avenir” grant managed by Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-0025). This paper partially incorporates earlier unpublished work written by E. Marcon and F. Puech (generalizing Ripley’s K function to inhomogeneous populations, halshs-00372631, version 1).

References

- [1] E. C. Pielou, “The use of point-to-plant distances in the study of the pattern of plant populations,” *Journal of Ecology*, vol. 47, no. 3, pp. 607–613, 1959.
- [2] B. W. Silverman, *Density Estimation For Statistics and Data Analysis*, Chapman & Hall, London, UK, 1986.
- [3] B. D. Ripley, “The second-order analysis of stationary point processes,” *Journal of Applied Probability*, vol. 13, pp. 255–266, 1976.
- [4] J. E. Besag, “Comments on Ripley’s paper,” *Journal of the Royal Statistical Society*, vol. B 39, no. 2, pp. 193–195, 1977.
- [5] R. Law, J. Illian, D. F. R. P. Burslem, G. Gratzner, C. V. S. Gunatilleke, and I. A. U. N. Gunatilleke, “Ecological information from spatial patterns of plants: insights from point process theory,” *Journal of Ecology*, vol. 97, no. 4, pp. 616–628, 2009.
- [6] J. Cuzick and R. Edwards, “Spatial clustering for inhomogeneous populations,” *Journal of the Royal Statistical Society*, vol. B 52, no. 1, pp. 73–104, 1990.
- [7] P. J. Diggle and A. G. Chetwynd, “Second-order analysis of spatial clustering for inhomogeneous populations,” *Biometrics*, vol. 47, no. 3, pp. 1155–1163, 1991.
- [8] A. J. Baddeley, J. Møller, and R. Waagepetersen, “Non- and semi-parametric estimation of interaction in inhomogeneous point patterns,” *Statistica Neerlandica*, vol. 54, no. 3, pp. 329–350, 2000.
- [9] P. J. Diggle, V. Gómez-Rubio, P. E. Brown, A. G. Chetwynd, and S. Gooding, “Second-order analysis of inhomogeneous spatial point processes using case-control data,” *Biometrics*, vol. 63, no. 2, pp. 550–638, 2007.
- [10] G. Duranton and H. G. Overman, “Testing for localization using micro-geographic data,” *Review of Economic Studies*, vol. 72, no. 4, pp. 1077–1106, 2005.
- [11] M. Brühlhart and R. Traeger, “An account of geographic concentration patterns in Europe,” *Regional Science and Urban Economics*, vol. 35, no. 6, pp. 597–624, 2005.
- [12] E. Marcon and F. Puech, “Measures of the geographic concentration of industries: improving distance-based methods,” *Journal of Economic Geography*, vol. 10, no. 5, pp. 745–762, 2010.
- [13] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan, *Statistical analysis and modelling of spatial point patterns*, Wiley-Interscience, Chichester, UK, 2008.
- [14] G. Lang and E. Marcon, “Testing randomness of spatial point patterns with the Ripley statistic,” *ESAIM: Probability and Statistics*.
- [15] P. Haase, “Spatial pattern analysis in ecology based on Ripley’s K function: Introduction and methods of edge correction,” *Journal of Vegetation Science*, vol. 6, no. 4, pp. 575–582, 1995.
- [16] B. D. Ripley, *Statistical Inference For Spatial Processes*, Cambridge University Press, 1988.
- [17] E. Marcon and F. Puech, “Evaluating the geographic concentration of industries using distance-based methods,” *Journal of Economic Geography*, vol. 3, no. 4, pp. 409–428, 2003.
- [18] S. P. Kingham, A. C. Gatrell, and B. Rowlingson, “Testing for clustering of health events within a geographical information

- system framework," *Environment & Planning A*, vol. 27, no. 5, pp. 809–821, 1995.
- [19] A. C. Gatrell and T. C. Bailey, "Interactive spatial data analysis in medical geography," *Social Science and Medicine*, vol. 42, no. 6, pp. 843–855, 1996.
 - [20] A. C. Gatrell, T. C. Bailey, P. J. Diggle, and B. S. Rowlingson, "Spatial point pattern analysis and its application in geographical epidemiology," *Transactions of the Institute of British Geographers*, vol. 21, no. 1, pp. 256–274, 1996.
 - [21] N. C. Kenkel, "Pattern of self-thinning in jack pine: testing the random mortality hypothesis," *Ecology*, vol. 69, no. 4, pp. 1017–1024, 1988.
 - [22] F. Goreaud and R. Péliissier, "Avoiding misinterpretation of biotic interactions with the intertype K 12-function: population independence versus random labelling hypotheses," *Journal of Vegetation Science*, vol. 14, no. 5, pp. 681–692, 2003.
 - [23] N. B. Loosmore and E. D. Ford, "Statistical inference using the G or K point pattern spatial statistics," *Ecology*, vol. 87, no. 8, pp. 1925–1931, 2006.
 - [24] P. J. Diggle, *Statistical Analysis of Spatial Point Patterns*, Edward Arnold, London, UK, 2003.
 - [25] S. Gourlet-Fleury, J. M. Guehl, and O. Laroussinie, *Ecology & Management of a neotropical rainforest*, Lessons drawn from Paracou, a long-term experimental research site in French Guiana, Elsevier, Paris, France, 2004.
 - [26] B. Matérn, "Spatial variation," *Meddelanden Från Statens Skogsforskningsinstitut*, vol. 49, no. 5, pp. 1–144, 1960.
 - [27] F. Goreaud, *Apports de l'analyse de la structure spatiale en forêt tempérée à l'étude de la modélisation des peuplements complexes [Ph.D. thesis]*, ENGREF, Nancy, France, 2000.
 - [28] T. Wiegand and K. A. Moloney, "Rings, circles, and null-models for point pattern analysis in ecology," *Oikos*, vol. 104, no. 2, pp. 209–229, 2004.
 - [29] K. Schiffrers, F. M. Schurr, K. Tielbörger, C. Urbach, K. Moloney, and F. Jeltsch, "Dealing with virtual aggregation—a new index for analysing heterogeneous point patterns," *Ecography*, vol. 31, no. 5, pp. 545–555, 2008.
 - [30] B. S. Rowlingson and P. J. Diggle, "Splancs: spatial point pattern analysis code in S-plus," *Computers and Geosciences*, vol. 19, no. 5, pp. 627–655, 1993.
 - [31] P. A. Jansen, F. Bongers, and P. J. Van Der Meer, "Is farther seed dispersal better? Spatial patterns of offspring mortality in three rainforest tree species with different dispersal abilities," *Ecography*, vol. 31, no. 1, pp. 43–52, 2008.
 - [32] J. H. Connell, "On the role of natural enemies in preventing competitive exclusion in some marine animals and in forest trees," in *Dynamics of Populations*, P. J. Den Boer and G. Gradwell, Eds., pp. 298–312, 1971.
 - [33] D. H. Janzen, "Herbivores and the number of species in tropical forests," *The American Naturalist*, vol. 104, no. 940, pp. 501–528, 1970.
 - [34] E. Marcon and F. Puech, "A typology of distance-based measures of spatial concentration," HAL, Working Paper no. halshs-00679993, 2012.
 - [35] T. Wiegand, K. A. Moloney, J. Naves, and F. Knauer, "Finding the missing link between landscape structure and population dynamics: a spatially explicit perspective," *American Naturalist*, vol. 154, no. 6, pp. 605–627, 1999.
 - [36] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2012.

APPENDIX J

Measures of the geographic concentration of industries: improving distance-based methods

Marcon, E. et F. Puech (2010). « Measures of the geographic concentration of industries: improving distance-based methods ». In : Journal of Economic Geography 10.5, p. 745–762.

Measures of the geographic concentration of industries: improving distance-based methods

Eric Marcon* and Florence Puech*,†

Abstract

We discuss a property of distance-based measures that has not been addressed with regard to evaluating the geographic concentration of economic activities. The article focuses on the choice between a probability density function of point-pair distances or a cumulative function. We begin by introducing a new cumulative function, M , for evaluating the relative geographic concentration and the co-location of industries in a non-homogeneous spatial framework. Secondly, some rigorous comparisons are made with the leading probability density function of Duranton and Overman (2005), Kd . The merits of the simultaneous use of Kd and M is proved, underlining the complementary nature of the results they provide.

Keywords: Geographic concentration, distance-based methods, K -density function, Ripley's K function, M function

JEL classifications: C40, C60, R12, L60

Date submitted: 8 February 2008 **Date accepted:** 28 September 2009

1. Introduction

Step back and ask, what is the most striking feature of the geography of economic activity? The short answer is surely *concentration*. Krugman (1991, 5)

As highlighted by Paul Krugman (1991) in the first pages of his book *Geography and Trade*, economic activities are definitely not homogeneously distributed. During the last decade, the appraisal of the degree of spatial concentration of economic activities has received an increasing amount of attention. Economists have improved the measurement of the geographic concentration of economic activities in several ways and various criteria have been suggested for a 'good concentration index' (Combes and Overman, 2004). Up to now, the measure that respects the largest number of these properties is the K -density function (denoted Kd) proposed by Duranton and Overman (2005). The Kd function (i) compares the geographic concentration results across industries, (ii) controls for industrial concentration, (iii) controls for the overall aggregation patterns of industries, (iv) tests the significance of the results and (v) keeps the empirical results unbiased across geographic scales.

The aim of this article is to discuss an additional important property of distance-based measures that has not yet been addressed in the economic literature.

*AgroParisTech ENGREF, UMR EcoFoG, BP 316, 97310 Kourou, French Guyana.

†Corresponding author: Florence Puech, LET (Université de Lyon, CNRS, ENTPE), Institut des Sciences de l'Homme, 14 av. Berthelot, 69363 Lyon Cedex 07, France. email <Florence.Puech@univ-lyon2.fr>

This is the choice between using a probability density function of point-pair distances or a cumulative function for evaluating geographic concentration. Surprisingly, the preference for one or the other has never been discussed in empirical economic articles.¹ The only exception is Duranton and Overman (2005) who claimed in the conclusion of their article that probability density functions reveal more information than cumulative functions. In this article, we shall demonstrate clearly that the two types of functions cannot be considered as substitutes for each other. To do this, we shall propose a new function, that we shall call the *M* function, that, like the *Kd* function, respects the five criteria listed previously. However, while *Kd* is a probability density function of point-pair distances because it is calculated on the basis of the average number of neighbours at each distance, the *M* function is cumulative, depending on the number of neighbours up to each distance.

We have reached two main conclusions. Firstly, both probability density functions and cumulative functions are viable approaches for analysing the geographic concentration of activities. Consequently, a switch to probability density functions is not compulsory in order to meet Duranton and Overman's criteria. Secondly, the simultaneous use of *Kd* and *M* is recommended because they provide complementary results concerning the distribution of economic activities.

Our study is organized as follows. Section 2 presents an overview of cumulative distance-based measures. The section after this introduces the *M* function and discusses its mathematical properties. Next comes a comparison with other similar distance-based methods in order to demonstrate the value of these two new measures (Section 4). The last section concludes.

2. An overview of distance-based measures

In order to evaluate the geographic distribution of establishments, economists have traditionally employed cluster-based methods, which measure the spatial concentration of economic activity according to pre-defined geographic limits (regions, counties ...). It is now widely accepted that the measures obtained with these methods, such as the Gini and the Ellison and Glaeser (EG) indices,² introduce a statistical bias resulting from the chosen concept of space. Cluster-based methods zone the area in question: dividing space into a set of geographical units raises the well-known Modifiable Areal Unit Problem (MAUP), which can be summarized as follows: 'the result will be sensitive to the shape, size, and position of the areal units chosen' (Morphet 1997, 1039). The use of cluster-based methods is therefore problematic as they violate property (v).³ The solution to this problem is to use a continuous approach to space, thus switching from cluster-based methods to distance-based methods.

Distance-based measures are a relatively new way of gauging the geographic concentration of activities (Arbia and Espa, 1996; Marcon and Puech, 2003; Duranton and Overman 2005, 2008). The idea of these functions is simple. Unlike cluster-based

1 See for instance Fratesi (2008) for an empirical application of the *Kd* function or Arbia and Espa (1996), Barff (1987), Ó hUallacháin and Leslie (2007), Arbia et al. (2008), Jensen (2006) for different applications of cumulative distance-based methods.

2 For instance, see Combes et al. (2008) for a review of the main spatial concentration indices.

3 An empirical estimation of the shape and size bias resulting from different French area zonings may be found in Briant et al. (forthcomomg).

methods, distance-based methods do not zone the area in question in a specific manner but consider continuous space. Basically, each plant in the sample is localized by its coordinates (x,y) and the Euclidean distances between plants are considered. Unlike measures that only describe the location of economic activity on a single scale, distance-based methods can detect spatial structure at every scale. The main advantage of these methods is that they detect the distance at which significant geographic concentration or dispersion of establishments occurs. Two types of distance-based methods currently coexist in the economic literature:

- (i) The probability density function of point-pair distances that is based on the average number of neighbours at each distance (r). This measure is then smoothed and normalized so that it sums up to 1.
- (ii) Cumulative distance-based methods that describe geographic concentration by counting the average number of neighbours of plants on a disc, i.e. 'within' a circle of a given radius (r). This operation is then repeated for all possible radii.

The Kd function of Duranton and Overman (2005) is the only probability density function used in economic geography. Before their work, cumulative functions were used systematically in studies. Ripley's K function (1976, 1977), Besag's L function (1977) and their extensions based on the second-order property of point patterns are used in empirical applications.⁴ However, while Ripley's functions are now widely applied in other scientific fields such as forestry and ecology,⁵ Marcon and Puech (2003) have pointed out that they do not respect Duranton and Overman's five criteria (2005) for wide application in economics. Firstly, Ripley's function measures absolute concentration: it is based on the null hypothesis of a completely random spatial distribution of establishments (i.e. plants are distributed uniformly and independently). In spite of the longstanding debate about implementing absolute or relative measures (see Haaland et al., 1999), relative measures are still more widely used. Comparing a sector distribution to that of the whole of industry is even one of the theoretical criteria—property (iii)—defined by Duranton and Overman (2005). Relative measures detect whether each industry is overrepresented or underrepresented with respect to a baseline distribution, for example, the overall location pattern of industries. In other words, statistical tools based on relative concentration effectively measure the existence of specialized areas.⁶ Secondly, Ripley's function does not control for industrial concentration, i.e. the productive concentration within an industry among plants belonging to the sector in question—property (ii)—as every establishment is considered to be a point on the plot, regardless of its size.

These two problematic points must be answered to permit direct comparisons between a cumulative function and Kd . In the next section, we propose two versions of a new statistical tool, the M function, for the measurement of intra- and inter-industry geographic concentration. We have called it the M function because it is an extension of

4 For instance see Barff (1987), Arbia (1989), Ó hUallacháin and Leslie (2007) and Arbia et al. (2008) for different applications of cumulative distance-based methods to describe the geographic concentration of industries.

5 A survey of empirical studies in ecology or forestry using the K or L function can be found in Puech (2003, 324).

6 A discussion on the limits of relative indices can be found in Appendix A of Mori et al. (2005).

the existing cumulative distance-based methods, namely Ripley's K function (1976, 1977) and Besag's L function (1977).

3. Improving Ripley's functions: an introduction to the M function

In what follows, we shall first give an intuitive presentation of the common framework. We shall then successively define the M function and discuss its properties.

3.1. An intuitive analysis

Our relative measure compares the location patterns of an economic sector to that of aggregate activity (represented by all sectors). For this, we develop a cumulative function that counts neighbouring points up to a chosen distance denoted r . Let us consider a map with points on it that represent plants. We choose:

- (i) a reference point type, say a specific sector, and
- (ii) a target neighbour type called T : the same sector for intra-industry concentration or another sector for inter-industry concentration.

The average number of target neighbours is compared to a benchmark to detect whether they are more or less frequent than if plants were distributed randomly and independently from each other. To control for variations in the local density of points, each number of target neighbours (T_i around a point i) is normalized by the number of all the neighbours in the same area (N_i). Around each reference point we obtain a ratio of target neighbours (T_i/N_i) within the distance r from each point i . The average of this ratio ($\overline{T_i/N_i}$) is compared to the global ratio of the target type (T/N) calculated for the entire area. If $\overline{T_i/N_i}$ is greater than T/N , we conclude that more plants of the target type are observed within a distance r around the reference points type than on average, if circles of radius r were drawn anywhere. In other words, target points are concentrated around reference points. The ratio $M = (\overline{T_i/N_i})/(T/N)$ will be used in our analysis for convenience because the benchmark is equal to one. M values are computed over a large range of distances and presented as a continuous function of r on a graph including confidence intervals for the null hypothesis of independence of plant locations (significance is checked by appropriate statistical tests). As a result of these successive normalizations any value of M can be interpreted immediately and compared across sectors and distances. Finally, points can also be weighted, counting, for example, the number of employees instead of the number of plants. Finally, significance tests must properly control for the non-independence of their distribution (i.e. industrial concentration).

3.2 Evaluating geographic intra-industry concentration

In mathematical terms, let us consider an area A containing a total of N plants belonging to a variety of industries. We shall focus on a particular industry S where N_S is the total number of establishments from that sector in area A . The description of the neighbourhood of the N_S plants follows. Consider a dummy variable $c_S(i,j,r)$ that is equal to 1 if the Euclidean distance between the two plants i and j from the sector S is less than the radius r ($c_S(i,j,r)=0$ otherwise). The number of neighbouring establishments of plant i , belonging to the same sector and located within a distance

r from it, is thus $\sum_{j=1, i \neq j}^{N_S} c_S(i, j, r)$. In the same way, we define the dummy $c(i, j, r)$ as equal to 1 if plant j (whatever its industry) is located at a distance inferior or equal to r from the establishment i (the dummy's value is 0 otherwise). Consequently, the number of establishments located at most at a distance r from business unit i is: $\sum_{j=1, i \neq j}^N c(i, j, r)$. Plant size may now be included in our analysis. The weight associated with each dummy is that of the neighbouring plant j , and it is denoted w_j . The weight associated with the plants may, for example, be their number of employees. The average proportion of employees of industry S within a given radius r is clearly:

$$\frac{1}{N_S} \sum_{i=1}^{N_S} \frac{\sum_{j=1, i \neq j}^{N_S} c_S(i, j, r) w_j}{\sum_{j=1, i \neq j}^N c(i, j, r) w_j}$$

In the same way, we can define the ratio of employees in industry S in the entire area A compared to the whole of industry by: $\frac{1}{N_S} \sum_{i=1}^{N_S} \frac{W_S - w_i}{W - w_i}$ where W_S is the total number of industry S employees in the area A ; W is the total number of employees in aggregate activity and w_i is the weight of plant i .⁷ The ratio of the above quantities, averaged over all the establishments in sector S , defines the M function for the intra-industry geographic concentration of sector S as:

$$M_S(r) = \sum_{i=1}^{N_S} \frac{\sum_{j=1, i \neq j}^{N_S} c_S(i, j, r) w_j}{\sum_{j=1, i \neq j}^N c(i, j, r) w_j} \bigg/ \sum_{i=1}^{N_S} \frac{W_S - w_i}{W - w_i} \quad (1)$$

The numerator corresponds to the relative weight of sector S in comparison with the whole industrial within circles of radius r . The denominator represents the relative size of the considered sector in comparison with all activities in area A . The benchmark for the M function is 1 and this defines the same location pattern for the specific sector as for aggregate activity. This means that whatever the considered radius, there are proportionally as many employees who belong to sector S as there are in the whole of area A . Thus, M values superior to 1 ($M_S(r) > 1$) indicate that there are proportionally more employees close to plants in sector S (within a distance r) than in the whole area. This corresponds to the definition of the relative geographic concentration of sector S at distance r . In contrast, the relative geographic dispersion of sector S at distance r is defined by $M_S(r) < 1$, indicating that there are relatively fewer employees in sector S within a distance r around the establishments than in the whole area. One can see that interpreting the M values is straightforward. For instance, $M_S(r) = 2$ indicates that, within a particular distance r , the relative density of employees in sector S is double that in the whole area. In the same way, $M_S(r) = 0.5$ indicates that within a given distance r around sector S plants the density of employees in this sector is on average half that of the whole area.

We shall now consider how the significance of the results can be tested and how we can control for the industrial concentration. Two types of confidence intervals for the null hypothesis are generated: local and global. The null hypothesis is that establishments belonging to sector S are located according to the same pattern as the others. To test this, we generate a series of random and independent distributions of the plant

⁷ w_i must be subtracted from W_S and W since we count the number of establishments around the plant i within a radius r : the reference establishment i itself should not be counted.

dataset based on the actual set of possible locations and the industry/plant size pairs (i.e. the industrial concentration as given). The local confidence interval is determined using the Monte–Carlo method. In practice, we generate a large number of simulations and choose a confidence level, say 5%. The 95% confidence interval of the M function for each value of r is bounded by the outer 5% of the randomly generated values. Nearly all empirical studies that use Ripley’s K function (or one of its extensions) compute only the local confidence intervals to test the significance of the results. However, Duranton and Overman (2005) have recently criticized the computation of local confidence intervals on their own, considering them as too ‘optimistic’, and highlighted the need for the global confidence intervals of the null hypothesis as well. If we assume that the values of the M function at different radii are independently distributed, one would expect a proportion of them equal to the confidence threshold to be outside the confidence interval even though the point process corresponds to the null hypothesis. For instance at a 5% threshold, complete spatial randomness should not be rejected when 5 points in a 100-point plot are outside the confidence interval. Successive values of Ripley’s functions are actually highly correlated: the risk of erroneous rejection of the null hypothesis is consequently reduced but cannot be quantified. A global confidence interval is defined such that the confidence threshold is the risk that the plot of a function generated by the null hypothesis exceeds the interval at least once. This may be chosen in many ways but it should have an equal weight at all distances. A simple method of computing global confidence intervals is to generate local confidence intervals at increasing confidence levels until the ratio of simulated plots that lie partly outside them reaches the predefined threshold. As an example, let us suppose that 1000 plots have been generated and the confidence level has been set at 5%. The outer values at each distance are eliminated, defining a local confidence interval at $2/1000 = 0.2\%$. The plots that lie partly outside this interval are counted. If there are 10 such plots, the global confidence level will be 1%. The process is repeated until the threshold is reached. If it is not reached exactly, interpolation is performed.

Five fundamental criteria characterizing a ‘good’ economic measure of geographic concentration were presented in the introduction. The M function respects them all. Moreover, one can note that an appreciable property of this index is that its values may be interpreted. Additionally, the M function can be calculated for any topology. Ripley’s function and its developments (see Goreaud and Pélissier, 1999) require edge-effect correction for points that are close to borders. Complex geographical shapes are consequently intractable hence the domain must always be a polygon or a disc.⁸ The M function provides an answer to this problem: comparing the number of neighbours in a certain industry to the total number of neighbouring establishments ‘in the same area’ avoids any need for correction. Working on complex geographical limits, such as national borders, is now possible. This last consideration justifies the somewhat complex computation of the M function. Software can be downloaded from the authors’ website⁹ to facilitate its implementation.

8 Sweeney and Feser (1998, 52) Figure 1, or Feser and Sweeney (2000, 361) Figure 2; Pancer-Koteja et al. (1998, 757) Figure 1; Rowlingson and Diggle (1993, 634) Figure 5.

9 Available online at: <http://e.marcon.free.fr/Ripley> (English and French versions).

3.3 Evaluating the co-location of industries

If the researcher suspects interactions between them it may be interesting to evaluate the co-location of different industries. The geographic concentration of different industries can be investigated using the inter-industry version of the M function that has the same properties as the intra-industry version. In what follows, we shall consider co-location between two sectors denoted S_1 and S_2 . A complete description of the spatial distribution of the co-location patterns of these industries leads not to one but to two definitions of the M function. The first, M_{S_1,S_2} , depicts the spatial distribution of plants belonging to sector S_2 around those of sector S_1 in non-homogenous space. The second function, M_{S_2,S_1} , describes the spatial distribution of plants belonging to sector S_1 around those of sector S_2 . The meaning of the co-location M functions is thus simple: they test whether the relative density of employees from one sector around establishments of another sector is on average greater or lesser than in the whole area.

Let us consider the same area A using the same notations as in the previous section. We shall now examine the Euclidean distances between plants belonging to two different industries. First, we shall consider the definition of M_{S_1,S_2} in which the reference plants (i.e. those at the centre of the circles) are from industry S_1 . The definition of the $M_{S_1,S_2}(r)$ co-location function is thus:

$$M_{S_1,S_2}(r) = \sum_{i=1}^{N_{S_1}} \frac{\sum_{j=1}^{N_{S_2}} c_{S_2}(i,j,r)w_j}{\sum_{n=1, i \neq n}^N c(i,n,r)w_n} / \sum_{i=1}^{N_{S_1}} \frac{W_{S_2}}{W - w_i} \quad (2)$$

The value of Equation (2) shows whether the relative density of plants S_2 located around those of sector S_1 is greater ($M_{S_1,S_2}(r) > 1$) or lesser ($M_{S_1,S_2}(r) < 1$) than in the entire area A . In the same manner, we can define the function $M_{S_2,S_1}(r)$ that describes the spatial structure of the S_1 plants located around those of sector S_2 . The alterations to the function are obvious:

$$M_{S_2,S_1}(r) = \sum_{i=1}^{N_{S_2}} \frac{\sum_{j=1}^{N_{S_1}} c_{S_1}(i,j,r)w_j}{\sum_{n=1, i \neq n}^N c(i,n,r)w_n} / \sum_{i=1}^{N_{S_2}} \frac{W_{S_1}}{W - w_i} \quad (3)$$

Concerning the significance of the results, both the local and global confidence intervals for the null hypothesis are computed but we shall pay particular attention to the null hypothesis. Monte-Carlo techniques are used to generate simulated distributions (the threshold and the number of simulations are exogenous). Nevertheless, the null hypothesis has to eliminate the sector-specific patterns in order to detect only interactions between the two industries. For instance, if S_1 is highly aggregated and S_2 completely randomly distributed, the relative importance number of S_2 plants around S_1 establishments is low and artificial segregation is detected. Under these conditions, the null hypothesis must control for the patterns of both S_1 and S_2 . The solution is as follows. The null-hypothesis set of plants for M_{S_1,S_2} is generated by keeping the S_1 establishments fixed and redistributing all the other plant size/sector pairs between all the other locations, thus controlling for the pattern of S_1 . To be sure that the structure of the industry S_2 is not responsible for any under- or over-estimation of the density of the plant's employees, we also need

to control for the structure of S_2 : the same process applied to M_{S_2, S_1} controls for the pattern of S_2 plants. Lastly, it is accepted that there is a significant interaction if both values are significantly different from their respective null hypothesis. Note that the null hypothesis excludes the detection of a ‘multi-concentration’ phenomenon: a situation in which we could observe a significant co-location that does not result from an interaction between these two industries (this would be the case for instance if both industries locate around the plants of another industry). This is undoubtedly a limitation of the inter-industry M function and is shared by all other distance-based functions.¹⁰

4. Towards an unified framework for distance-based methods

At this stage, we can compare the statistical properties of distance-based methods and especially those of the two leading geographic concentration measures namely Kd and M .¹¹ Our aim is to reveal in which cases one measure should be preferred to the other. As underlined previously, our discussion focuses on the implications of using probability density functions rather than cumulative functions because this remains the main difference between Kd and M .

4.1 Common statistical framework

Historically, methods of characterizing the structure of point processes as a function of bilateral distances between pairs of points have been developed by Ripley (1976, 1977).

Ripley defined the function $g(r)$ as the ratio of the probabilities of finding two points at a distance r from each other to the product of the probabilities of finding each of them. If points are distributed independently, $g(r) = 1$; higher values show that point pairs at this distance are more frequent than under the null hypothesis of independence. The integral function $K(r) = \int_{\rho=0}^r g(\rho) 2\pi\rho \, d\rho$ is easy to estimate.¹² Assuming the point density is uniform on an area A and denoting by N the total number of points on the domain A , we find (like Sweeney and Feser, 1998, for example):

$$\hat{K}(r) = \frac{A}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j>i}^N c(i,j,r) \quad (4)$$

In a space where edge effects do not occur (say a torus), $c(i,j,r)$ is a dummy: its value is 1 if the distance between points i and j is less than r . In the real world, it is corrected for edge effects: when point i is close to the boundary of the domain, it has fewer neighbours because those outside the domain are not observed. After computing \hat{K} , $g(r)$ can be estimated by $\hat{g}(r) = \hat{K}(r + \Delta r) - \hat{K}(r - \Delta r) / 2\Delta r$, taking Δr as arbitrarily small. \hat{g} is proportional to the number of point pairs whose distance is close to r .

10 The term ‘co-localization’ is used by Duranton and Overman (2005) when co-agglomeration results from attraction on the part of both industries whereas they prefer the term ‘joint-localization’ when both industries exhibit significant co-agglomeration for another reason. However, they mention that they cannot disentangle these location patterns empirically.

11 As stated in the introduction, the debate about implementing absolute or relative measures for evaluating geographic concentration has recently been settled, as one of the criteria of a ‘good’ concentration measure states that relative indices must be preferred (Combes and Overman, 2004; Duranton and Overman, 2005). This article does not question this view.

12 See Marcon and Puech (2003) for a concise presentation of this function.

K is a cumulative function, while g is a local function. \hat{K} has been widely used in the literature, but \hat{g} has not. Both are restricted to homogenous point processes. Further mathematical developments are necessary in order to characterize inhomogeneous point sets and to control for the spatial distribution of the whole economy. Duranton and Overman (2005) chose to define the Kd function as the probability density function of point-pair distances. Kd is also proportional to the number of point pairs whose distance is close to r . The differences between Kd and \hat{g} are: (i) Kd integrates appropriate smoothing, but this is only a technical improvement, and (ii) Kd does not correct for any edge effects. Its value is compared to those of point distributions with the same geometry, which have the same edge effects. We chose another approach. From Equation (4) and using the same definitions of T_i and T as those given in Section 2.1., it follows that the expression for \hat{K} can be rearranged as:

$$\frac{\hat{K}(r)}{\pi r^2} = \frac{\sum_{i=1}^{N-1} \frac{\sum_{j>i}^N c(i,j,r)}{\pi r^2}}{N} \bigg/ \frac{N-1}{A} = \frac{\sum_{i=1}^N T_i/N_i}{N} / T/N \quad (5)$$

It is thus a particular case of M .

To summarize, all these functions are derived from the raw data that is the number of point pairs at a given distance. Ripley's \hat{g} is normalized so that its value is 1 when points are distributed independently. Duranton and Overman's Kd is normalized to be a probability density function. Ripley's \hat{K} is the cumulative function of \hat{g} . It can be interpreted as the ratio of the observed number of neighbours to the number of neighbours there would be if the points were distributed independently. The M function is its generalization in non-homogenous space.

4.2 Kd and M functions as complementary measures of geographic concentration

In what follows, we need to re-examine a well-known dilemma, namely whether a probability density function or a cumulative function should be used. Some examples will be given to illustrate the advantages and limitations of both measures. We shall then explain why Kd and M can be considered as useful complements to each other in economic geography.

4.2.1 Fundamental differences between Kd and M

4.2.1.1. Property 1: like any probability density function, Kd detects local density more precisely at different spatial scales. The idea here is simple. A probability density function like Kd can detect structures at specific spatial scales and thus easily identifies local patterns. Like any cumulative function, M accumulates spatial information on the distribution of points up to a certain scale: its local density estimates then become less precise (Condit et al., 2000). Duranton and Overman (2005) underlined this important property when they asserted in the conclusion of their article that ‘ K -densities are more informative than [Ripley's] K -functions with respect to the scale of localization’. In their working article, the reason for preferring a probability density function is even clearer:

Traditionally, spatial statisticians have used the cumulative of the K -density, the K -function. [...] Given that a major objective of our analysis is to distinguish the spatial scale(s) at which

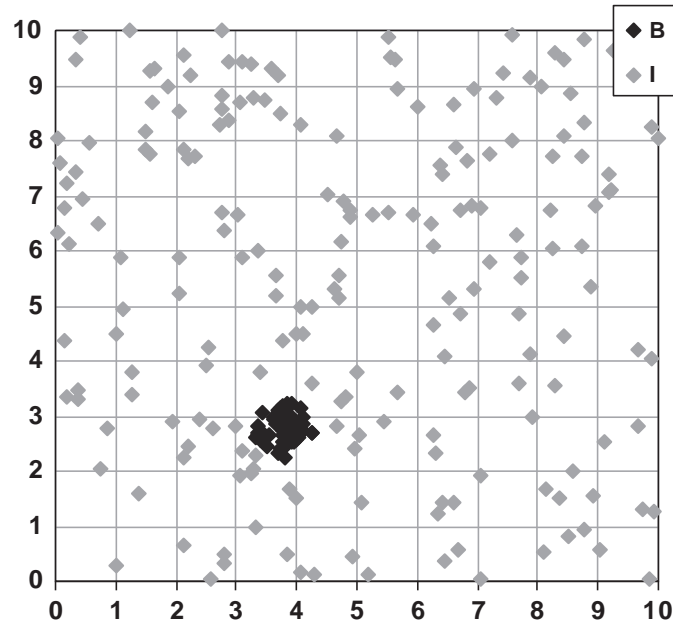


Figure 1. First theoretical distribution of establishments on a 10×10 area.

excess-localisation takes place, the focus on the density distribution rather than its cumulative is warranted. (Duranton and Overman, 2001, 7)

Let us turn to an example to understand what issue is at stake. Consider two industries localized on a 10×10 territory. The first industry I has 200 production units randomly distributed (Poisson distribution) over the whole area. The second industry B is generated by a Matérn process (Matérn, 1960, cited by Stoyan et al., 1987): 50 points are uniformly generated in one cluster of radius 0.5. The location of establishments of both industries is shown in Figure 1. As in the other illustrative theoretical cases elsewhere in the article, the weight of each plant is equal to 1 to simplify the examples and the confidence intervals are computed at a 1% threshold, from 10,000 simulations (only global confidence intervals are shown in the figures). Kd , M and global intervals are computed at intervals of 0.5 up to a radius of 10.¹³ The Kd and M functions for industry B are respectively given in Figures 2 and 3. Finally, note that Duranton and Overman (2005) recommend analysing the spatial pattern up to the median distance between all pairs of plants. However, in what follows, the results are given for all possible radii to describe completely the behaviour of the functions (even though this should not be done in empirical studies).

What can we learn from this example? As we can see in Figures 2 and 3, the first positive significant peak of the Kd and the M functions appears at a radius of 0.5.¹⁴ At this distance, both measures successfully detect the circular cluster of the Matérn

¹³ It is interesting to note that $Kd(0)$ is defined by smoothing, whereas $M(r)$ may be undefined, if no plant has any neighbor less than r apart.

¹⁴ Note that contrarily to Kd , M values are not defined below $M(0.5)$. As we previously underlined, this result is coherent: as any cumulative function, M does not use smoothing techniques (as Kd does) so it is not defined for distances lower than that of the closest point pair.

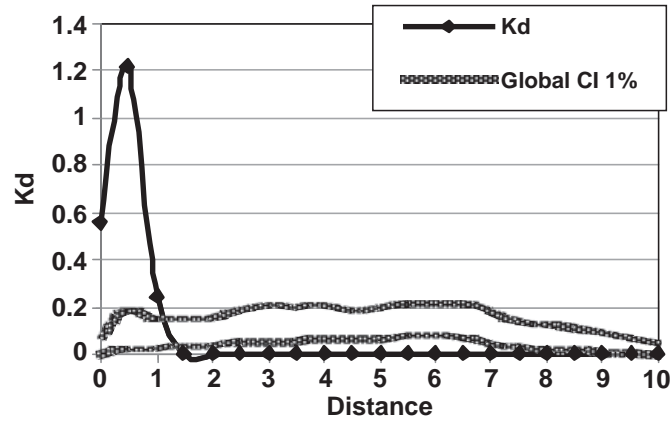


Figure 2. Kd function for industry B.

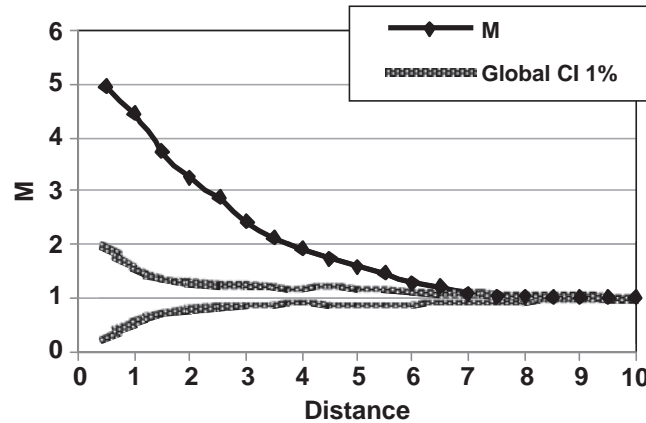


Figure 3. M function for industry B.

process. However beyond the distance of 0.5, the Kd and M plots are clearly different. Since there are no industry B neighbours beyond a radius of 0.5, the Kd values rapidly decrease while the M plot gradually (and not suddenly like Kd) returns to within the confidence interval bands. The M plot gives the impression of being less precise than the Kd plot because its slope is not steep beyond a radius of 0.5. However, the M plot is completely in accordance with the definition of a cumulative function and there is no misinterpretation. M perfectly detects the lack of industry B neighbours' because the M values decrease beyond a distance of 0.5. However, the Kd results seem more intuitive because for any cumulative function 'aggregation at smaller scales influences the estimate at larger scales' (Wiegand and Moloney, 2000, 220).

4.2.1.2. Property 2: like any cumulative density function, the M function identifies spatial structures of point patterns better. This property highlights the fact that only a cumulative function can reveal a superposition of different spatial point patterns. We illustrate this second property with two theoretical examples: independent distribution of clusters and spatial repulsion between clusters. In the first example, the clusters of an industry C are generated by a Matérn process: 50 plants are uniformly generated in 9 clusters of radius 0.5 randomly distributed on a 10×10 domain.

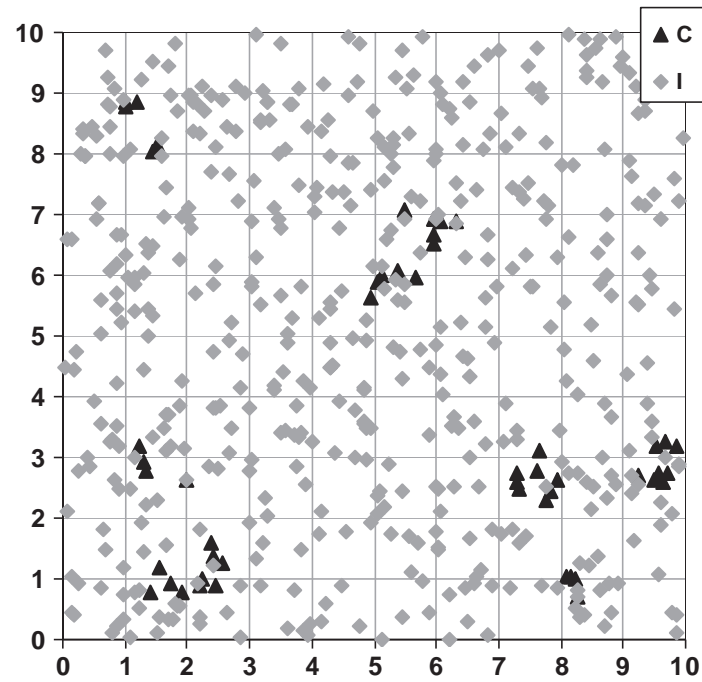


Figure 4. Second theoretical distribution of establishments on a 10×10 area.

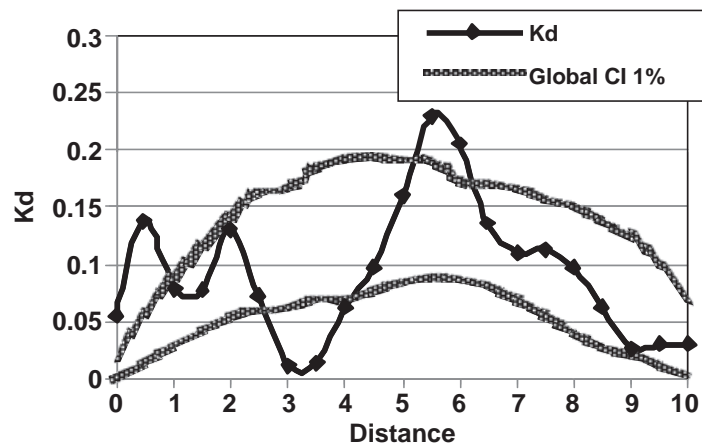


Figure 5. Kd function for industry C.

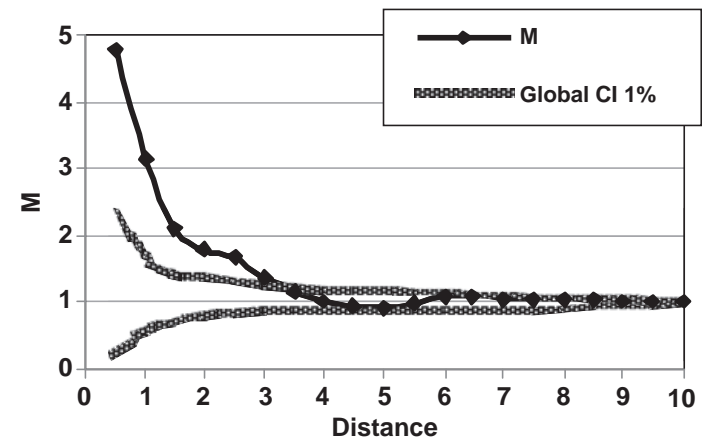


Figure 6. M function for industry C.

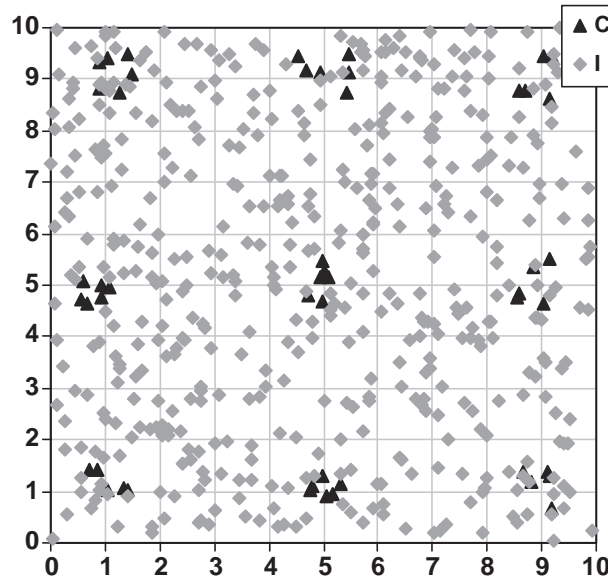


Figure 7. Third theoretical distribution of establishments on a 10×10 area.

500 establishments of another industry (called I) are randomly distributed according to a Poisson distribution. The map of the distribution of plants is given in Figure 4 and the results of Kd and M for industry C are shown in Figures 5 and 6. In the second example, the same industry I composed of 500 randomly distributed plants is located in the same area. However, the 50 establishments of industry C are now located in 9 clusters regularly distributed on a squared grid (Figure 7).

To begin with, let us consider the Kd plots for industry C (Figures 5 and 8). A first significant peak is detected within a radius equal to 0.5, corresponding to the size of the clusters. The M function corroborates the findings with the Kd function. At larger distances, differences between the plots appear. The positive significant peaks of both Kd functions reveal several excessive point-pair distances corresponding to the relative position of aggregates.¹⁵ Between the clusters, both Kd plots indicate that there is a lack of industry C plants: it can be shown in Figures 5 and 8 that negative significant peaks emerge around a radius of 1.5–3. Without looking at the maps, the information given by the Kd values is insufficient for us to tell whether the observed lack of point-pair distances is caused by repulsion between clusters or a random distribution. This is a clear and important weakness of Kd . Estimations of the M function solve the dilemma: in the second case ‘only’, significant repulsion is detected between 2.5 and 4 (and between 7 and 8.5). The M plot is below the confidence interval at these distances, detecting the regular position of the clusters on the grid.

4.2.1.3 Property 3: M values are easier to interpret. Kd and M evaluate the spatial distribution of plants and, for every distance, summarize the location patterns at a certain level of concentration. It would be of value to a researcher for the results to be easy to understand. Ideally, the values that are obtained should be comparable (i) at several radii (ii) across sectors and (iii) for several points in time. In this respect, M is

15 For instance, in Figure 8, note that the first neighbours of industry C plants outside the cluster appear at a radius approximately equal to 4 (the grid size) and then around 8 (twice the grid size).

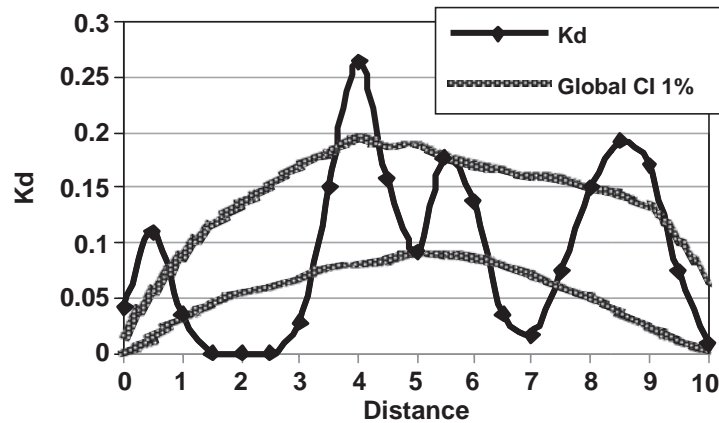


Figure 8. Kd function for industry C.

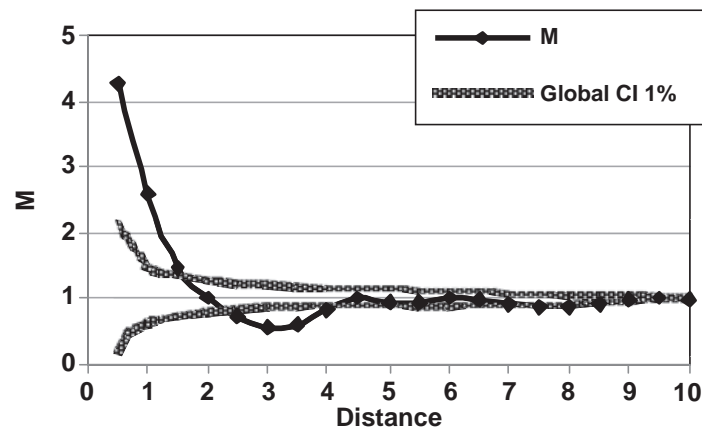


Figure 9. M function for industry C.

undoubtedly more useful than Kd . M function values can effortlessly be interpreted and compared whatever the distance. To give an example, let us consider the case where the M plot is outside the confidence interval (centered on 1) and M reaches the value of 5 at a given radius. This means that the relative density of neighbours from the considered industry ‘within this radius’ is 5-times higher than it would be if its plants were distributed as they are in the whole industry. This value is a measure of concentration, so it can be understood independently from the data it was calculated from. The Kd values are the probability density of a plant having a neighbour at a given distance. They depend both on the geometry of the plant distribution and the deviation of the sector under scrutiny from the benchmark distribution. The Kd values are consequently unsuitable. Duranton and Overman (2005) suggest using the difference between Kd and the upper $\overline{Kd(r)}$ or lower $\underline{Kd(r)}$ band of the confidence interval as a normalized measure. Nevertheless, the resulting normalized values still have no meaning.

4.2.2 Implications for measuring the geographic concentration of economic activities

We can now examine the implications of the above properties for the field of economic geography. Basically, like any geographic concentration measure, both Kd and M could

be useful for resolving two types of issues. The first concerns the characterization of the spatial distribution of economic activity. The second is related to identifying the determinants of agglomeration.

- (i) Kd and M are useful for exploring the location patterns of economic activities.

As a general rule, for analysing a point pattern in economics we suggest first using M to have a global view of the spatial structure of the distribution, and then Kd to obtain more detailed information. There are three reasons for this. Firstly, Kd provides more precise results for the local density of establishments (Property 1). Secondly, a density function such as Kd is not able to evaluate the global effect of the superposition of spatial structures (Property 2). The absence of neighbour plants at a given distance may be the consequence of repulsion (Figure 7) or just compensation for very strong attraction at another scale (Figure 8).¹⁶ Thirdly, we suggest in the theoretical examples that an overview of the map may avoid any misinterpretation of Kd . This technique is unhelpful for evaluating the geography of production because in real life things are more complex than in the theoretical cases considered above. Both the number of establishments and the number of sectors under study are higher, which means that looking at the map is less informative. Here too, the simultaneous use of both measures is advisable.

An exception to the general rule will be given by Property 3. If the only reason for using a geographic concentration index is to quantify the spatial deviation from locational randomness, the appropriate measure is undoubtedly the M function. The level of concentration of dispersion is certainly more intelligible for a decision-maker than any normalized results without concrete meaning.

- (ii) M seems more appropriate for understanding the determinants of the geographic concentration of activities.

Should the influence of the determinants of agglomeration be analysed *at* a given distance or *up to* a certain distance? The answer is at the heart of the dilemma between choosing M or Kd .¹⁷

The vast majority of previous studies that have analysed the determinants of agglomeration clearly used a cumulative approach. Authors systematically regress the observed level of a cluster-based measure of spatial concentration with respect to different local factors (Kim, 1995; Ellison and Glaeser, 1999; Rosenthal and Strange, 2001; Co, 2002). Such studies are based on a particular zoning of the area, individual data is thus aggregated up to a certain level (such as a region or state). Is the availability of data the only reason for this? We do not think so: the proof is given by authors who deliberately avoid pre-defined zoning, preferring the geometric form of a disc in order to evaluate the global impact of the surrounding plants. For example, Holmes

16 Duranton and Overman (2005, 1086) suggest studying the geography of production only up to an economically pertinent distance. They define this maximum distance as the median radius between all pairs of plants. It is interesting to note that our recommendation of using both measures is all the more valid in the example that illustrates Property 2 because the problematic negative peak of Kd plots appears below the median distance for industry C in Figure 7.

17 However, the way of testing the determinants of agglomeration is far from being settled. Several live debates still exist in the literature. For example, are concentration indexes the appropriate variables to answer this question (Combes et al., 2008, chapter 11)? If so, should a continuous-space framework or a discrete-space one be preferred (Ellison et al., forthcoming)?

(1999) estimates the link between vertical disintegration and the level of geographic concentration of plants by applying an integrative approach. However, probability density functions may also be useful but only in capturing the marginal effect of the factors of agglomeration. Some examples are given by Rosenthal and Strange (2003, 2008) who consciously use several concentric rings to estimate the spatial scope of externalities.

5. Conclusion

The aim of this article was to improve some existing relative statistical tools for testing the spatial concentration of industries. We have shown that M functions constitute first-class instruments for evaluating intra- or inter-industry geographic concentration. We have proved that the cumulative M function must be implemented with the probability density function Kd to give a complete and good description of the distribution of economic activities. Nonetheless, some intrinsic limits of these new tools suggest that other research approaches are needed to fill the gap between the theoretical and empirical literatures. Despite the considerable recent interest of researchers in an ‘ideal’ concentration index, and even though significant progress has been made, work still needs to be done to meet the most difficult criteria suggested by Combes and Overman (2004): the complete integration of the tools to economic theory, and the independence of geographic concentration measures from the industrial classification. In this article, we have enhanced existing distance-based methods but further investigations are still required.

Acknowledgements

We are grateful to Richard Arnott, Gilles Duranton, Pablo Jensen, John McBreen, many seminar and conference participants, the editor Henry Overman and three anonymous referees for their very helpful suggestions and comments.

References

- Arbia, G. (1989) *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Dordrecht: Kluwer.
- Arbia, G., Espa, G. (1996) *Statistica Economica Territoriale*. Padua: Cedam.
- Arbia G., Espa G., Quah, D. (2008) A class of spatial econometric methods in the empirical analysis of clusters of firms in the space. *Empirical Economics*, 34: 81–103.
- Barff, R. A. (1987) Industrial clustering and the organization of production: a point pattern analysis of manufacturing in Cincinnati, Ohio. *Annals of the Association of American Geographers*, 77: 89–103.
- Besag, J. E. (1977) Comments on Ripley’s paper. *Journal of the Royal Statistical Society B*, 39: 193–195.
- Briant, A., Combes, P.-P., Lafourcade, M. (forthcoming) Dots to boxes: do the size and shape of spatial units jeopardize economic geography estimations? *Journal of Urban Economics*.
- Co, C. Y. (2002) The agglomeration of US-owned and foreign-owned plants across the US States. *The Annals of Regional Science*, 36: 575–592.
- Combes, P.-P., Mayer, T., Thisse, J.-F. (2008) *Economic Geography: The Integration of Regions and Nations*. Princeton: Princeton University Press.
- Combes, P.-P., Overman, H. (2004) The spatial distribution of economic activities in the European Union. In J. V. Henderson, J.-F. Thisse (eds) *Handbook of Urban and Regional Economics*. North Holland, Amsterdam: Elsevier.

- Condit, R. et al. (2000) Spatial patterns in the distribution of tropical tree species. *Science*, 288: 1414–1418.
- Duranton, G., Overman, H. G. (2001) Localisation in UK Manufacturing Industries: Assessing Non-Randomness Using Micro-Geographic Data. Working Paper.
- Duranton, G., Overman, H. G. (2005) Testing for localisation using micro-geographic data. *Review of Economic Studies*, 72: 1077–1106.
- Duranton, G., Overman, H. G. (2008) Exploring the detailed location patterns of UK manufacturing industries using microgeographic data. *Journal of Regional Science*, 48: 213–243.
- Ellison, G., Glaeser, E. L. (1999) The geographic concentration of industry: does natural advantage explain agglomeration? *The American Economic Review, American Economic Association Papers and Proceedings*, 89: 311–316.
- Ellison, G., Glaeser, E. L., Kerr, W. R. (forthcoming) What causes industry agglomeration? Evidence from coagglomeration patterns. *The American Economic Review*.
- Feser, E. J., Sweeney, S. H. (2000) A test for the coincident economic and spatial clustering of business enterprises. *Journal of Geographical Systems*, 2: 349–373.
- Fratesi, U. (2008) Issues in the measurement of localization. *Environment and Planning A*, 40: 733–758.
- Goreaud, F., Pélissier, R. (1999) On explicit formulas of edge effect correction for Ripley's *K*-function. *Journal of Vegetation Science*, 10: 433–438.
- Haaland, J. I., Kind, H. J., Midelfart-Knarvik, K. H., Torstensson, J. (1999) What determines the economic geography of Europe? Centre for Economic Policy Research. Discussion paper, 2072.
- Holmes, T. J. (1999) Localization of industry and vertical disintegration. *The Review of Economics and Statistics*, 81: 314–325.
- Jensen, P. (2006) Network-based predictions of retail store commercial categories and optimal locations. *Physical Review*, E 74: 035101(R).
- Kim, S. (1995) Expansion of markets and the geographic distribution of economic activities: the trends in US regional manufacturing structure, 1860–1987. *The Quarterly Journal of Economics*, 110: 881–908.
- Krugman, P. (1991) *Geography and Trade*. London: MIT Press.
- Marcon, E., Puech, F. (2003) Evaluating the geographic concentration of industries using distance-based methods. *Journal of Economic Geography*, 3: 409–428.
- Matérn, B. (1960) Spatial variation. *Meddelanden från Statens Skogsforskningsinstitut*, 49: 1–144.
- Mori, T., Nishikimi, K., Smith, T. E. (2005) A divergence statistic for industrial localization. *Review of Economics and Statistics*, 87: 635–651.
- Morphet, C. S. (1997) A statistical method for the identification of spatial clusters. *Environment and Planning A*, 29: 1039–1055.
- Ó hUallacháin, B., Leslie, T. F. (2007) Producer services in the urban core and suburbs of Phoenix, Arizona. *Urban Studies*, 44: 1581–1601.
- Pancer-Koteja, E., Szwagrzyk, J., Bodziarczyk, J. (1998) Small-scale spatial pattern and size structure of *Rubus hirtus* in a canopy gap. *Journal of Vegetation Science*, 9: 755–762.
- Puech, F. (2003) Concentration géographique des activités industrielles: Mesures et enjeux. PhD thesis, Université de Paris I.
- Ripley, B. D. (1976) The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13: 255–266.
- Ripley, B. D. (1977) Modelling spatial patterns. *Journal of the Royal Statistical Society B*, 39: 172–212.
- Rosenthal, S. S., Strange, W. C. (2001) The determinants of agglomeration. *Journal of Urban Economics*, 50: 191–229.
- Rosenthal, S. S., Strange, W. C. (2003) Geography, industrial organisation, and agglomeration. *Review of Economics and Statistics*, 85: 377–393.
- Rosenthal, S. S., Strange, W. C. (2008) The attenuation of human capital spillovers. *Journal of Urban Economics*, 64: 373–389.
- Rowlingson, B. S., Diggle, P. J. (1993) SPLANCS: Spatial Point Pattern Analysis Code in S-Plus. *Computers and Geosciences*, 19: 627–655.

- Stoyan, D., Kendall, W. S., Mecke, J. (1987) *Stochastic Geometry and its Applications*. New York: John Wiley & Sons.
- Sweeney, S. H., Feser, E. J. (1998) Plant size and clustering of manufacturing activity. *Geographical Analysis*, 30: 45–64.
- Wiegand, T., Moloney, K. A. (2006) Rings, circles, and null-models for point pattern analysis in ecology. *Oikos*, 104: 209–229.

APPENDIX K

Assessing foliar chlorophyll contents with the SPAD-502 chlorophyll meter: a calibration test with thirteen tree species of tropical rainforest in French Guiana

Coste, S., C. Baraloto, C. Leroy, E. Marcon, A. Renaud, A. D. Richardson, J.-C. Roggy, H. Schimann, J. Uddling et B. Hérault (2010). « Assessing foliar chlorophyll contents with the SPAD-502 chlorophyll meter: a calibration test with thirteen tree species of tropical rainforest in French Guiana ». In : *Annals of Forest Science* 67.6, p. 607.

Assessing foliar chlorophyll contents with the SPAD-502 chlorophyll meter: a calibration test with thirteen tree species of tropical rainforest in French Guiana

Sabrina COSTE¹, Christopher BARALOTO¹, Céline LEROY², Éric MARCON³, Amélie RENAUD³,
Andrew D. RICHARDSON⁴, Jean-Christophe ROGGY¹, Heidy SCHIMANN¹, Johan UDDLING⁵,
Bruno HÉRAULT^{6*}

¹ INRA – Unité Mixte de Recherche Écologie des Forêts de Guyane, Kourou, France

² CNRS – Unité Mixte de Recherche Écologie des Forêts de Guyane, Kourou, France

³ AgroParisTech-ENGREF- Unité Mixte de Recherche Écologie des Forêts de Guyane, Kourou, France

⁴ Harvard University, Department of Organismic and Evolutionary Biology, Harvard University Herbaria,
22 Divinity Avenue Cambridge, MA 02138, USA

⁵ University of Gothenburg, Department of Plant and Environmental Sciences, Göteborg, Sweden

⁶ Université des Antilles et de la Guyane, Unité Mixte de Recherche Ecologie des Forêts de Guyane, Kourou, France

(Received 27 March 2009; accepted 21 December 2009)

Keywords:

chlorophyll estimate /
model calibration /
homographic functions /
neotropical trees

Abstract

- Chlorophyll meters such as the SPAD-502 offer a simple, inexpensive and rapid method to estimate foliar chlorophyll content. However, values provided by SPAD-502 are unitless and require empirical calibrations between SPAD units and extracted chlorophyll values.
- Leaves of 13 tree species from the tropical rain forest in French Guiana were sampled to select the most appropriate calibration model among the often-used linear, polynomial and exponential models, in addition to a novel homographic model that has a natural asymptote.
- The homographic model best accurately predicted total chlorophyll content ($\mu\text{g cm}^{-2}$) from SPAD units ($R^2 = 0.89$). Interspecific differences in the homographic model parameters explain less than 7% of the variation in chlorophyll content in our data set.
- The utility of the general homographic model for a variety of research and management applications clearly outweighs the slight loss of model accuracy due to the abandon of the species' effect.

1. INTRODUCTION

Traditional extraction-based methods of measuring chlorophyll content in forest trees are lengthy and expensive. Non-destructive methods have therefore been developed and inexpensive optical chlorophyll meters, such as the SPAD-502 (Konica Minolta, Osaka, Japan), are frequently used (Hawkins et al., 2009; Marengo et al., 2009; Pinkard et al., 2006; Uddling et al., 2007). Although these meters are portable and suitable for field use, they remain underutilised in forest science, perhaps because SPAD units are difficult to interpret. Indeed, SPAD value depends not only on chlorophyll content but also on other aspects of leaf optics, which may be influenced by various environmental and biological factors (Manetas et al., 1998; Markwell et al., 1995). The establishment of reference curves relating SPAD-unit and total foliar chlorophyll un-

der controlled environmental conditions is therefore a high priority.

To our knowledge, apart from the recent work of Marengo et al. (2009), research evaluating the use of SPAD units to estimate chlorophyll content in trees has been limited to certain fruit trees (e.g. Schaper and Chacko, 1991) or ornamental species (Pinkard et al., 2006; Richardson et al., 2002; Uddling et al., 2007). More generally, the current literature suffers from at least two limitations. First, the SPAD meter is unable to measure very low transmittances. However, all studies to date have based the calibration model on either exponential or, more often, polynomial relationships with no saturation level at high chlorophyll content. To correct this, we propose a new calibration model based on a homographic function that has a natural asymptote. A second limitation is the use of different calibration models in different studies which complicates rigorous comparisons of calibrations among a large panel of species. Tropical trees present an opportune system to study

* Corresponding author: bruno.herault@ecofog.gf

Table I. List of the studied tree species and their systematic position. Mean values ($\pm 95\%$ confidence interval) of leaf mass-per-area (LMA, g m^{-2}) and leaf thickness (μm) for leaves sampled under 20% of full irradiance are displayed. The number of samples (n), the parameters of the homographic models (α and β) and the coefficient of determination (R^2) are indicated for the 13 studied species.

Species name	Family name	Leaf structural characteristics		n	Homographic model		
		LMA (g m^{-2})	Thickness (μm)		α	β	R^2
<i>Amanoa guianensis</i> J.B. Aublet	Euphorbiaceae	102 ± 7.9	313 ± 15	30	124.1 ± 13.7	158.4 ± 10.6	0.98
<i>Bagassa guianensis</i> J.B. Aublet	Moraceae	38.2 ± 1.8	165 ± 12	30	147.6 ± 39.4	170.0 ± 35.8	0.97
<i>Carapa procera</i> A.P. de Candolle	Meliaceae	81.8 ± 4.4	220 ± 5	30	147.6 ± 24.5	193.7 ± 22.9	0.98
<i>Cecropia obtusa</i> Trécul.	Cecropiaceae	56.4 ± 6.6	254 ± 30	29	49.6 ± 10.0	88.2 ± 8.8	0.92
<i>Eperua falcata</i> J.B. Aublet	Caesalpiniaceae	67.4 ± 4.7	158 ± 4	32	205.5 ± 38.5	299.2 ± 45.8	0.78
<i>Hymenaea courbaril</i> Linnaeus	Caesalpiniaceae	59.8 ± 6.4	115 ± 8	34	195.0 ± 77.2	201.0 ± 61.6	0.91
<i>Inga thibaudiana</i> D.C.	Mimosaceae	75.7 ± 9.3	167 ± 18	34	331.2 ± 108.5	274.1 ± 70.4	0.94
<i>Pouteria</i> sp. J.B. Aublet	Sapotaceae	119 ± 11	289 ± 22	28	133.0 ± 51.2	150.9 ± 29.1	0.92
<i>Protium opacum</i> Swart	Burseraceae	92.0 ± 7.6	228 ± 20	31	165.4 ± 70.4	197.6 ± 52.6	0.94
<i>Sextonia rubra</i> (Mez) van der Weff	Lauraceae	73.9 ± 7.4	259 ± 12	30	169.9 ± 20.2	198.6 ± 17.2	0.99
<i>Symphonia globulifera</i> Linnaeus f.	Clusiaceae	71.2 ± 7.4	314 ± 21	30	154.1 ± 35.8	197.6 ± 31.9	0.98
<i>Tachigali melinonii</i> (Harms) Barneby	Caesalpiniaceae	59.0 ± 5.6	121 ± 8	23	164.7 ± 33.2	158.3 ± 23.0	0.97
<i>Vouacapoua americana</i> J.B. Aublet	Caesalpiniaceae	49.2 ± 4.4	118 ± 7	30	69.2 ± 60.5	112.8 ± 57.9	0.91

the potential for differential SPAD-chlorophyll relationships among species because they include a broad gradient of leaf types with different anatomical and physiological properties (Coste et al., 2005; Rozendaal et al., 2006).

In this study, we apply a novel homographic function to a data set of SPAD and chlorophyll measures of 13 tropical tree species representing a three-fold gradient in specific leaf area, and chlorophyll content ranging from ≈ 0 to $\approx 150 \mu\text{g}/\text{cm}^2$ total chlorophyll (Tab. I). In particular, we address the following questions: (1) is a homographic model appropriate for calibration of relationships between SPAD unit and chlorophyll content? (2) How does a general model perform relative to multiple species-specific models?

2. MATERIAL AND METHODS

2.1. Plant material

Our study was conducted in a greenhouse at Kourou, French Guiana ($5^\circ 10' \text{ N}$, $52^\circ 40' \text{ W}$) between May and July 2005. Thirteen tropical rainforest tree species, with various light requirements, were selected to cover a broad range of leaf structural characteristics (Tab. I). Seeds or saplings from at least five parent trees per species were collected from March to July 2003. They were grown in the greenhouse from December 2003 in 30 L pots with a 1:2 (v/v) mixture of a brown ferrallitic clay soil and a white sand of podzolic origin. During July 2004, all pots received 40 g slow-release complete fertilizer (Multicote 4, 17/17/17 N/P/K). Seedlings were irrigated daily with drip irrigation in order to maintain soil at field capacity ($0.25 \text{ m}^3 \text{ m}^{-3}$).

Shading nets were used to establish treatments that represent the range of light conditions experienced by seedlings in the lowland forests of French Guiana (Baraloto et al., 2005). The three irradiance treatments corresponded to 5, 10 and 20% of full sun, as recorded during 3 d-long measurement campaigns with inter-calibrated quantum-sensors for photosynthetically active radiation (PAR CBE 80 Solems,

Palaiseau, France) compared to an external sensor. Ten replicates per species were grown under each irradiance level.

2.2. Sampling

For each studied species, approximately 30 leaf samples were selected to cover the range of SPAD units exhibited across the light gradient. A sample consisted of a 0.78 cm^2 leaf disk. Three measurements using a SPAD-502 (Konica-Minolta, Osaka, Japan) were taken from the leaf disk and disk fresh weight (mg) was recorded. Immediately after measurements, samples were immersed in DMSO (dimethylsulfoxide) and kept in the dark.

2.3. Chlorophyll extraction

We used the DMSO extraction technique of Hiscox and Israelstam (1979). Samples were incubated at 65°C until leaf disks were completely colourless and the DMSO had turned green. Absorbance of the DMSO-chlorophyll extractions was then measured at 647 nm and 664 nm, relative to a DMSO blank, using a spectrophotometer (Jenway 6305 UV/Vis Spectrophotometer, Jenway, Essex, UK). The spectrophotometer was previously calibrated using commercial solutions of spinach chlorophylls a and b (BioChemika 10865 and 25740, respectively). Calibration procedure provided the Equations (1) and (2).

$$\text{Chl}_a (\text{mg ml}^{-1}) = 22 \text{ DO}_{664} - 9.1 \text{ DO}_{647} \quad (1)$$

$$\text{Chl}_b (\text{mg ml}^{-1}) = 29.5 \text{ DO}_{647} - 10.2 \text{ DO}_{664} \quad (2)$$

with Chl_a and Chl_b , the total content of chlorophyll a and b respectively; DO_{664} and DO_{647} , the optical densities read for wavelengths of 664 nm and 647 nm respectively.

2.4. Data analysis

We first calibrate an empirical relationship between SPAD unit and chlorophyll content for all 13 species together. Throughout the literature, three models have been used: linear, polynomial, or exponential. We developed a new homographic model (Eq. (3)) and tested which one out of these four models fit our data best.

$$Chl_i = (\alpha SPAD_i / (\beta - SPAD_i)) + \epsilon_i \quad (3)$$

where Chl is the total content in chlorophyll (a and b) of the sample i , $SPAD$ is the unitless reading from the SPAD-502 meter, α and β the fitted model parameters, and ϵ the model residuals.

The four competitive models (linear, polynomial, exponential or homographic) were compared for:

- (1) their accuracy (i.e. their ability to minimize the residual variance) by using the Akaike Information Criterion (AIC);
- (2) their robustness (i.e. their ability to predict an unseen data set) using a cross-validation technique.

We randomly split the data set into training and testing data sets of equal sample size. Each model was calibrated on the training data set and the values of the testing data set were then predicted using this calibration. Robustness was estimated using the coefficient of determination (R^2) and root mean square error of prediction (RMSEP) between predicted and observed values of the testing set. The procedure was repeated 1 000 times.

The species' effect was investigated by comparing AIC values for the general and species-specific model parameterization, with the one having the lowest AIC being the best. A residual bootstrap procedure (Efron and Tibshirani, 1993) was used to estimate the variance of the fitted parameters for each model.

Finally, to test the performance of our newly-calibrated model against other published models calibrated on different species, we compared these competing models in two ways. First, we tested the ability of the published models to predict our data. Second, we tested the ability of our newly-calibrated model to predict some raw data. Model performance was assessed using the R^2 and the RMSEP.

3. RESULTS

The relationship between total chlorophyll content and SPAD units was curvilinear (Fig. 1). The linear model, displaying the highest AIC value, was therefore the least accurate (Tab. II). Among the other models, the homographic model minimized the residual variance (Tab. II), displayed the lowest root mean square error of prediction (RMSEP = 201) and the higher R^2 ($R^2 = 0.89$). Among the 11 different models from the literature applied to our data set, the new homographic model best predicts our data (Tab. III). The RMSEP were generally lowest for models calibrated on tropical trees (*Mangifera indica*, *Eucalyptus* spp.). The performance of our newly-calibrated general homographic model to predict independent data sets gathered on other species is similar to the original models developed specifically for these species (Appendix¹ 1).

¹ Appendices are available on line only at www.afs-journal.org.

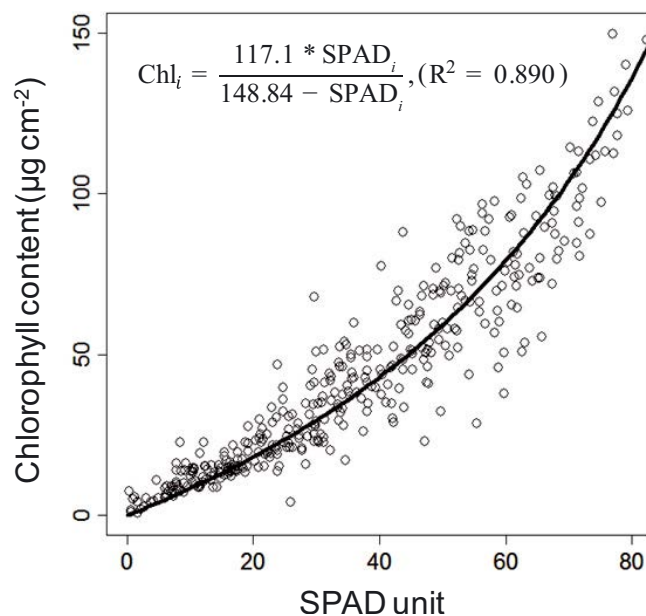


Figure 1. Relationships between leaf chlorophyll content and SPAD units for 13 neotropical trees. Homographic model was parameterized on the whole data set ($n = 391$ samples from 13 neotropical tree species). Equation of the homographic model and coefficient of determination (R^2) are displayed.

When the effect of species was taken into account, the AIC of the relationship between total chlorophyll contents and SPAD units decreased from 3 029 to 2 661. Among the 13 studied neotropical tree species, α ranged from 50 to 331 and β from 88 to 299 (Tab. I). All species-specific models had coefficients of determination (R^2) above 0.9, except *Eperua falcata* ($R^2 = 0.78$). Whereas the general homographic model obtained an R^2 value of 0.89 (Fig. 1), adding a species-effect increased the R^2 value to 0.96.

4. DISCUSSION

Linear regressions resulted in a higher Akaike Information Criterion highlighting lower predictability in the high SPAD range as has been shown by Uddling et al. (2007). Indeed, at least 11 out of 13 tropical trees (omitting *Inga thibaudiana* and *Eperua falcata*) had a pronounced non-linear SPAD-chlorophyll relationship (Appendix¹ 2). Several methods have been used to model such a curvilinear behaviour. In the literature, most authors employ second-order polynomials (Monje and Bugbee, 1992; Richardson et al., 2002) or exponential equations (Uddling et al., 2007). We introduced a new model that provided more readily interpretable parameters. Moreover, this new model best fit our data, was the most robust (Tab. II) and we were able to properly estimate the asymptotic value (β) above which the SPAD-502 does not work because of a high light absorption level. Our choice of a homographic model with no intercept was first based on a close examination of our data and on its fewer parameter number. To our knowledge, most published models have used an intercept (see

Table II. Comparison of the accuracy and the robustness of 4 competitive general models developed to predict chlorophyll content (Chl) with a SPAD-502. Accuracy was estimated using the Akaike Information Criterion (AIC). Robustness was estimated using both the coefficient of determination R^2 and the Root Mean Square Error of Prediction (RMSEP) throughout a cross-validation procedure.

Model	α	β	AIC	Min R^2	Mean R^2	Max R^2	Min RMSEP	Mean RMSEP	Max RMSEP
Linear									
$Chl_i = \alpha * SPAD_i$	1.310 ± 0.009	—	3215.9	0.8057	0.8454	0.8868	157.7	207.2	251.3
Polynomial									
$Chl_i = \alpha * SPAD_i + \beta * SPAD_i^2$	0.664 ± 0.040	0.012 ± 0.001	3075.3	0.8485	0.8794	0.9174	141.4	175.0	208.8
Exponential									
$Chl_i = \alpha(e^{\beta * SPAD_i} - 1)$	0.018 ± 0.001	42.85 ± 4.48	3049.2	0.8554	0.8853	0.9201	139.1	170.7	204.3
Homographic									
$Chl_i = (\alpha * SPAD_i / (\beta - SPAD_i))$	117.10 ± 6.53	148.84 ± 8.69	3029.7	0.8598	0.8899	0.9211	137.3	164.9	201.4

Table III. Comparison of the ability of 11 published models, linking total chlorophyll content (units are given for each model; Y) to Spad units (X), to predict our 13 neotropical trees data set. Relative performance of models was assessed using Root Mean Square Error of Prediction (RMSEP).

Species	Reference	Model	Unit	RMSEP
<i>Betula papyrifera</i> Marsh.	Richardson et al. (2002)	$Y = 5,52E - 04 + 4.04E - 04 * X + 1.25E - 05 * X^2$	mg cm ⁻²	14.67
<i>Betula pendula</i> Roth	Uddling et al. (2007)	$Y = 0.0641 * \exp(0.0467X)$	g m ⁻²	35.93
<i>Eucalyptus globules</i> Labill.	Pinkard et al. (2006)	$Y = \exp(-8.79 + 2.08 * \ln(X))$	μg mm ⁻²	15.38
<i>Eucalyptus nitens</i> (H. Deane & Maiden) Maiden	Pinkard et al. (2006)	$Y = \exp(-8.55 + 0.97 * \ln(X))$	μg mm ⁻²	17.74
<i>Mangifera indica</i> L.	Schaper and Chacko (1991)	$Y = -5.4 + 11.1 * X$	mg m ⁻²	17.08
<i>Malus domestica</i> Borkh	Campbell et al. (1990)	$Y = -83 + 2.37 * X$	μg cm ⁻²	47.86
<i>Oriza sativa</i> L.	Monje and Bugbee (1992)	$Y = 1.034 + 0.308 * X + 0.11 * X^2$	mg m ⁻²	30.31
<i>Solanum tuberosum</i> L.	Uddling et al. (2007)	$Y = 0.0913 * \exp(0.0415 * X)$	g m ⁻²	33.60
<i>Triticum aestivum</i> L.	Uddling et al. (2007)	$Y = 0.0599 * \exp(0.0493 * X)$	g m ⁻²	46.18
<i>Vigna unguiculata</i> (L.) Walp.	Murillo-Amador et al. (2004)	$Y = -2.79 + 0.88 * X$	μg cm ⁻²	24.98
6 neotropical trees	Marenco et al. (2009)	$Y = 0.53 * \exp(0.0364 * X)$	mg m ⁻²	24.21
13 neotropical trees	Present study	$Y = 117.1 * X / (148.84 - X)$	μg cm ⁻²	11.59

Tab. III). Markwell et al. (1995) justified such an intercept by citing a stronger absorbance of chlorophyll b than of chlorophyll a at 650 nm, but this statement was not supported by recent work by Uddling et al. (2007).

The addition of a species parameter to the model reduced the residual variance (from 11 to 4%) because most of the species-specific models diverged substantially at high SPAD units (see Appendix 2¹). The differences observed between species may be related to differences in the distribution of chlorophyll in tree leaves as a result of the structural organization of chlorophyll molecules in chloroplasts, chloroplasts in cells, and cells in leaves (Fukshansky et al., 1993). It is generally accepted that with increasing heterogeneity of chlorophyll distribution inside a leaf, the absorption of a given amount of chlorophyll decreases (Uddling et al., 2007) and the relationship between SPAD and chlorophyll content deviates more from linearity. Another potential source of heterogeneity in the chlorophyll distribution may be the veins and veinules network (McClendon and Fukshansky, 1990). Even though we were careful to avoid veins when measuring SPAD, it remains possible that some measurements were biased because they were made along veins, reinforcing the curvilinear shape of the calibration curve. Alternatively, it has been sug-

gested that differences in LMA (Leaf Mass per Area), among or within species, may lead to different calibration curves (Thompson et al., 1996) through different degrees of mutual shading among chloroplasts. Assuming that higher LMA is related to a greater leaf thickness, mutual chloroplast shading may be avoided in thicker leaves (Hikosaka, 2004) even if our data suggest no effect of LMA *per se* (we found no relationship between alpha-beta and LMA, results not shown). All in all, interspecific differences in the homographic model were clear but explain a relatively small proportion of variation in chlorophyll content in our data set (less than 7%).

Despite the increasing frequency with which leaf functional trait studies are employing optical meters to estimate chlorophyll content, few studies have reported results on the relative statistical performance of different models to describe the SPAD-chlorophyll relationships. Tropical tree species appear to represent the broadest range of SPAD and chlorophyll values reported in the literature, and our results demonstrated that foliar chlorophyll of tropical trees could be reliably estimated using the SPAD-502 meter. The utility of the general homographic model (Fig. 1) clearly outweighs limitations due to the loss of model accuracy. The SPAD-502 meter provides a simple, non-destructive method for estimating foliar chlorophyll

that quickly reports a large number of readings, thus paving the road for immediate assessment of physiological variables (Hawkins et al., 2009). In this way, it should be possible to use the SPAD-502 as a tool for a variety of research and management applications, including the assessment of physiological changes over time, the assessment of relative health status or to delineate the effects of management and logging practices on the photosynthetic performance.

Acknowledgements: The authors are indebted to Pascal Imbert, Jean-Yves Goret, Saintano Dufort, Marcel Blaize, Jean-Louis De Kerpeyron and Henry Grootfaam (UMR Ecofog, Kourou) for their help throughout the experiment. E. Pinkard graciously provided the raw data from her published study.

REFERENCES

- Baraloto C., Goldberg D.E., and Bonal D., 2005. Performance trade-offs among tropical tree seedlings in contrasting microhabitats. *Ecol.* 86: 2461–2472.
- Campbell R.J., Mobley K.N., Marini R.P., and Pfeiffer D.G., 1990. Growing conditions alter the relationship between SPAD-502 values and apple leaf chlorophyll. *Hortsci.* 25: 330–331.
- Coste S., Roggy J.-C., Imbert P., Born C., Bonal D., and Dreyer E., 2005. Leaf photosynthetic traits of 14 tropical rain forest species in relation to leaf nitrogen concentration and shade tolerance. *Tree Physiol.* 25: 1127–1137.
- Efron B. and Tibshirani R.J., 1993. An introduction to the bootstrap, Chapman-Hall, London, 456 p.
- Fukshansky, Martinez A., Remisowsky V., McClendon J., Ritterbusch A., Richter T., and Mohr H., 1993. Absorption spectra of leaves corrected for scattering and distributional error: a radiative transfer and absorption statistics treatment. *Photochem. Photobiol.* 57: 538–555.
- Hawkins T.S., Gardiner E.S., and Comer G.S., 2009. Modeling the relationship between extractable chlorophyll and SPAD-502 readings for endangered plant species research. *J. Nat. Conserv.* 17: 123–127.
- Hikosaka K., 2004. Interspecific difference in the photosynthesis–nitrogen relationship: patterns, physiological causes, and ecological importance. *J. Plant Res.* 117: 481–494.
- Hiscox J. and Israelstam G., 1979. A method for the extraction of chlorophyll from leaf tissue without maceration. *Can. J. Bot.* 57: 1332–1334.
- Manetas Y., Grammatikopoulos G., and Kypris A., 1998. The use of the portable, non-destructive, SPAD-502 (Minolta) chlorophyll meter with leaves of varying trichome density and anthocyanin content. *J. Plant Physiol.* 153: 513–516.
- Marenco R.A., Antezana-Vera S.A., and Nascimento H.C.S., 2009. Relationship between specific leaf area, leaf thickness, leaf water content and SPAD-502 readings in six Amazonian tree species. *Photosynth.* 47: 184–190.
- Markwell J., Osterman J., and Mitchell J., 1995. Calibration of the Minolta SPAD-502 leaf chlorophyll meter. *Photosynth. Res.* 46: 467–472.
- McClendon J.H. and Fukshansky L., 1990. On the interpretation of absorption spectra of leaves I. The introduction and the correction of leaf spectra for surface reflection. *Photochem. Photobiol.* 51: 203–210.
- Monje O.A. and Bugbee B., 1992. Inherent limitations of nondestructive chlorophyll meters - a comparison of 2 types of meters. *Hortscience* 27: 69–71.
- Murillo-Amador B., Avila-Serrano N.Y., Garcia-Hernandez J.L., Lopez-Aguilar R., Troyo-Dieguez E., and Kaya C., 2004. Relationship between a nondestructive and an extraction method for measuring chlorophyll contents in cowpea leaves. *J. Plant Nutr. Soil Sc.* 167: 363–364.
- Pinkard E.A., Patel V., and Mohammed C., 2006. Chlorophyll and nitrogen determination for plantation-grown *Eucalyptus nitens* and *E. globulus* using a non-destructive meter. *For. Ecol. Manage.* 223: 211–217.
- Richardson A.D., Duigan S.P., and Berlyn G.P., 2002. An evaluation of noninvasive methods to estimate foliar chlorophyll content. *New Phytol.* 153: 185–194.
- Rozendaal D.M.A., Hurtado V.H., and Poorter L., 2006. Plasticity in leaf traits of 38 tropical tree species in response to light; relationships with light demand and adult stature. *Funct. Ecol.* 20: 207–216.
- Schaper H. and Chacko E.K., 1991. Relation between extractable chlorophyll and portable chlorophyll meter readings in leaves of eight tropical and subtropical fruit tree species. *J. Plant Physiol.* 138: 674–677.
- Thompson J.A., Schweitzer L.E., and Nelson R.L., 1996. Association of specific leaf weight, an estimate of chlorophyll, and chlorophyll concentration with apparent photosynthesis in soybean. *Photosynth. Res.* 49: 1–10.
- Uddling J., Gelang-Alfredsson J., Piikki K., and Pleijel H., 2007. Evaluating the relationship between leaf chlorophyll concentration and SPAD-502 chlorophyll meter readings. *Photosynth. Res.* 91: 37–46.

APPENDIX L

Integrating functional diversity into tropical forest plantation designs to study ecosystem processes

Baraloto, C., E. Marcon, F. Morneau, S. Pavoine et J.-C. Roggy (2010). « Integrating functional diversity into tropical forest plantation designs to study ecosystem processes ». In : *Annals of Forest Science* 67, p. 303.

Integrating functional diversity into tropical forest plantation designs to study ecosystem processes

Christopher BARALOTO^{1*}, Eric MARCON², François MORNEAU², Sandrine PAVOINE³,
Jean-Christophe ROGGY¹

¹ INRA, UMR “Ecologie des Forêts de Guyane”, Kourou, French Guiana

² AgroParisTech, UMR “Ecologie des Forêts de Guyane”, Kourou, French Guiana

³ UMR “Conservation des espèces, restauration et suivi des populations” Muséum National d’Histoire Naturelle, Paris, France

(Received 16 May 2009; accepted 15 September 2009)

Keywords:

complementarity /
ecosystem function /
functional groups /
leaf economics spectrum /
nitrogen fixation /
quadratic entropy

Abstract

- The elucidation of relationships between biodiversity and ecosystem processes has been limited by the definition of metrics of biodiversity and their integration into experimental design. Functional trait screening can strengthen the performance of these designs.
- We suggest the use of Rao’s quadratic entropy to measure both functional diversity and phylogenetic diversity of species mixtures proposed for an experimental design, and demonstrate how they can provide complementary information.
- We also present an index assessing the statistical performance of these independent variables in different experimental designs. Measurement of independent variables as continuous vs. discrete variables reduces statistical performance, but improves the model by quantifying species differences masked by group assignments.
- To illustrate these advances, we present an example from a tropical forest tree community in which we screened 38 species for nine functional traits. The proposed TropiDEP design is based on the relative orthogonality of two multivariate trait axes defined using principal component analysis.
- We propose that independent variables describing functional diversity might be grouped to calculate independent variables describing suites of different traits with potentially different effects on particular ecosystem processes. In other systems these axes may differ from those reported here, yet the methods of analysis integrating functional and phylogenetic diversity into experimental design could be universal.

Mots-clés :

complémentarité /
fonctions de l’écosystème /
groupes fonctionnels /
schéma universel de fonctionnement
foliaire des végétaux /
fixation biologique du N /
entropie quadratique

Résumé – Diversité fonctionnelle et processus écosystémiques dans des assemblages synthétiques d’espèces d’arbres de forêt tropicale.

- La compréhension des relations pouvant exister entre biodiversité et fonctionnement des écosystèmes a été longtemps limitée par la définition de méthodes de quantification de la diversité biologique et la mise en œuvre de dispositifs expérimentaux permettant sa mesure. L’identification de syndromes de traits fonctionnels clefs influençant des fonctions écosystémiques particulières peut renforcer la performance de ces dispositifs.
- Nous suggérons l’utilisation de l’entropie quadratique de Rao pour mesurer la diversité fonctionnelle et phylogénétique dans des assemblages synthétiques d’espèces, et montrons comment ces mesures de diversité sont complémentaires.
- Nous présentons également un indice permettant de tester la performance statistique de ces variables indépendantes dans différents modèles expérimentaux. L’utilisation de variables indépendantes continues plutôt que discrètes réduit la performance statistique mais améliore le modèle en quantifiant les différences fonctionnelles entre espèces ; différences généralement masquées lors de leur assignation en groupes fonctionnels.

* Corresponding author: Institut National de la Recherche Agronomique, BP 709, 97387 Kourou, France, chris.baraloto@ecofog.gf

- Pour illustrer ces avancées, nous présentons un exemple d'assemblages synthétiques à partir de 38 espèces d'arbres de forêt tropicale sélectionnées pour 9 traits fonctionnels (TropiDEP). Le plan d'expérience de TropiDEP est basé sur l'orthogonalité relative de deux axes multivariés de traits fonctionnels définis par analyse en composantes principales.
- Nous proposons que les variables décrivant la diversité fonctionnelle soient groupées pour calculer des variables indépendantes, divisées en plusieurs axes décrivant des combinaisons de différents traits pouvant influencer des processus différents de l'écosystème (e.g. processus du N et du C). Dans d'autres systèmes, ces axes peuvent différer de ceux présentés ici, mais les méthodes d'analyse peuvent être universelles.

1. INTRODUCTION

The rapid and pervasive loss of biodiversity over the past century has provoked debate about the consequences of species loss for ecosystem function and the stability of biogeochemical cycles (e.g., Schwartz et al., 2000). Understanding the relationship between biodiversity and ecosystem processes (B-EP) has thus become one of the critical issues in contemporary ecology (Chapin et al., 2000; Loreau et al., 2001). Although some experiments have found links between biodiversity and ecosystem parameters such as primary productivity or nitrogen retention, other experiments have not found these effects (Hooper et al., 2005). These conflicting results may be due in part to the means by which the B-EP relationship is investigated and analyzed (Diaz and Cabido, 2001; Hooper et al., 2005; Huston et al., 2000; Loreau and Hector, 2001; Wright et al., 2006). Such discussions have moved the central question from whether biodiversity affects ecosystem processes, to what mechanisms underly these relationships and how do they differ between systems (Cardinale et al., 2007; Ewel, 2006; Fargione et al., 2007; Gamfeldt et al., 2008; Hector and Bagchi, 2007; Hillebrand et al., 2008; Hooper and Dukes, 2004; Isbell et al., 2009; Naeem and Wright, 2003; Polley et al., 2007; Reich et al., 2004; Roscher et al., 2004; Zhang and Zhang, 2007).

In this paper, we suggest improvements for studying the biodiversity-ecosystem process relationship (hereafter B-EP) using experiments manipulating species and functional diversity. We focus on forest ecosystems, using a tropical tree community as a model system for several reasons. First, tropical forests are among the most diverse plant communities described and thus offer a large pool of species, including hundreds of nitrogen-fixing legumes that differ widely in functional traits related to carbon and water cycling (Bonal et al., 2000; Roggy et al., 1999). Second, tropical forests play a major role in global biogeochemical cycles; they may account for more than a third of global net primary productivity (e.g., Phillips et al., 1998). Finally, tropical forests are undergoing rapid conversion to deforested areas for livestock, agriculture and mining, and recent attention has focused on how to rehabilitate converted tropical forests (Lamb et al., 2005; Parrotta and Knowles, 1999). Yet, only a handful of experimental plantations manipulating mixtures of forest trees exist, all of which incorporate some limitations for studying B-EP related to the choice and number of species tested and the design of experimental plots (Ewel, 2006; Scherer-Lorenzen et al., 2005).

In particular, we address two general limitations of current B-EP research that are particularly lacking in the experimental approach used in extant forest plantation studies. First, we focus on how biodiversity is defined and measured in designing and evaluating B-EP experiments (Petchey and Gaston 2006; Wright et al. 2006). Second, we examine the statistical performance of different experimental designs and propose a compromise between practical implementation and statistical rigor using a new performance measure. We illustrate our proposed improvements using as a case study TropiDEP, an experimental design based on a matrix of traits and phylogenetic information assembled for French Guianan tree species.

2. DEFINING AND MEASURING BIODIVERSITY

It is now well accepted that the general relationships between species number and ecosystem processes such as productivity are the result of functional differences among species. Accordingly, the use of species richness as a proxy for functional diversity has been criticized as too coarse a measure for predicting ecosystem parameters (Petchey et al., 2004; Petchey and Gaston, 2006; Roscher et al., 2004). Still, the best means to estimate functional diversity for both the design and analysis of B-EP experiments remains contentious (Petchey and Gaston, 2006; Ricotta, 2005; Wright et al., 2006). Although analyses can be performed using post-hoc attributions of species to groups or post-hoc measures of diversity, designs generally incorporate some a priori designation to maintain balance and to avoid problems of circularity (Wright et al., 2006).

In general, two approaches have been used to estimate *a priori* functional diversity in B-EP experiments. Most studies have used a broad designation of species groups based on key traits such as growth form, photosynthetic pathway, or N-fixing capacity (Ewel, 2006; Hooper and Dukes, 2004; Reich et al., 2004; Tilman et al., 2001). Broad designations are often easy to employ because they rely on “soft” traits that are readily distinguished for most species and are easily scored as categorical variables (Hooper et al., 2005). Broad designations also permit the identification of types of species that have particular effects on ecosystem processes or that may complement species from other groups (e.g., Ewel, 2006; Hooper and Dukes, 2004; Reich et al., 2004). However, broad designations mask within-group trait variability, and such fine-scale differences may also be of consequence for ecosystem processes (Craine et al., 2002).

A fine-scale approach relies on quantitative differences between species in the values of particular functional traits that are hypothesized to affect ecosystem processes (Mason et al., 2003; Ricotta, 2005). For example, Petchey and Gaston (2006) proposed a quantitative measure of functional diversity (FD) analogous to a similar measure of phylogenetic diversity (PD; Faith, 1992) that is based on the branch length of the functional dendrogram of species clustered in trait space. Their index appears to be a better predictor of aboveground productivity than species richness or other measures of functional distance (Petchey et al., 2004).

Biodiversity can be represented not only by species and functional diversity but also by the diversity in evolutionary relationships among taxa (Faith, 1992; Forest et al., 2006). If traits are conserved within lineages, then functional diversity should be tightly correlated with phylogenetic diversity; for example, among vascular plants nitrogen-fixation occurs almost exclusively within legumes (Wojciechowski et al., 2005). However, the extent to which other functional traits are evolutionarily conserved remains under debate. Wood density appears to be highly conserved within higher taxonomic units (Chave et al., 2006), whereas seed mass shows strong divergence within genera (Moles et al., 2005).

A third problem encountered by traditional designs is that even continuous measures of functional diversity such as FD do not account for differences in the relative abundances of species in experimental plots. A recently popularized measure of diversity, the quadratic entropy (Rao, 1982), takes into account both species abundances and pairwise distances among species (Botta-Dukát, 2005; Pavoine et al., 2005). When all species are considered functionally equivalent, the index is equivalent to the Gini-Simpson index (Pavoine et al., 2005). Botta-Dukát (2005) showed that in addition to accounting for abundance and integrating multiple traits, the Rao index satisfies a priori criteria proposed by Mason et al. (2003). The Rao index is thus suitable to contrast diversity in terms of species richness, functional diversity and phylogenetic diversity.

3. DESIGNING EXPERIMENTS ADDRESSING BIODIVERSITY AND ECOSYSTEM PROCESSES

The classical statistical model used in the literature to assess the relationship between diversity and ecosystem processes is a linear regression (Loreau and Hector, 2001). By constructing a sample design of multiple plots, one controls the values of a set of variables such as species richness in each plot. These values are contained in a vector \mathbf{x} , where $\mathbf{x}' = (x_1, \dots, x_j, \dots, x_p)$. After a period of time, a response y is measured in each plot, for example net primary productivity (NPP), and this response is assumed to be given by $y = \mathbf{x}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\theta}$ is the vector of parameters to estimate and $\boldsymbol{\varepsilon}$ is the error vector of variance σ^2 . For improved estimates, several (n) plots must be investigated, each of which is submitted to a defined sample design. Together, these plots can be described by a matrix \mathbf{X} whose rows give the n vectors corresponding to the values fixed for the p variables in the n plots and a vector \mathbf{y} of n observed responses. Consequently, the model can be written as

a matrix formula as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}. \quad (1)$$

The precision of the estimation of $\boldsymbol{\theta}$ is given by the matrix of variance/covariance $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$. One can not control for the model error σ^2 , due to the unobserved factors, but the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ depends solely on the design. We denote V_{jk} its terms. The variance of the estimator $\hat{\alpha}_j$ is $\sigma_j^2 = V_{jj}\sigma^2$. The optimal design is characterized by smaller parameter estimation variance and smaller correlations between parameter estimations. By definition, the simplest factorial experimental design, in which only extreme values of the factors are retained and combined in all possible ways, is statistically optimal (Cochran and Cox, 1992).

However, an experimental design must also be able to confirm the assumed linearity of the statistical model. Linearity of ecosystem processes such as NPP has been shown only for species richness (expressed as a logarithm). Tilman et al. (2001) estimate that the relation is probably asymptotic for functional diversity; it would be linear for low diversity only. Intermediate values of the factors that are not included in the basic factorial design are necessary to describe nonlinear relationships. An alternative is a complete design, which includes all possible combinations of factor levels.

Two other considerations are important to experimental design. First, because functional diversity and the number of species are partly correlated, some designs (e.g., high diversity with a single species) are impossible. Second, the effects of particular species may dominate the response of all plots where they are present (Loreau and Hector, 2001). Replication with different species is therefore necessary to avoid biases in estimations. For each level of specific and functional diversity, several plots with alternative species have been proposed (e.g., Roscher et al., 2004). For biodiversity experiments, a potential solution is to replicate the complete design with different species combinations. However, when the species pool is large, as is the case in diverse systems, it becomes impractical to test all possible combinations. Consequently, a strategy for choosing species combinations is necessary.

The different constraints on experimental design are often contradictory, so compromises must be reached. Adding intermediate values of factors (e.g., species number) can help to detect non-linearity, but it also decreases factors' variance and consequently estimation precision. To allow a rational choice, we define a performance measure for experimental designs investigating diversity effects on ecosystem processes.

Assuming normality and denoting t the Student variable, we know the confidence interval for each parameter: $CI(\alpha_j) = \left[\hat{\alpha}_j - t \frac{\sigma_j}{\sqrt{n}}; \hat{\alpha}_j + t \frac{\sigma_j}{\sqrt{n}} \right]$. We rearrange it so that the confidence interval of the estimation of the parameter j is the product of four independent terms: σ , the standard deviation of the model's error term; n , the number of experimental plots; S_j , a scale factor reflecting the units chosen; and P_j , the performance of the design. We can write:

$$CI(\alpha_j) = \left[\hat{\alpha}_j - t \frac{\sigma}{\sqrt{n}S_jP_j}; \hat{\alpha}_j + t \frac{\sigma}{\sqrt{n}S_jP_j} \right].$$

The number n of plots can be considered as an economic variable. Doubling the design divides the estimators' confidence interval by $\sqrt{2}$. After choosing the minimum (unit) design, repeating it is a matter of finance.

$S_j = \sqrt{(x_j^{\max} - x_j^{\min})^2 / 2}$, where x_j^{\max} and x_j^{\min} are the extreme values of \mathbf{X}_j . It is a scale factor which only ensures the homogeneity of the equation.

$P_j = 1 / (S_j \sqrt{V_{jj}})$ is the performance of the design for variable j ¹. It is 1 for the factorial design (the one containing extreme values only), and less than 1 for other designs. It can be easily computed for each potential design to evaluate its relative efficiency. For example, $P_j = 50\%$ means that everything else equal, the confidence interval will be doubled, or that twice more plots will be necessary to achieve the same precision in the estimation of the parameter as in the factorial design. This metric thus allows a comparison of the same model with different values of exogenous variables and/or a different number of replications.

The other point of interest is the correlation between parameters j and k , given directly by $\frac{V_{jk}}{\sqrt{V_{jj}V_{kk}}}$. Performances for different variables and correlations may vary in opposite directions when the design is changed. Yet, we have the necessary information to evaluate the ratio of performance to infrastructure cost. Practically, the factorial design is taken as a reference since its cost is the lowest; variances are as low as possible (all P_j equal 1 by construction), and covariances are null. The real designs, which face other constraints, have higher costs and correlations.

4. A CASE STUDY EXAMPLE: TROPIDEP

As a case study, we present TropiDEP, an experimental design for B-EP in a tropical forest ecosystem, that differs in two key ways from other B-EP designs reported to date. It incorporates the multiple axis approach to functional distances, and it can be modified to strengthen its statistical power using discrete or continuous independent variables, based on the statistical performance metrics described above. In this way, it represents a compromise to multiple replications of the complete design.

4.1. Functional trait measurement

We measured a series of traits (see Tab. 1) for a set of 38 focal species that are common in lowland tropical forests in French Guiana and that represent the most abundant tree families in the Guiana Shield. We made a particular effort to include legume species of the subfamily Mimosoideae that are known to maintain associations with nitrogen fixing *Rhizobium* symbionts (Roggy et al., 1999).

¹ The form of P_j is appropriate because in the case of the factorial design, S_j is the variance of \mathbf{X}_j and $V_{jj} = 1/S_j^2$. P_j is actually a normalized ratio of variances of estimators.

Table 1. Functional traits measured for a regional species pool of functional diversity. All foliar traits have been standardized to a leaf mass basis, and were measured on juveniles of two years age under controlled conditions in shadehouses.

Attribute (Abbreviation)	Unit	Measurement
Foliar [C]:[N] (C_m - N_m)	g g ⁻¹	CHN autoanalyzer
Foliar [N] (N_m)	μg g ⁻¹	CHN autoanalyzer
Foliar delta ¹⁵ N	μg μg ⁻¹	Mass spec. analysis (Roggy et al., 1999)
Assimilation Rate (A_m)	μmol CO ₂ g ⁻¹ s ⁻¹	CIRAS-1 System at 360 ppm CO ₂ and 700 μmol m ⁻² s ⁻¹ PAR
Stomatal conductance (G_m)	mmol H ₂ O g ⁻¹ s ⁻¹	CIRAS-1 System at 360 ppm CO ₂ and 700 μmol m ⁻² s ⁻¹ PAR
Relative Growth Rate (RGR)	mg g ⁻¹ d ⁻¹	For a harvest period from 24–30 months age
Root nodules (Nodules)	Presence-absence	On roots at final harvest
Root-Shoot Ratio (R-S)	g g ⁻¹	root biomass/shoot biomass at final harvest
Specific Leaf Area (SLA)	cm ² g ⁻¹	leaf area/leaf biomass for new leaves at final harvest

Traits were measured on at least eight juvenile plants per species of two years age (45–203 cm tall with basal diameter of 5.4–21.3 mm) grown from seed collected from at least three parent trees. The juvenile stage was chosen for two reasons. First, to control for known environmental effects on functional traits (e.g., Bonal et al., 2000), we chose to measure traits under controlled conditions in a shadehouse, which limited the study to juveniles. Second, saplings represent a size at which individuals begin to interact in plantation settings (Scherer-Lorenzen et al., 2005) and are thus relevant to initial measurements of ecosystem processes, such as nitrogen retention and biomass accumulation. Nevertheless, trait values change with ontogeny especially in trees (e.g., Roggy et al., 1999), and subsequent analyses of experimental designs should update trait measurements in concert with measures of ecosystem processes.

All plants were grown in individual 7 L pots containing a 2:1 mixture of forest loam and sand soil, placed in a shadehouse in Kourou, French Guiana. Light availability was approximately 20% of full sun photosynthetically active radiation (about 300 μmol m⁻² s⁻¹, on a cloudless day, with daily integrated level of about 5 mol m⁻² d⁻¹). Leaf traits were chosen to represent the primary axis of foliar trait variation described in the literature (Cornelissen et al., 2003; Wright et al., 2004). In addition, we described nitrogen nutrition for each species using two variables. First, we analyzed ¹⁵N isotope concentrations for leaf tissue collected in the absence of any fertilization. The delta ¹⁵N ratio can be used to discriminate between the different N sources used by the species (NH₄⁺, NO₃⁻ and N₂) (Roggy et al., 1999). Second, we scored

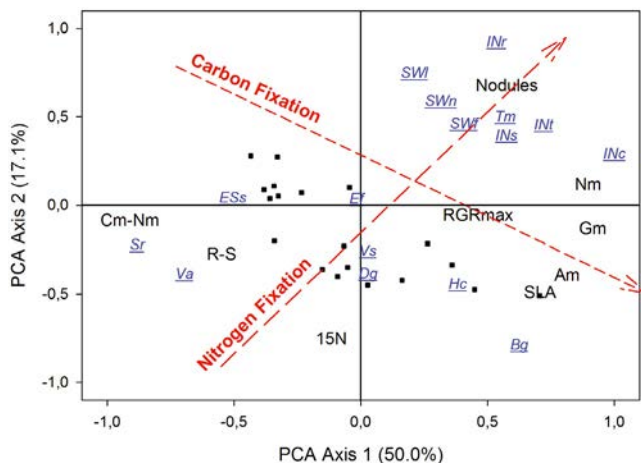


Figure 1. Results of a principal components analysis for correlations among the nine functional traits highlighted in Table I, for the 38 species regional pool in French Guiana. Two axes were defined that explain 67.1% of variation. Along these factors, four putative functional groups can be defined representing combinations of rates of carbon and nitrogen fixation. Abbreviations for variables are explained in Table I. Species positions in trait space are indicated with squares, except for the 16 focal species focal (four within each functional group), to be manipulated in the proposed TropiDEP experimental design, for which species codes are given (see Tab. II). Note that the nitrogen-fixing legumes do not cover the entire gradient of variation along the carbon diversity axis comprised by the non-fixing species, so for TropiDEP we chose the extreme species with very thick leaves and relatively low photosynthetic capacity within the non-N-fixing slow-growers.

the presence of *Rhizobium*-containing root nodules on plant with negative values of delta ^{15}N ratio in order to distinguish N_2 -fixing legumes from species using NO_3^- (Schimann et al., 2008); when nodules were present, they were abundant and occurred on all individuals. Table I summarizes the methods and units of measure for the nine traits.

4.2. Functional trait correlations and focal species selection

We used a principal component analysis (PCA) to examine correlations among the measured traits, to define interpretable multivariate trait axes, and to project species differences in multivariate trait space. All analyses were conducted in the ade4 module of the R statistical package (Chessel et al., 2004).

The PCA identified two principal axes explaining 67.1% of trait variation among the 38 study species (Fig. 1). A rotation of the first axis corresponds with what has been described as a global axis of leaf types among plants (Wright et al., 2004), with strong positive loadings of photosynthetic capacity, stomatal conductance for water vapor, and specific leaf area; and strong negative loadings for leaf carbon-nitrogen ratio. A rotation of the second axis segregates N-fixing legumes from other species, with strong positive loading for the nodulation vari-

able and a strong negative loading for the ^{15}N isotope values, confirming this pattern.

We also examined relationships among the species using hierarchical clustering algorithms with Ward's minimum variance method. We calculated Euclidean distances among species pairs derived from three dissimilarity matrices: a phylogenetic matrix based on the angiosperm supertree (Davies et al., 2004); and matrices of traits related to carbon fixation or nitrogen uptake (Tab. I).

From the 38 study species, we chose 16 species representing four broad "functional groups" that combine the carbon and nitrogen axes (Tab. II; projected in Fig. 1) for the TropiDEP design. Within our species pool, the nitrogen-fixing legumes do not cover the entire gradient of variation along the carbon diversity axis comprised by the non-fixing species (Fig. 1). We chose to retain the latter diversity by selecting the extreme species with very thick leaves and relatively low photosynthetic capacity within the non-N-fixing slow-growers. As a result, some species are actually closer in trait space to species assigned to a different group.

The chosen focal species also illustrate how measures of distances between species can be correlated despite the clear separation of trait axes. The strong phylogenetic constraint on N-fixation within legumes results in a slight positive correlation between phylogenetic distance and functional distance along the nitrogen axis. This can be seen in dendrograms of the hierarchical cluster analysis performed on the 16 TropiDEP species (Fig. 2), with the lower cluster in Figure 2c also being clustered in Figure 2a. However, this correlation is weakened because not all legumes in the species pool are N-fixing. A majority of the N-fixing legumes have high foliar nitrogen contents, and tend to grow quickly and have rapid carbon assimilation rates, even if they have thicker leaves. As a result, clusters of species with high values along the carbon axis (Fig. 2b) also tend to be clustered on the nitrogen axis (Fig. 2c). With further trait screening of N-fixing legumes, we might be able to identify species with lower values along the carbon axis to improve the design presented here. However, we can still account for the variability in axis distances for our independent variables by calculating measures of diversity for each experimental plot we create.

4.3. The TropiDEP design

The TropiDEP design is based on the relative orthogonality of the two multivariate trait axes defined using principal component analysis (Fig. 1) and their potential effects on particular ecosystem processes (Tab. II). We hypothesize that the consequences of competition and facilitation for ecosystem processes will depend at least in part on independent resource-use complementarity along each of these axes, such that a global distance measure as proposed by Petchey and Gaston (2002) may mask relationships between functional diversity and ecosystem processes.

To separate these effects, we propose to combine species mixtures that independently include variation along each axis of functional diversity for each level of species richness. This

Table II. A classification for functional groups of French Guianan trees based on leaf morphology and physiology and nitrogen nutrition status. Also shown are the predicted species properties relevant to ecosystem processes of carbon and nitrogen cycling. The ordination of traits and example species are presented in Figure 1.

Traits	Light-demanding N-fixers	Light-demanding	Shade-tolerant N-fixers	Shade-tolerant
Leaf Structure	High SLA	High SLA	Low SLA	Low SLA
Leaf Allocation	High Leaf N	High Leaf N	High Leaf N	Low Leaf N
Growth Rate	Fast	Moderate	Moderate	Slow
Biomass Turnover	Rapid	Rapid	Moderate	Slow
Processes				
Litter Quality	Excellent	Fair	Good	Poor
Nitrogen Availability	High	Moderate	Moderate	Low
Rooting Depth	Shallow	Variable	Shallow	Variable
Species	<i>Inga cayennensis</i> (INc) <i>Inga stipularis</i> (INs) <i>Inga thibaudiana</i> (INT) <i>Tachigali melinonii</i> (Tm)	<i>Bagassa guianensis</i> (Bg) <i>Dicorynia guianensis</i> (Dg) <i>Hymenaea courbaril</i> (Hc) <i>Viola surinamensis</i> (Vs)	<i>Inga rubiginosa</i> (INr) <i>Swartzia grandifolia</i> (SWf) <i>Swartzia leblondii</i> (SWl) <i>Swartzia panacoco</i> (SWn)	<i>Eschweilera sagotiana</i> (ESs) <i>Vouacapoua americana</i> (Va) <i>Sextonia rubra</i> (Sr) <i>Eperua falcata</i> (Ef)

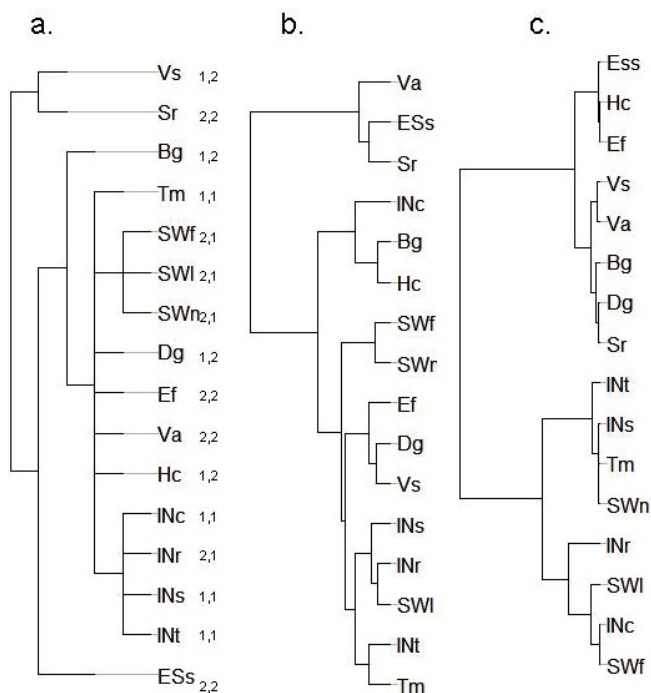


Figure 2. Dendrograms for the 16 focal species in the TropiDEP design based on (a) phylogenetic distance, after the angiosperm supertree of Davies et al. (2004); (b) functional trait distance of traits related to carbon fixation and leaf type (see Fig. 1); and (c) functional trait distance related to nitrogen nutrition. Species abbreviations are given in Table II. Subscripts indicate assignments to carbon and nitrogen functional groups, respectively, along the axes presented in Figure 1 and described in Table II.

permits us to study three independent variables – a species variable (S), into which phylogenetic relationships can be integrated; a carbon functional diversity variable (FDc); and a nitrogen functional diversity variable (FDn). The linear model might be written as:

$$EP = \alpha S + \beta FDc + \gamma FDn + \varepsilon. \quad (2)$$

To recognize the model of equation (1), this linear model may also be written as

$$EP = [S | FDc | FDn] \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} + \varepsilon. \quad (3)$$

We used the Rao index of quadratic entropy (Rao, 1982) to calculate the independent variables in the linear model that describe phylogenetic diversity (S) and functional diversity related to carbon fixation (FDc) and nitrogen uptake (FDn). For each plot, the Rao index can be calculated based on a vector of species abundances in the plot, and matrices of species pair dissimilarities calculated for phylogenetic position or functional trait values. We considered the total number of individuals planted in the plots to be 240 based on a planting density of 4 m^{-2} in a 1 ha plot. In mixed plots, species contributions are equal.

In a classical experimental design, the possible values of factors would be two, four or eight species, and one or two groups of functional diversity (i.e., FDc and FDn may equal 1 or 2). Within this context, a factorial design would include two or eight species, each with one or two groups, treated as discrete categories, for FDc and FDn . The complete design would include all combinations of two, four or eight species with one or two discrete groups for each axis of functional diversity, except for the impossible combination of eight species in a single group. A 16-species plot necessarily contains all groups so it is not adapted to either the factorial or the complete design. To account for species identity effects, monoculture plots for each species in the pool must be added (Loreau and Hector, 2001), even though they are not included in the classical designs.

The TropiDEP design is roughly the complete design repeated four times to eliminate species effects and to permit for continuous independent variables (Fig. 3). To avoid replicating identical plots, those with eight species and two groups for FDc (or FDn) and one group for FDn (or FDc) are repeated only twice, and a 16-species plot has been added.

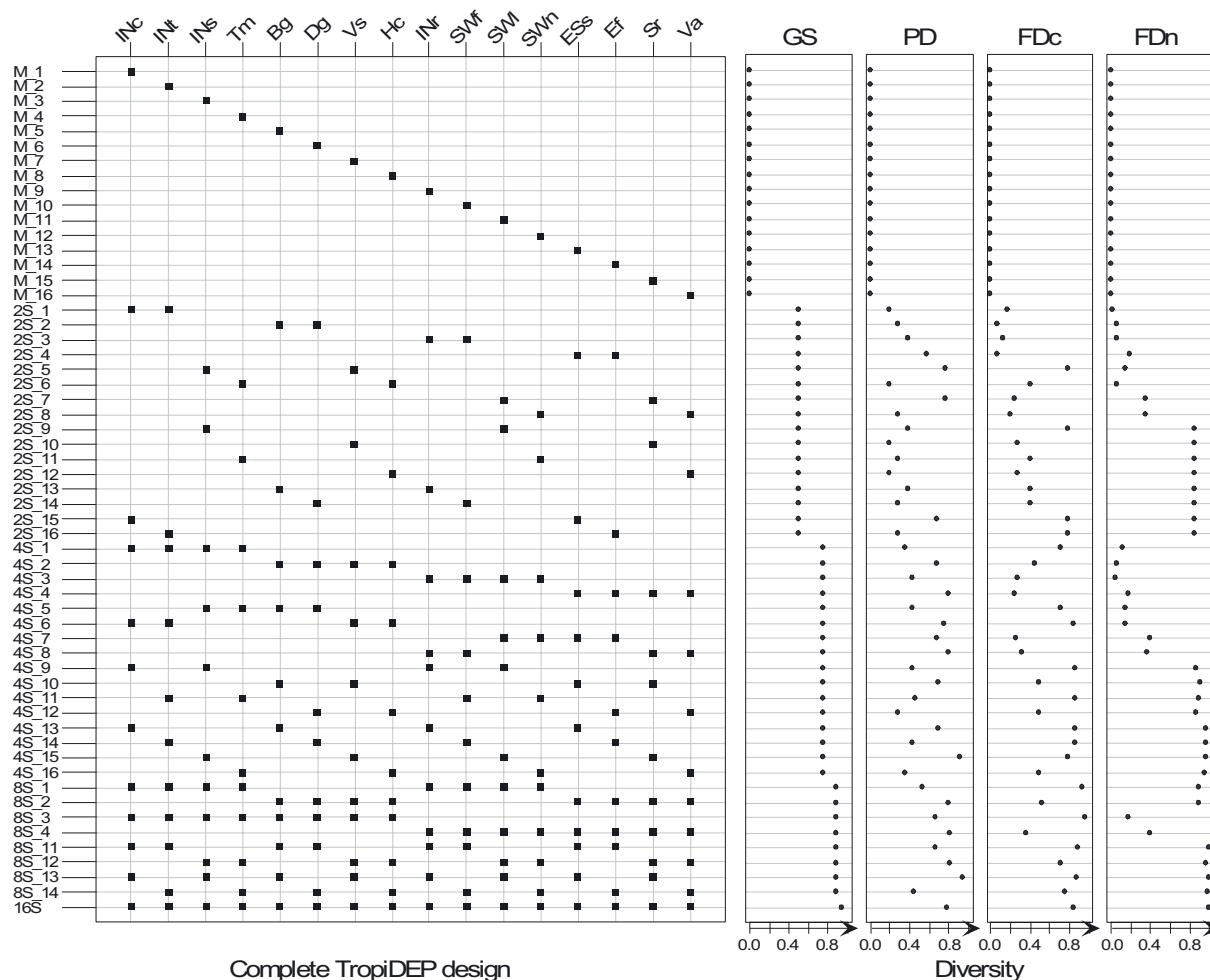


Figure 3. A summary of the 57 proposed plots in the complete TropiDEP design, including the 41 plots in the main design, 16 monocultures and one plot with all 16 species. Shown are the species planted in each plot (with equal relative densities), along with diversity estimates calculated as quadratic entropies with: equal distances among species (Gini-Simpson; GS), phylogenetic distances (PD), and functional distances along trait axes related to carbon (*FDC*) and nitrogen (*FDN*) cycling. Species abbreviations are given in Table II. All species are planted with the same number of repetitions within richness levels and overall.

4.4. Comparative performance of TropiDEP and other experimental designs

Table III presents a summary of statistical performance measures for different experimental designs. We evaluated the TropiDEP design with continuous values as they have been defined in the paper. Its performance estimates using our index are much lower than when these variables are estimated as discrete categories because their variance is reduced. Yet, it is highly probable that the model will fit better with continuous values, so its error σ^2 will be reduced. The actual effect on the variance of the parameter estimators cannot be evaluated before actual experimentation. Nevertheless, we chose this approach because it may permit the identification of diversity effects masked by group designation (cf. Wright et al., 2006).

The first three designs use discrete values including the logarithm of the number of species for species diversity, and the

number of functional groups (1 or 2) for functional diversities *FDC* and *FDN*, so they can be compared directly (Tab. III). For example, suppose we have resources for planting about 80 plots. We can choose to repeat the factorial design 10 times (80 plots), the complete design 7 times (84 plots) or the TropiDEP design twice (82 plots). The variance of the PD estimator will be 25% ($1/.78 \cdot 80/82$) greater in the TropiDEP design, compared to the factorial design. The PD estimator will also be slightly correlated to the other estimators. This can be considered as the price to be able to verify linearity. Other estimators will be almost as accurate. As such, the performance of TropiDEP is similar to that of the complete design.

The limit of our performance index is that it can not be used to compare completely different models, as we do not know anything about the model error. But it is very useful to evaluate the effect of adding or deleting plots and thus allows fine tuning of a design. For example, we might consider eliminating all of the 4-species and 16-species plots to simplify the

Table III. A comparison of statistical performances for experimental designs with sixteen species representing four functional groups along two axes of functional diversity. The factorial, complete and TropiDEP design (with categorical or continuous independent variables) are compared. Performances are relative to the reference factorial design. Corr(V1, V2) is the correlation between the estimators of the effects of variables 1 and 2. The complete design excludes the impossible combination of eight species in a single functional group. The TropiDEP design with continuous variables is based on calculations of quadratic entropies for each plot rather than discrete assignments of presence-absence of a functional group, or number of species.

	Factorial design	Complete design	TropiDEP design, discrete values	TropiDEP design, continuous values
PD ^a Performance	1	89%	78%	36%
FDc ^b Performance	1	95%	97%	44%
FDn ^b Performance	1	95%	97%	67%
Corr(PD, FDc)	0	-0.27	-0.21	-0.44
Corr(PD, FDn)	0	-0.27	-0.21	-0.37
Corr(FDc, FDn)	0	0.10	0.03	-0.01
N (plot number)	8	12	41	41
Advantages	Most efficient	Can verify linearity		Considers continuous factors
Disadvantages	Does not verify linearity	Less efficient		

^a Equivalent to $\log_2(\text{species number})$ in discrete analyses.

^b Equivalent to the number of groups (1 or 2) in discrete analyses.

TropiDEP design more towards a factorial design. The result (not shown in Tab. III) is a negligible performance improvement (1% for all factors), but a higher correlation between *FDc* and *FDn* (0.11 instead of -0.01). This 28-plot design could be repeated three times (84 plots) to be compared to the TropiDEP design repeated twice (82 plots). Estimation accuracy is not improved and linearity against the number of species can not be verified, so that design would not be retained.

5. DISCUSSION

5.1. The novelty of TropiDEP

The TropiDEP design incorporates three levels of analysis not examined to date in other studies. First, it incorporates separate functional trait axes that are predicted to influence ecosystem processes in different ways. Although each of the functional trait axes we observed is consistent with trait associations found in this and other plant communities (Roggy et al., 1999; Wright et al., 2004), the relative orthogonality of the two axes has not been reported to date. Together these two axes distinguish species that differ markedly in a suite of traits that could influence carbon and nitrogen cycling in this system (Tab. II). We suggest that distances along each of these axes would provide more interpretable results of the effects of functional diversity than the global distance measure suggested by Petchey and colleagues (Petchey and Gaston, 2002). As a result, we propose that independent variables describing functional diversity might be divided into several axes describing suites of different traits with potentially different effects on particular ecosystem processes. In other sites or ecosystems, these axes may differ from those reported here, yet the methods of analysis could be universal. The definition of separate axes of functional diversity also can improve the choice of focal species for experimental design, such that functional dis-

tances among species combinations (cf. Roscher et al., 2004) are varied deliberately along one or more axes.

A second contribution of the TropiDEP design is its consideration of phylogenetic diversity. Biodiversity can be represented not only by species and functional diversity but also by the diversity in evolutionary relationships among taxa (Faith, 1992; Forest et al., 2006). Mixtures of closely related species would then be considered less diverse than those with more distantly related species. Inclusion of this level of analysis might depend on the experimental system and the degree to which the researchers have confidence in their a priori knowledge of functional traits and the phylogenetic constraints on these traits. For example, in our species pool not all legumes fixed nitrogen, but legumes may share other traits we did not measure, such as anti-herbivore defense compounds (Wojciechowski et al., 2005). In this case, we may wish to account for correlations of unmeasured but evolutionarily constrained traits that might influence ecosystem processes, by considering the phylogenetic distance among species pairs in our sample.

We have also shown that both a multivariate measure of functional diversity and a measure of phylogenetic diversity that account for species abundances in experimental plots can be estimated using the Rao quadratic entropy. The Rao index is particularly suited to experimental designs in forest plantations where costs prohibit varying abundances experimentally (e.g., Roscher et al., 2004), and where abundances may vary through time due to self-thinning (Scherer-Lorenzen et al., 2005).

5.2. Suggestions for experimental designs

A major limitation to B-EP studies in forest ecosystem involves a trade-off between statistical rigor and feasibility of

implementation, due to the larger scale and substantial infrastructure required in tree plantations. Several general lessons can be drawn from the construction of the TropiDEP model that should advance the general field of study of functional diversity and ecosystem processes. First, regarding the definition of independent variables using a priori trait measures, researchers may want to account for particular functional traits, the results of which can be easily interpreted (Craine et al., 2002). We suggest screening traits for all potential species, from which one or multiple axes of trait variation can be defined and along which focal species can be chosen.

Second, regarding the measurement of independent variables, we suggest the relationship between diversity and ecosystem processes can only be improved by more precise measurement of the independent variables defining diversity (Wright et al., 2006). In particular, we believe that these variables can often be separated (as in our example for *FDc* and *FDn*). Moreover, the community composition of plots can be measured such that replicates for discrete variables are assigned unique calculated values for independent variables. In addition, measures such as Rao's quadratic entropy (Rao, 1982) can be employed to account not only for continuous distances among species in mixtures but also their relative abundances in each experimental unit.

A third lesson is perhaps obvious but often ignored, and is based on the wide acceptance that complete designs for a linear model are most appropriate (Gotelli and Graves, 1996). We recommend that statistical rigor should be evaluated prior to design implementation. In particular, designs should incorporate not only monoculture plots but also intermediate levels of both species-level diversity and functional diversity along the chosen axes. We propose the use of a performance measure such as that described here to compare among potential designs with particular focal species combinations and relative abundances, to evaluate completely the tradeoffs of particular designs and to choose the experimental design that best meets the objectives of a particular site and project.

Acknowledgements: This work was conducted with funding to CB from NSF OISE 03-01937. We thank Claude Millier and S. Hattenschwiler for productive discussions during the development of this project, and J. Ewel, D. Hibbs, S. Naeem, and F. Putz for constructive comments on the manuscript.

REFERENCES

- Bonal D., Sabatier D., Montpied P., Tremeaux D., and Guehl J.M., 2000. Interspecific variability of $\delta^{13}\text{C}$ among canopy trees in rainforests of French Guiana: Functional groups and canopy integration. *Oecologia* 124: 454–468.
- Botta-Dukát Z., 2005. Rao's quadratic entropy as a measure of functional diversity based on multiple traits. *J. Veg. Sci.* 16: 533–540.
- Cardinale B.J., Wrigh J.P., Cadotte M.W., Carroll I.T., Hector A., Srivastava D.S. et al., 2007. Impacts of plant diversity on biomass production increase through time because of species complementarity. *Proc. Natl. Acad. Sci., USA*, 104: 18123–18128.
- Chapin F.S., Zavaleta E.S., Eviner V.T., Naylor R.L., Vitousek P.M., Reynolds H.L., Hooper D.U., Lavorel S., Sala O.E., Hobbie S.E., Mack M.C., and Diaz S., 2000. Consequences of changing biodiversity. *Nature* 405: 234–242.
- Chave J., Muller-Landau H.C., Baker T.R., Easdale T.A., Ter Steege, H., and Webb C.O., 2006. Regional and phylogenetic variation of wood density across 2456 neotropical tree species. *Ecol. Appl.* 16: 2356–2367.
- Chessel D., Dufour, A.B., and Thioulouse, J., 2004. The ade4 package-I- One-table methods. *R News.* 4: 5–10.
- Cochran W.G. and Cox G.M., 1992. *Experimental Designs*, John Wiley and Sons, New York, 428 p.
- Cornelissen J.H.C., Lavorel S., Garnier E., Diaz S., Buchmann N., Gurvich D.E., Reich P.B., ter Steege H., Morgan H.D., van der Heijden M.G.A., Pausas J.G., and Poorter H., 2003. A handbook of protocols for standardised and easy measurement of plant functional traits worldwide. *Aust. J. Bot.* 51: 335–380.
- Craine J.M., Tilman D., Wedin D., Reich P., Tjolkter M., and Knops J., 2002. Functional traits, productivity and effects on nitrogen cycling of 33 grassland species. *Funct. Ecol.* 16: 563–574.
- Davies T.J., Barraclough T.G., Chase M.W., Soltis P.S., Soltis D.E., and Savolainen V., 2004. Darwin's abominable mystery: Insights from a supertree of the angiosperms. *Proc. Natl. Acad. Sci. USA* 101: 1904–1909.
- Diaz S. and Cabido M., 2001. Vive la difference: plant functional diversity matters to ecosystem processes. *Trends Ecol. Evol.* 16: 646–655.
- Ewel J.J., 2006. Species and rotation frequency influence soil nitrogen in simplified tropical plant communities. *Ecol. Appl.* 16: 490–502.
- Faith D.P., 1992. Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* 61: 1–10.
- Fargione J., Tilman D., Dybzinski R., Lambers J.H.R., Clark C., Harpole W.S. et al., 2007. From selection to complementarity: shifts in the causes of biodiversity-productivity relationships in a long-term biodiversity experiment. *Proc. Roy. Soc. B.* 274: 871–876.
- Forest F., Grenyer R., Rouget M. et al., 2006. Preserving the evolutionary potential of floras in biodiversity hotspots. *Nature* 445: 757–760.
- Gamfeldt L., Hillebrand H., and Jonsson P.R., 2008. Multiple functions increase the importance of biodiversity for overall ecosystem functioning. *Ecology* 89: 1223–1231.
- Gotelli N.J. and Graves G.R., 1996. *Null Models in Ecology*. Smithsonian Institution Press, Washington, 368 p.
- Grime J.P., Thompson K., Hunt R., Hodgson J.G., Cornelissen J.H.C., Rorison I.H. et al., 1997. Integrated screening validates primary axes of specialization in plants. *Oikos* 79: 259–281.
- Hector A. and Bagchi R., 2007. Biodiversity and ecosystem multifunctionality. *Nature* 448: 188–190.
- Hillebrand H., Bennett D.M., and Cadotte M.W., 2008. Consequences of dominance: a review of evenness effects on local and regional ecosystem processes. *Ecology* 89: 1510–1520.
- Hooper D.U. and Dukes J.S., 2004. Overyielding among plant functional groups in a long-term experiment. *Ecol. Lett.* 7: 95–105.
- Hooper D.U., Chapin F.S., Ewel J.J., Hector A., Inchausti P., Lavorel S., Lawton J.H., Lodge D.M., Loreau M., Naeem S., Schmid B., Setälä H., Symstad A.J., Vandermeer J., and Wardle D.A., 2005. Effects of biodiversity on ecosystem functioning: a consensus of current knowledge and needs for future research. *Ecol. Monogr.* 75: 3–36.
- Huston M.A., Aarssen L.W., Austin M.P., Cade B.S., Fridley J.D., Garnier E., Grime J.P., Hodgson J., Lauenroth W.K., Thompson K., Vandermeer J.H., and Wardle D.A., 2000. No consistent effect of plant diversity on productivity. *Science* 289: 1255.
- Isbell F.I., Polley H.W., and Wilsey, B.J., 2009. Biodiversity, productivity and the temporal stability of productivity: patterns and processes. *Ecol. Lett.* doi: 10.1111/j.1461-0248.2009.01299.x.
- Lamb D., Erskine P.D., and Parotta J., 2005. Restoration of degraded tropical forest landscapes. *Science* 310: 1628–1632.

- Loreau M. and Hector A., 2001. Partitioning selection and complementarity in biodiversity experiments. *Nature* 412: 72–76.
- Loreau M., Naeem S., Inchausti P., Bengtsson J., Grime J.P., Hector A. et al., 2001. Biodiversity and ecosystem functioning: current knowledge and future challenges. *Science* 294: 804–808.
- Mason N.W.H., MacGillivray K., Steel J.B., and Wilson J.B., 2003. An index of functional diversity. *J. Veg. Sci.* 14: 571–578.
- Moles A.T., Ackerly D.D., Webb C.O., Tweddle J.C., Dickie J.B., and Westoby M., 2005. A brief history of seed size. *Science* 307: 576–580.
- Naeem S. and Wright J.P., 2003. Disentangling biodiversity effects on ecosystem functioning: deriving solutions to a seemingly insurmountable problem. *Ecol. Lett.* 6: 567–579.
- Parotta J.A. and Knowles O.H., 1999. Restoration of tropical moist forest on bauxite-mined lands in the Brazilian Amazon. *Restor. Ecol.* 7: 103–116.
- Pavoine S., Ollier S., and Dufour A.-B., 2005. Is the originality of a species measurable? *Ecol. Lett.* 8: 579–586.
- Petchey O.L. and Gaston K.J., 2006. Functional diversity: back to basics and looking forward. *Ecol. Lett.* 9: 741–758.
- Petchey O.L., Hector A., and Gaston K.J., 2004. How do different measures of functional diversity perform? *Ecology* 85: 847–857.
- Phillips O.L., Malhi Y., Higuchi N., Laurance W.F., Núñez P.V., Vázquez R.M., Laurance S.G., Ferreira L.V., Stern M., Brown S., and Grace J., 1998. Changes in the carbon balance of tropical forests: evidence from long-term plots. *Science* 282: 439–442.
- Polley H.W., Wilsey B.J., and Derner J.D., 2007. Dominant species constrain effects of species diversity on temporal variability in biomass production of tallgrass prairie. *Oikos* 116: 2044–2052.
- Rao C.R., 1982. Diversity and dissimilarity coefficients: a unified approach. *Theor. Popul. Biol.* 21: 24–43.
- Reich P., Tilman D., Naeem S., Ellsworth D., Knops J., Craine J. et al., 2004. Species and functional group diversity independently influence biomass accumulation and its response to CO₂ and N. *Proc. Natl. Acad. Sci. USA* 101: 10101–10106.
- Ricotta C., 2005. A note on functional diversity measures. *Basic Appl. Ecol.* 6: 479–486.
- Roggy J.C., Prévost M.F., Gourbière F., Casabianca H., and Garbaye J., 1999. Leaf natural ¹⁵N abundance and total N concentration as potential indicators of plant N nutrition in legumes and pioneer species in a rain forest of French Guiana. *Oecologia* 120: 171–182.
- Roscher C., Schumacher J., Baade J., Wilcke W., Gleixner G., Weisser W.W. et al., 2004. The role of biodiversity for element cycling and trophic interactions: an experimental approach in a grassland community. *Basic Appl. Ecol.* 5: 107–121.
- Scherer-Lorenzen M., Potvin C., Koricheva J., Schmid B., Hecto, A., Bornik Z., Reynolds G., and Schulze E.-D., 2005. The design of experimental tree plantations for functional biodiversity research. In: Scherer-Lorenzen M., Körner C, and Schulze E.-D. (Eds.), *Forest Diversity and Function: Temperate and Boreal Systems*, Springer-Verlag, Berlin, pp. 347–376.
- Schimann H., Ponton S., Hattenschwiler S., Ferry B., Lensi R., Domenach A.M., and Roggy J.C., 2008. Differing nitrogen use strategies of two tropical rainforest late successional tree species in French Guiana: Evidence from ¹⁵N natural abundance and microbial activities. *Soil Biol. Biochem.* 40: 487–494.
- Schwartz M.W., Brigha, C.A., Hoeksema J.D., Lyons K.G., Mills M.H., and van Mantgem P.J., 2000. Linking biodiversity to ecosystem function: implications for conservation ecology. *Oecologia* 122: 297–305.
- Simpson E.H., 1949. Measurement of diversity. *Nature* 163: 688.
- Tilman D., Reich P.B., Knops J., Wedin D., Mielke T., and Lehman C., 2001. Diversity and productivity in a long-term grassland experiment. *Science* 294: 843–845.
- Wojciechowski M.F., Lavin M., and Sanderson M.J., 2005. A phylogeny of legumes (*Leguminosae*) based on analysis of the plastid matK gene resolves many well-supported subclades within the family. *Am. J. Bot.* 91: 1846–1862.
- Wright I.J., Reich P.B., Westoby M., Ackerly D.D., Baruch Z., Bongers F. et al., 2004. The worldwide leaf economics spectrum. *Nature* 428: 821–827.
- Wright J.P., Naeem S., Hector A., Lehman, C., Reich P.B., Schmid B., and Tilman D., 2006. Conventional functional classification schemes underestimate the relationship with ecosystem functioning. *Ecol. Lett.* 9: 111–120.
- Zhang Q.G. and Zhang D.Y., 2007. Colonization sequence influences selection and complementarity effects on biomass production in experimental algal microcosms. *Oikos* 116: 1748–1758.

APPENDIX M

Dynamics of aboveground carbon stocks in a selectively logged tropical forest

Blanc, L., M. Echard, B. Hérault, D. Bonal, E. Marcon, J. Chave et C. Baraloto (2009). « Dynamics of aboveground carbon stocks in a selectively logged tropical forest ». In : Ecological Applications 19.6, p. 1397–1404.

Dynamics of aboveground carbon stocks in a selectively logged tropical forest

LILIAN BLANC,¹ MARION ECHARD,¹ BRUNO HERAULT,² DAMIEN BONAL,³ ERIC MARCON,⁴ JÉRÔME CHAVE,⁵
AND CHRISTOPHER BARALOTO^{3,6}

¹CIRAD, UMR "Ecologie des Forêts de Guyane," 97379 Kourou, French Guiana

²Université des Antilles et de la Guyane, UMR "Ecologie des Forêts de Guyane," 97379 Kourou, French Guiana

³INRA, UMR "Ecologie des Forêts de Guyane," 97379 Kourou, French Guiana

⁴ENGREF, UMR "Ecologie des Forêts de Guyane," 97379 Kourou, French Guiana

⁵Université Paul Sabatier, CNRS, Laboratoire Evolution et Diversité Biologique, Toulouse, France

Abstract. The expansion of selective logging in tropical forests may be an important source of global carbon emissions. However, the effects of logging practices on the carbon cycle have never been quantified over long periods of time. We followed the fate of more than 60 000 tropical trees over 23 years to assess changes in aboveground carbon stocks in 48 1.56-ha plots in French Guiana that represent a gradient of timber harvest intensities, with and without intensive timber stand improvement (TSI) treatments to stimulate timber tree growth. Conventional selective logging led to emissions equivalent to more than a third of aboveground carbon stocks in plots without TSI (85 Mg C/ha), while plots with TSI lost more than one-half of aboveground carbon stocks (142 Mg C/ha). Within 20 years of logging, plots without TSI sequestered aboveground carbon equivalent to more than 80% of aboveground carbon lost to logging (−70.7 Mg C/ha), and our simulations predicted an equilibrium aboveground carbon balance within 45 years of logging. In contrast, plots with intensive TSI are predicted to require more than 100 years to sequester aboveground carbon lost to emissions. These results indicate that in some tropical forests aboveground carbon storage can be recovered within half a century after conventional logging at moderate harvest intensities.

Key words: aboveground biomass; carbon sequestration; deforestation; French Guiana; global change; timber stand improvement; tropical forests.

INTRODUCTION

Tropical forests represent a major reservoir of global carbon, accounting for up to half of the estimated 558 Pg of carbon stored in vegetation (Houghton 2005), with an estimated 86 Pg of carbon in the Amazon basin alone (Saatchi et al. 2007). Land use changes in the tropics have become an increasing concern for their potential impacts on the global carbon cycle and climate change (Carpenter et al. 2006, Solomon et al. 2007).

Remote sensing studies, combined with models of land-cover dynamics, have revealed that over the past 20 years Amazonian deforestation contributed an estimated 0.30 Pg of carbon per year to the atmosphere (Ramankutty et al. 2007). These estimations ignore selective timber logging, which is largely invisible to classic remote sensing technologies. Degradation of tropical forests by selective logging, fuelwood removal, and fire encroachment contributes even further to these emissions (Nepstad et al. 1999, DeFries et al. 2002, Peres 2006, Achard et al. 2007).

Recently, high-resolution remote sensing studies have suggested that the contribution of selective logging in

the Amazon could release up to 0.08 Pg C/yr, or 25% of the loss due to land use change through deforestation (Asner et al. 2005). And models from unlogged plots predict that selectively logged forests may lose of up to 70% of carbon storage potential (Bunker et al. 2005). Yet few direct data exist to evaluate the long-term contributions of selective logging to the carbon balance of tropical forests, and the trajectory of carbon sequestration in stands regenerating after logging.

In this paper, we analyze data collected over a 20-yr period following selective logging in replicated permanent plots in French Guiana. We examine the contribution of tree growth and recruitment vs. tree harvesting and death to changes in aboveground carbon fluxes over this period, and we use a bookkeeping model to extrapolate the long-term consequences of differences in forest dynamics after logging to the aboveground carbon balance of this tropical forest.

METHODS

All inventories were conducted at the Paracou experimental site (5°18' N, 52°55' W), a lowland tropical rain forest near Sinnamary, French Guiana (Gourlet-Fleury et al. 2004). The site receives nearly two-thirds of the annual 3041 mm of precipitation between mid-March and mid-June, and <50 mm per month in

Manuscript received 21 August 2008; revised 8 December 2008; accepted 19 December 2008. Corresponding Editor: E. Cuevas.

⁶ Corresponding author. E-mail: chris.baraloto@ecofog.gf

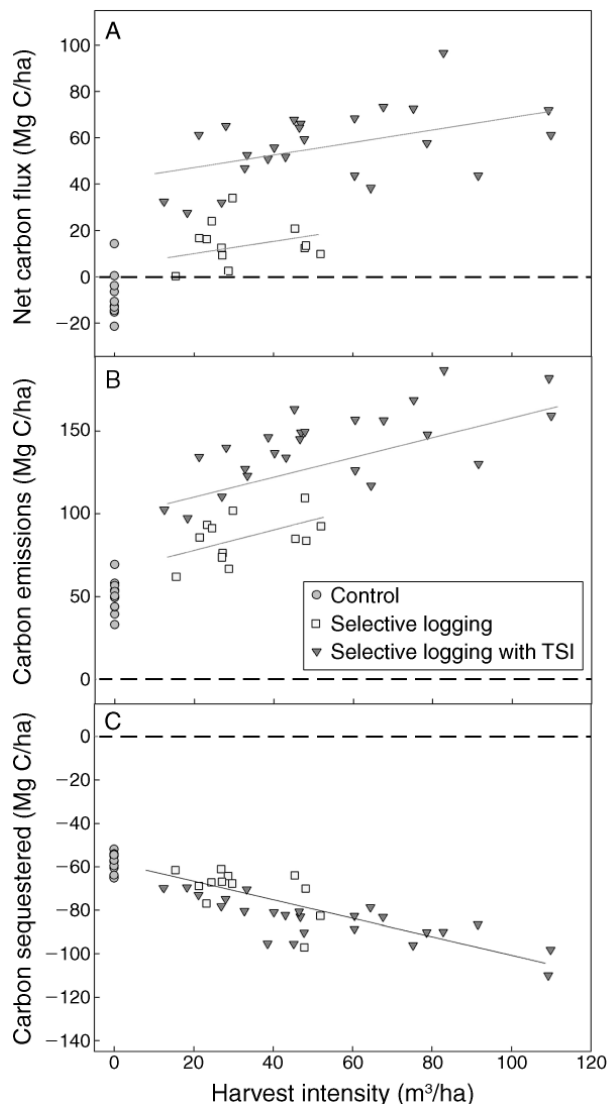


FIG. 1. Effects of logging intensity on carbon flux (the dashed line corresponds to a flux of 0) for aboveground biomass over a 20-yr period after selective logging. Each point represents the cumulative sum of carbon gains and losses for the period 1987–2007 from a 1.56-ha plot subjected to one of three treatments: control, selective logging only, and selective logging with timber stand improvement (TSI). (A) Net carbon flux; (B) carbon emissions from tree harvesting and death, estimated using decomposition constants and species-specific wood density; and (C) carbon sequestered by tree growth and recruitment, estimated from allometric measures incorporating tree diameter and species-specific wood density. Lines illustrate linear models fitted from ANCOVA performed to test the effect of harvest intensity and treatment on net flux, carbon emissions, and carbon sequestered over the 20-yr observation period; all coefficients were significant at $P < 0.0001$. For logged plots without TSI, flux = $5.96 + 0.257 \times \text{intensity}$, with the intercept changing to 43.0 for plots with TSI. For plots without TSI, carbon emissions increased with harvest intensity; flux = $63.9 + 0.560 \times \text{intensity}$, with the intercept changing to 101.0 for plots with TSI. Across all logged plots, carbon sequestered increased with harvest intensity (sequestered flux = $-57.9 - 0.313 \times \text{intensity}$).

September and October (Gourlet-Fleury et al. 2004). The most common soils in Paracou are the shallow ferralitic soils limited in depth by a more or less transformed loamy saprolithe (Gourlet-Fleury et al. 2004). Some very thick ferralitic soils, with free vertical drainage, are primarily encountered on the highest residual summits of the area (~ 40 m above sea level). Surface soils at the site exhibit similar carbon and nitrogen properties to other eastern South American lowland forest soils, but have very low phosphorus availability (Baraloto and Goldberg 2004).

Beginning in 1984, 48 permanent tree plots totaling 75 ha of tropical rain forest were established at the site. All trees ≥ 10 cm diameter breast height (dbh) have been identified, tagged, mapped, and measured in these plots (see Appendix A). From 1986 to 1988 different logging treatments were applied to 36 plots, with 12 plots established as controls. In 12 logged plots, selected timbers were extracted, with an average of 10.4 trees (from 5.8 to 15.4 trees) ≥ 50 cm dbh removed per hectare, corresponding to a timber volume average of $32.5 \text{ m}^3/\text{ha}$ (from 15.4 to $51.8 \text{ m}^3/\text{ha}$). In 24 plots in which intensive timber stand improvement (TSI) was applied, logging intensity averaged 20.6 trees (from 5.1 to 41.7 trees) ≥ 50 cm dbh removed per hectare, corresponding to a timber volume average of $53.4 \text{ m}^3/\text{ha}$ (from 12.4 to $109.8 \text{ m}^3/\text{ha}$). Subsequent poison girdling of selected noncommercial species killed an average of 16.6 trees ≥ 40 cm dbh/ha (Appendix A: Table A1). Complete inventories were conducted annually until 1995, then every two years, with a most recent census in 2007.

We estimated the aboveground pools of carbon in trees from allometric formulas using wood specific gravity and trunk diameter (Chave et al. 2005), and we converted dry biomass into carbon assuming a carbon : dry biomass ratio of 0.5 (Penman et al. 2003). Aboveground carbon emissions were estimated by integrating carbon stored in each tree with census data for tree death and a model of coarse woody debris decomposition (Chambers et al. 2000). An additional source of aboveground carbon emissions was due to harvested logs, which we separated into decomposing and harvested proportions using field reports of the experimental logging operation in which the fate of 2-m segments of all harvested trees was followed. For harvested segments that arrived in sawmills, we estimated that only one-third of the aboveground carbon was stored in woody commercial products. The remaining two-thirds of aboveground carbon were considered as immediate emissions following wood transformation practices employed in most parts of tropical America (Keller et al. 2004).

RESULTS AND DISCUSSION

During the 1987–2007 observation period, logged plots did not recover their original aboveground carbon stock. Over this period, net carbon flux was positive and ranged between 0.3 and 96.7 Mg C/ha (Fig. 1A; carbon

emissions are described as positive fluxes according to IPCC conventions; Penman et al. 2003), depending on the timber volume harvested and on the logging treatment. Conventionally logged plots had an average net flux of 14.3 Mg C/ha. In contrast, logged plots with timber stand improvement (TSI) had almost four times the average net positive carbon flux over the period (56.7 Mg C/ha). Differences in net aboveground carbon flux among treatments were due more to differential emissions from dead and harvested trees (Fig. 1B) than to differential sequestration from subsequent tree growth and recruitment (Fig. 1C). Plots logged with TSI released two-thirds more aboveground carbon (141.2 Mg C/ha) over the 20-yr period than plots without TSI (85.0 Mg C/ha).

All 36 logged plots remained sources of aboveground carbon emissions for 10–12 years following harvesting activities (Fig. 2). Emissions from harvested trees peaked immediately after logging as an estimated two-thirds of the harvested biomass was burned in sawmills (see Appendix B). Annual emissions from harvested trees peaked at 9.1 Mg C/ha one year after logging, with the most intensively logged plots releasing >20 Mg C/ha in that year. Emissions due to increased mortality after logging were less severe but more sustained, both because of damaged trees that died several years after harvesting and because these emissions arose from coarse woody decomposition within the forest rather than from burning immediately after the logging operation. Plots with TSI maintained emissions of >10 Mg C/ha from tree mortality for several years after logging, as many of the poisoned trees did not die until several years after treatment and the canopy gaps they created resulted in additional tree mortality (Gourlet-Fleury et al. 2004).

Beginning about 10 years after logging, most logged plots began to sequester more aboveground carbon than they released from decomposing dead and harvested trees, and some logged plots actually had more positive net annual aboveground carbon flux than unlogged plots (Fig. 2). This reaction was due not only to reduced emissions from dead and harvested trees, but also to enhancement of tree growth and recruitment by >50% in logged vs. unlogged plots over a 20-yr period. Aboveground carbon sequestration due to tree recruitment (Appendix A: Fig. A1) peaked in most plots about 10 years after logging and this peak compensated for the

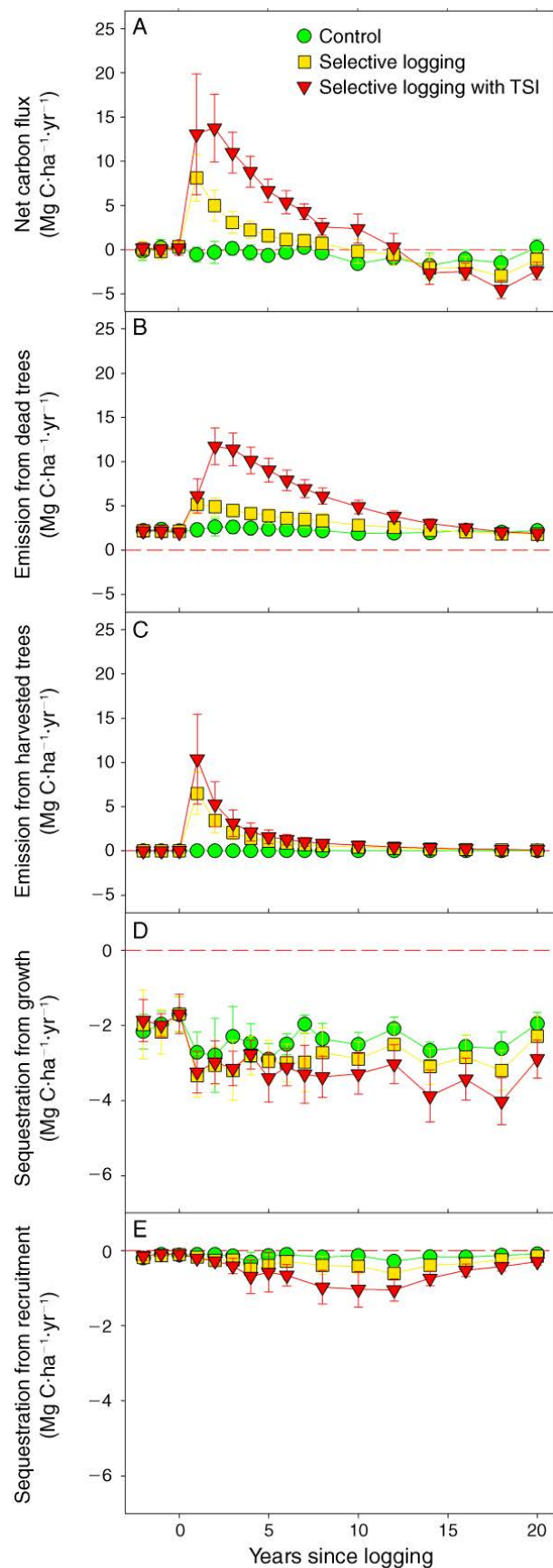


FIG. 2. Net annual flux of carbon from aboveground biomass and its components across 23 years in logged and unlogged forest. Shown are the means (\pm SD) of 12 (or 24 for logging + TSI) 1.56-ha plots in each of three treatments including control (green circles), selective logging only (yellow squares), and selective logging with timber stand improvement (red triangles). (A) Net annual carbon flux; (B) annual carbon emissions from tree death; (C) annual carbon emissions from tree harvesting; (D) annual carbon stored by tree growth; and (E) annual carbon stored by tree recruitment.

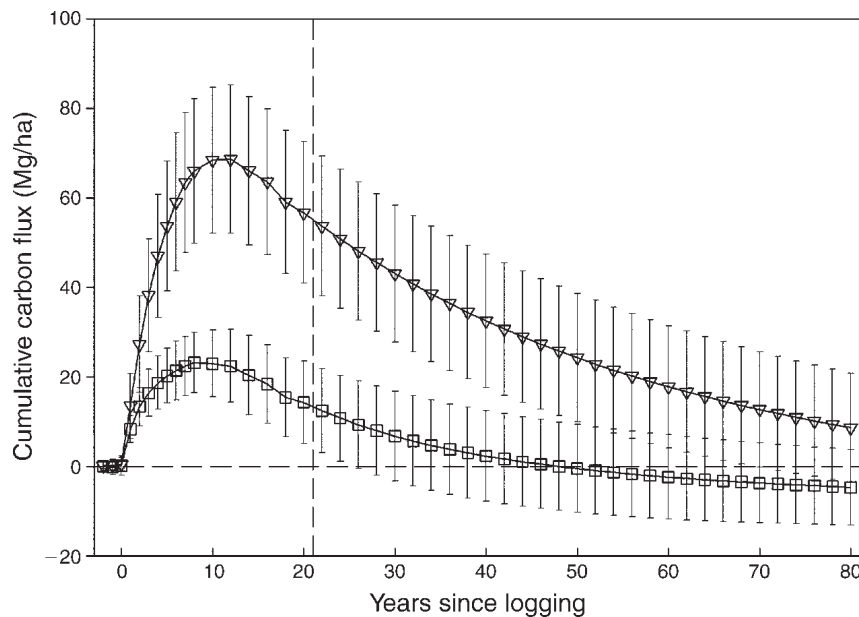


FIG. 3. Projected cumulative carbon flux of aboveground biomass in logged forest plots. Points indicate observed (to 2007; indicated by the vertical dashed line) and simulated mean values for 1.56-ha plots that were logged only (squares) or logged with timber stand improvement (triangles). Error bars indicate the maximum and minimum values among plots of each treatment. Simulations projected reduced carbon sequestration through time with maximum carbon stocks equivalent to mean values observed in unlogged plots at the most recent (2001–2007) census interval.

continued increased emissions from earlier tree mortality in these plots. Enhanced tree growth in logged plots occurred within two years of harvesting and was maintained until the most recent census.

To determine the long-term consequences of differences in forest dynamics after logging to the aboveground carbon balance of this tropical forest, we extrapolated the time trajectories of net aboveground carbon fluxes in the plots into the future using a bookkeeping model of aboveground carbon recovery through time (see Appendix B). The average time to recover the original aboveground carbon stock is projected to be 45 years for logged plots without TSI (Fig. 3). However, plots with timber stand improvement would take more than 100 years to recover their pre-logging aboveground carbon stocks.

We believe these results represent a conservative estimate of aboveground carbon stock recovery in selectively logged tropical forests for two reasons. First, logging operations at Paracou were typical of conventional logging (CL) systems, with no effort to implement any of the recommendations suggested for reduced-impact logging (RIL) techniques, including reduced damage to forest structure and future crop trees, wasted timber left in the forest, and poor sawmill transformation efficiency (Appendix A: Fig. A3). No concerted effort was made to reduce impacts on the residual stand. Indeed, the surface areas damaged in the plots (Appendix A: Table A1) are among the highest values reported in the literature for similar extracted timber volumes (Feldpausch et al. 2005). In other tropical forests, where reduced-impact logging techniques have

been implemented, the contribution to emissions from incidentally killed or damaged trees can be reduced by more than one-third (Pinard and Cropper 2000).

In addition, about one-third of the timber volume harvested at Paracou ($34\% \pm 8\%$ SD) was left to decompose on site. Training chainsaw operators in directional felling techniques could lead to marked reductions in this waste rate (Sist and Brown 2004). Recent improvements in chain-of-custody accompanying the certification of forest products in many tropical RIL operations will result in less transformation waste and thus reduced emissions from harvested trees. In operations attempting to maximize transformation efficiency, waste is estimated to be $<30\%$. In summary, the use of RIL techniques at Paracou could have reduced initial emissions by $\sim 50\%$.

Our results from Paracou may also represent conservative patterns of aboveground carbon sequestration relative to other Neotropical forests, as evidenced from three factors of forest dynamics that drive carbon accumulation after logging: growth, mortality, and recruitment. The degree to which our results can be extrapolated to other sites depends on the site specificity of the response through time of these three factors to logging-associated disturbance. To our knowledge, long-term data in Neotropical logged permanent plots are available only for six sites in Brazil and one in Suriname (Table 1). The average stand-level diameter increment in the control (unlogged) plots at Paracou was equal to 1.2 mm/yr. This is low compared to other experimental studies (range, 1.1–3.2 mm/yr; Table 1). Stand-level mortality was also low (1.1%; range of the other study

TABLE 1. A comparison of forest dynamics in unlogged and logged plots at Paracou and six other Neotropical forest sites.

Forest dynamics	Paracou	CELOS 67/9B	Jari	Tapajos km 67	Tapajos km 114	Paragominas	BIONTE
Growth (mm/yr)							
Unlogged	1.2	1.1	2.1	2.0	1.4	3.2	1.6
Initial	2.2	4.1	3.8	3.0	3.6	2.9	2.9
Subsequent	2.1	4.2	3.1	2.5	2.2	3.4	1.9
Mortality (%)							
Unlogged	1.1	2.0	1.2	1.7	1.2	1.8	1.0
Initial	2.2	2.1	2.6	2.4	2.7	2.4	2.9
Subsequent	1.6	2.1	1.2	2.6	2.1		
Recruitment (%)							
Unlogged	1.1		1.5	4.8	1.2	1.7	
Initial	4.2		2.6	5.2	7.8	2.4	
Subsequent	5.1		2.0	1.8	1.6		

Notes: "Initial" refers to averages across the first 6–8 years after logging; "subsequent" values are for available data >8 years after logging. Data are from plots where no timber stand improvement (TSI) treatments were employed. Size classes and data sources for each site (other than Paracou) are: CELOS, unrefined logged plots, stems >20 cm dbh (de Graaf 1986, Jonkers 1987, de Graaf et al. 1999); Jari, stems >20 cm dbh (De Azevedo 2006); Tapajos km 67, stems >5 cm dbh (Silva et al. 1995); Tapajos km 114, stems >5 cm dbh (De Carvalho et al. 2004, De Oliveira 2005); Paragominas, control and conventional logging sites, stems >10 cm dbh (Vidal et al. 2002, Vidal 2003); BIONTE, stems >10 cm dbh (Chambers et al. 2004, Rice et al. 2004, Vieira et al. 2004, 2005).

sites, 1.0–2.0%). In most of these forests, diameter growth rate increased up to fourfold after logging. What is the contribution of maintaining higher growth rates to aboveground biomass accumulation after logging? An important contrast among sites occurs for the subsequent response of the forest to logging after about six years post logging (Table 1). After this period, growth slowed in four of the five Amazonian forests, but not in the Guiana Shield sites (Paracou and CELOS; Table 1). In the BIONTE experiment, the only Amazonian study site with long-term data comparable with Paracou, aboveground biomass stocks returned to pre-logging levels after 16 years (Chambers et al. 2004). Hence the results of our study are consistent with other slow-growing Amazonian forest sites.

A second process influencing aboveground carbon sequestration that may differ among forests is tree mortality. Annualized mortality rates in unlogged Neotropical forests range between 0.8% and 2% per year (Swaine et al. 1987, Lugo and Scatena 1996, Lewis et al. 2004), within which fall the observations in unlogged plots at Paracou (Table 1). Following an initial period of high mortality directly after logging at Paracou, mortality rates remained ~30% higher than those observed prior to logging for about eight years after logging. Thereafter (since 1997), mortality rates between logged and unlogged plots have been similar. Reports from other permanent plots show increases in mortality after logging ranging from ~33% at Tapajos km 67 to nearly 200% at BIONTE, Brazil (Table 1). However, the few data that exist for longer term mortality rates show contrasting patterns. At CELOS and Tapajos km 67, mortality remained high up to 16 years after logging, whereas the Jari plots (and to a lesser extent the Tapajos km 114 plots) showed declines through time similar to those observed at Paracou (Table 1).

A third process that might influence aboveground carbon sequestration is recruitment of new stems. Recruitment rates in unlogged forest at Paracou (Appendix A: Fig. A1) are on the lower end of the range reported for Neotropical forests (0.8–2.32%, mean = 1.84%; Phillips and Gentry 1994, Laurance et al. 1998, Lewis et al. 2004). Given the local dominance of species with heavier wood from families including Chrysobalanaceae and Lecythydaceae, recruited stems appear to result in equal contributions of aboveground carbon as in other Amazonian sites with higher recruitment rates but lower community wood densities (Fig. 2E; ter Steege et al. 2003, 2006, Chave et al. 2006).

Few studies have reported recruitment rates after logging in permanent plots, and comparisons are difficult because these studies were using different minimum dbh for inventories. At Paracou, where the minimum dbh for recruitment was 10 cm, rates nearly quadrupled. A similar result was found at Tapajos km 114, but entry dbh there was 5 cm. More modest gains were observed at Paragominas (10 cm) and at Jari (20 cm). Results from Tapajos km 67 are difficult to interpret, with a very high rate of recruitment in unlogged plots and a large reduction in recruitment rates in the second 6-yr period following logging (Table 1). Nonetheless, the consistency in available data on recruitment rates and their response to logging disturbance suggests that results from the Paracou site can be extrapolated to other Neotropical forests.

CONCLUSION

Conservation strategies in tropical forests must compromise among multiple objectives, including not only the preservation of biodiversity and the integrity of global biogeochemical cycles, but also the economic prosperity of landholders (Soares et al. 2006, Foley et

al. 2007). Selective logging has been promoted as a tool enabling local landholders to maintain forest cover while deriving economic benefits from timber extraction (Holmes et al. 2002). However, critics have maintained that logging will negatively impact ecosystem processes, including aboveground carbon storage potential (Keller et al. 2004, Asner et al. 2005, Bunker et al. 2005). Our results suggest that selectively logged forest can recover aboveground carbon lost to emissions within a few decades if several conditions are met.

In French Guiana, mean harvest intensity is 14 m³/ha and maximum harvest never exceeds 43 m³/ha (Gourlet-Fleury et al. 2004). This is consistent with other Amazonian forests, where harvest intensity averages 23 m³/ha (Keller et al. 2004, Asner et al. 2005, Feldpausch et al. 2005). If other Amazonian forests demonstrate similar or more rapid aboveground carbon recovery potential as Paracou, the average 0.08 Pg/yr of aboveground carbon released by selective logging in the Amazon (Asner et al. 2005) could be balanced by subsequent sequestration in the subsequent 40 years (committed flux; Ramankutty et al. 2007). Although cutting cycles in French Guiana (where demand for timber is low) are currently set at 65 years (Gourlet-Fleury et al. 2004), in most neighboring countries cycles of less than 30 years are employed (Zarin et al. 2007). On this schedule, it is unlikely that aboveground carbon stocks will be replenished, even if conventional logging is conducted at low harvest intensity.

Under conventional timber production, a large proportion of aboveground carbon emissions into the atmosphere are believed to be caused by unsustainable logging practices (Gullison et al. 2007). A recent analysis estimates that the use of improved timber harvest practices in tropical forests could retain 0.16 Tg C/yr, or >10% of the carbon released by deforestation (1.5 Tg/yr; Putz et al. 2008). These results strongly argue for promoting improved forest harvesting practices in Amazonian forests.

Timber stand improvement methods have been found to increase tree growth rates by 9–100% in several long-term tropical forestry concessions, including Paracou (Jonkers 1987, Gourlet-Fleury et al. 2004, Peña-Claros et al. 2008). However, our results suggest a possible trade-off between such rapid recovery of timber volume and the time to recovery of aboveground carbon stocks. At Paracou, where extreme TSI treatments were implemented (additional reduction of basal area by an average of 20%; Appendix A: Table A1), growth rates of some commercial species doubled (Gourlet-Fleury et al. 2004), but so does the estimated time to recovery of aboveground carbon stocks. In La Chonta, Bolivia, for example, moderate growth rate increases of up to 60% were observed in plots receiving less intensive TSI (reducing basal area by 10%; Peña-Claros et al. 2008), where we might predict a more rapid recovery of aboveground carbon stocks. Thus, the effects of TSI treatments may vary under different forest management

operations. In particular, the use of TSI treatments in concert with reduced-impact logging (RIL) techniques merits exploration as a means by which to stimulate timber production with less reduction in aboveground carbon storage than we observed. For example, if the increased carbon sequestration observed under TSI from increased growth (Fig. 2D) were maintained for an additional decade, the estimated recovery time for carbon stocks would be reduced by >10%. For now, our results suggest that an appropriate compromise between timber production and carbon flux recovery will be necessary to meet regional management objectives. An additional consideration for TSI involves its potential to reduce the local abundance of tree species of limited timber value but exceptional conservation value for animal habitat and food.

Finally, selective logging is rarely an isolated activity and is frequently accompanied by other anthropogenic disturbances that may affect carbon sequestration and other important conservation goals (Asner et al. 2006, Peres 2006, Soares et al. 2006, Foley et al. 2007). Often, the road system used to remove timber from forest sites is subsequently used by colonists who hunt, remove more timber, or clear land for agriculture, which can significantly affect aboveground carbon storage potential (Asner et al. 2006, Foley et al. 2007). At Paracou these threats were mitigated by protection of the research site, and similar protection will likely be necessary in other tropical forests, either by local landholders and/or regional managers, to enable selectively logged forests to recover aboveground carbon lost to emissions (Nepstad et al. 2006). Nonetheless, we believe our results support a continued debate regarding the conservation value of management strategies that include selective logging.

ACKNOWLEDGMENTS

The project received financial support from the Ministry of Overseas Departments (CORDET), the National Forestry Fund (FFN), the State–Region Plan Contract, the BGF section of Ecofor, and the European Structural Funds (Guyafor project). The project had logistic support from SILVOLAB and Ecofor. However nothing would have been possible without the basic annual funding provided by CIRAD, the institution in charge of Paracou.

The authors dedicate this paper to L. Schmitt who greatly contributed to the Paracou experimental site. Along with P. Pétronelli, he supervised plot establishment and led forest inventories in Paracou since 1984. This study would not have been possible without their tremendous involvement in Paracou. The authors also gratefully acknowledge the CIRAD field assistants for their contribution in data collection in the Paracou experimental site: D. Max, O. N'Gwete, Mo. Baisie, Mi. Baisie, K. Ficadici, A. Etienne, F. Kwasie, K. Martinus, P. Naisso, and R. Santé. The authors thank N. Blanc, N. Haumont, S. Vrot, B. Rosset, and A. Jolivot for their work in data checking and management. Valuable contributions were made by J. Fabre and F. Wagner for statistical analyses. Finally, the authors thank C. E. T. Paine, C. Rockwell, J. Ewel, S. Gourlet-Fleury, F. Putz, E. G. Schurr, and P. Sist for valuable reviews and comments.

LITERATURE CITED

- Achard, F., R. DeFries, H. Eva, M. Hansen, P. Mayaux, and H. J. Stibig. 2007. Pan-tropical observations and mid-resolution monitoring of deforestation. *Environmental Research Letters* 2:045022.
- Asner, G. P., E. N. Broadbent, P. J. C. Oliveira, M. Keller, D. E. Knapp, and J. N. M. Silva. 2006. Condition and fate of logged forests in the Brazilian Amazon. *Proceedings of the National Academy of Sciences (USA)* 103:12947–12950.
- Asner, G. P., D. E. Knapp, E. N. Broadbent, P. J. C. Oliveira, M. Keller, and J. N. Silva. 2005. Selective logging in the Brazilian Amazon. *Science* 310:480–482.
- Baraloto, C., and D. E. Goldberg. 2004. Microhabitat associations and seedling bank dynamics in a neotropical forest. *Oecologia* 141:701–712.
- Bunker, D. E., F. DeClerck, J. C. Bradford, R. K. Colwell, I. Perfecto, O. L. Phillips, M. Sankaran, and S. Naeem. 2005. Species loss and aboveground carbon storage in a tropical forest. *Science* 310:1029–1031.
- Carpenter, S. R., R. DeFries, T. Dietz, H. A. Mooney, S. Polasky, W. V. Reid, and R. J. Scholes. 2006. Millennium ecosystem assessment: research needs. *Science* 314:257–258.
- Chambers, J. Q., N. Higuchi, J. P. Schimel, L. V. Ferreira, and J. M. Melack. 2000. Decomposition and carbon cycling of dead trees in tropical forests of the central Amazon. *Oecologia* 122:380–388.
- Chambers, J. Q., N. Higuchi, L. M. Teixeira, J. dos Santos, S. G. Laurance, and S. E. Trumbore. 2004. Response of tree biomass and wood litter to disturbance in a Central Amazon forest. *Oecologia* 141:596–611.
- Chave, J., et al. 2005. Tree allometry and improved estimation of carbon stocks and balance in tropical forests. *Oecologia* 145:87–99.
- Chave, J., H. Muller-Landau, T. Baker, T. Easdale, H. ter Steege, and C. O. Webb. 2006. Regional and phylogenetic variation of wood density across 2456 neotropical tree species. *Ecological Applications* 16:2356–2367.
- De Azevedo, C. P. 2006. Dinâmica de florestas submetidas a manejo na Amazônia oriental: experimentação e simulação. Tese de doutorado. Universidade Federal do Paraná, Curitiba, Brasil.
- De Carvalho, J. O. P., J. N. M. Silva, and J. C. A. Lopes. 2004. Growth rate of a terra firma rain forest in Brazilian Amazonia over an eight-year period in response to logging. *Acta Amazonia* 34:209–217.
- DeFries, R. S., R. A. Houghton, M. C. Hansen, C. B. Field, D. Skole, and J. Townshend. 2002. Carbon emissions from tropical deforestation and regrowth based on satellite observations for the 1980s and 1990s. *Proceedings of the National Academy of Science (USA)* 99:14256–14261.
- de Graaf, N. R. 1986. A silvicultural system for natural regeneration of tropical rainforest in Suriname. Dissertation. Agricultural University Wageningen, The Netherlands.
- de Graaf, N. R., R. L. H. Poels, and R. Van Rampaey. 1999. Effect of silvicultural treatment on growth and mortality of rainforest in Suriname over long periods. *Forest Ecology and Management* 124:123–135.
- De Oliveira, L. C. 2005. Efeito da exploração da madeira e de diferentes intensidades de desbastes sobre a dinâmica da vegetação de uma área de 136 ha na floresta nacional do Tapajós. Dissertation. Universidade São Paulo, Brazil.
- Feldpausch, T. R., S. Jirka, C. A. M. Passos, F. Jasper, and S. J. Riha. 2005. When big trees fall: damage and carbon export by reduced impact logging in southern Amazonia. *Forest Ecology and Management* 219:199–215.
- Foley, J. A., G. P. Asner, M. H. Costa, M. T. Coe, R. DeFries, H. K. Gibbs, E. A. Howard, S. Olson, J. Patz, N. Ramankutt, and P. Snyder. 2007. Amazonia revealed: forest degradation and loss of ecosystem goods and services in the Amazon Basin. *Frontiers in Ecology* 5:2255–2277.
- Gourlet-Fleury, S., J.-M. Guehl, and O. Laroussinie, editors. 2004. Ecology and management of a neotropical forest. Lessons drawn from Paracou, a long-term experimental research site in French Guiana. Elsevier, Paris, France.
- Gullison, R. E., P. C. Frumhoff, J. G. Canadell, C. B. Field, D. C. Nepstad, K. Hayhoe, R. Avissar, L. M. Curran, P. Friedlingstein, C. D. Jones, and C. Nobre. 2007. Tropical forests and climate policy. *Science* 316:985–986.
- Holmes, T. P., G. M. Blate, J. C. Zweede, R. Pereira, P. Barreto, F. Boltz, and R. Bauch. 2002. Financial and ecological indicators of reduced impact logging performance in the eastern Amazon. *Forest Ecology and Management* 163:93–110.
- Houghton, R. A. 2005. Aboveground forest biomass and the global carbon balance. *Global Change Biology* 11:945–958.
- Jonkers, W. B. J. 1987. Vegetation structure, logging damage and silviculture in a tropical rain forest in Suriname. Backhuys Publishers, Leiden, The Netherlands.
- Keller, M., G. P. Asner, N. Silva, and M. Palace. 2004. Sustainability of selective logging of upland forests in the Brazilian Amazon: carbon budgets and remote sensing as tools for evaluating logging effects. Pages 41–63 in D. J. Zarín, J. R. R. Alavalapati, F. E. Putz, and M. Schmink, editors. *Working forests in the Neotropics*. Columbia University Press, New York, New York, USA.
- Laurance, W. F., L. V. Ferreira, J. M. Rankin de Merona, S. G. Laurance, R. W. Hutchings, and T. E. Lovejoy. 1998. Effects of forest fragmentation on recruitment patterns in Amazonian tree communities. *Conservation Biology* 12:460–464.
- Lewis, S. L., et al. 2004. Concerted changes in tropical forest structure and dynamics: evidence from 50 South American long-term plots. *Philosophical Transactions of the Royal Society B* 359:421–436.
- Lugo, A. E., and F. N. Scatena. 1996. Background and catastrophic mortality in tropical moist, wet and rain forests. *Biotropica* 28:585–599.
- Nepstad, D., S. Schwartzman, B. Bamberger, M. Santilli, D. Ray, P. Schlesinger, P. Lefebvre, A. Alencar, E. Prinz, G. Fiske, and A. Rolla. 2006. Inhibition of Amazon deforestation and fire by parks and indigenous lands. *Conservation Biology* 20:65–73.
- Nepstad, D. C., A. Verissimo, A. Alencar, C. Nobre, E. Lima, P. Lefebvre, P. Schlesinger, C. Potter, P. Moutinho, E. Mendoza, M. Cochrane, and V. Brooks. 1999. Large-scale impoverishment of Amazonian forests by logging and fire. *Nature* 398:508–508.
- Peña-Claros, M., T. S. Fredericksen, A. Alarcón, G. M. Blate, U. Choque, C. Leão, J. C. Licona, B. Mostacedo, W. Pariona, Z. Villegas, and F. E. Putz. 2008. Beyond reduced-impact logging: silvicultural treatments to increase growth rates of tropical trees. *Forest Ecology and Management* 256:1458–1467.
- Penman, J., M. Gytarsky, T. Hiraishi, T. Krug, D. Kruger, R. Pipatti, L. Buendia, K. Miwa, T. Ngara, K. Tanabe, and F. Wagner. 2003. Good practice guidance for land use, land-use change, and forestry. Institute for Global Environmental Strategies, Intergovernmental Panel on Climate Change, National Greenhouse Gas Inventories Programme, Kanagawa, Japan.
- Peres, C. A. 2006. Detecting anthropogenic disturbance in tropical forests. *Trends in Ecology and Evolution* 21:227–229.
- Phillips, O., and A. Gentry. 1994. Increasing turnover through time in tropical forests. *Science* 263:954–958.
- Pinard, M., and W. P. Cropper. 2000. A simulation model of carbon dynamics following logging of dipterocarp forest. *Journal of Applied Ecology* 37:267–283.
- Putz, F. E., P. A. Zuidema, M. A. Pinard, R. G. A. Boot, J. A. Sayer, D. Sheil, P. Sist, Elias, and J. K. Vanclay. 2008. Improved tropical forest management for carbon retention. *PLoS Biology* 6:1368–1369.

- Ramankutty, N., H. K. Gibbs, F. Achard, R. Defriess, J. A. Foley, and R. A. Houghton. 2007. Challenges to estimating carbon emissions from tropical deforestation. *Global Change Biology* 13:51–66.
- Rice, A. H., E. H. Pyle, S. R. Saleska, L. Hutya, M. Palace, M. Keller, P. B. de Camargo, K. Portilho, D. F. Marques, and S. C. Wofsy. 2004. Carbon balance and vegetation dynamics in an old-growth Amazonian forest. *Ecological Applications* 14(Supplement):S55–S71.
- Saatchi, S. S., R. A. Houghton, R. C. Dos Santos Alvala, J. V. Soares, and Y. Yu. 2007. Distribution of aboveground live biomass in the Amazon basin. *Global Change Biology* 13: 816–837.
- Silva, J. N. M., J. O. P. de Carvalho, J. do C.A. Lopes, B. F. de Almeida, D. H. M. Costa, L. C. de Oliveira, J. K. Vanclay, and J. P. Skovsgaard. 1995. Growth and yield of a tropical rainforest in the Brazilian Amazon 13 years after logging. *Forest Ecology and Management* 71:267–274.
- Sist, P., and N. Brown. 2004. Silvicultural intensification for tropical forest conservation: a response to Fredericksen and Putz. *Biodiversity and Conservation* 13:2381–2385.
- Soares, B. S., D. C. Nepstad, L. M. Curran, G. C. Cerqueira, R. A. Garcia, C. A. Ramos, E. Voll, A. McDonald, P. Lefebvre, and P. Schlesinger. 2006. Modelling conservation in the Amazon basin. *Nature* 440:520–523.
- Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K. Averyt, M. M. B. Tignor, and H. L. Miller, editors. 2007. *Climate change 2007. The physical science basis working group I contribution to the fourth assessment report of the intergovernmental panel on climate change*. Cambridge University Press, Cambridge, UK.
- Swaine, M., D. Lieberman, and F. Putz. 1987. The dynamics of tree populations in tropical forest: a review. *Journal of Tropical Ecology* 3:359–366.
- ter Steege, H., N. C. A. Pitman, O. L. Phillips, J. Chave, D. Sabatier, A. Duque, J. F. Molino, M. F. Prevost, R. Spichiger, H. Castellanos, P. von Hildebrand, and R. Vasquez. 2006. Continental-scale patterns of canopy tree composition and function across Amazonia. *Nature* 443: 444–447.
- ter Steege, H., et al. 2003. A spatial model of tree alpha-diversity and tree density for the Amazon. *Biodiversity and Conservation* 12:2255–2277.
- Vidal, E. J. 2003. *Dinâmica de florestas manejadas e sob exploração convencional na Amazônia oriental*. Dissertation. Escola de Engenharia de São Carlos/Universidade de São Paulo, São Paulo, Brazil.
- Vidal, E. J., V. M. Viana, and J. L. F. Batista. 2002. Regrowth of a tropical rain forest in Eastern Amazonia three years after planned and unplanned logging. *Scientia Forestalis* 61:133–143.
- Vieira, S., P. B. de Camargo, D. Selhorst, R. da Silva, L. Hutya, J. Q. Chambers, I. F. Brown, N. Higuchi, J. dos Santos, S. C. Wofsy, S. E. Trumbore, and L. A. Martinelli. 2004. Forest structure and carbon dynamics in Amazonian tropical rain forests. *Oecologia* 140:468–479.
- Vieira, S., S. Trumbore, P. B. Camargo, D. Selhorst, J. Q. Chambers, N. Higuchi, and L. A. Martinelli. 2005. Slow growth rates of Amazonian trees: Consequences for carbon cycling. *Proceedings of the National Academy of Sciences (USA)* 102:18502–18507.
- Zarin, D. J., M. D. Schulze, E. J. Vidal, and M. Lentini. 2007. Beyond reaping the first harvest: management objectives for timber production in the Brazilian Amazon. *Conservation Biology* 21:916–925.

APPENDIX A

Detailed site description (*Ecological Archives* A019-058-A1).

APPENDIX B

Carbon bookkeeping model (*Ecological Archives* A019-058-A2).

APPENDIX N

The successional status of tropical rainforest tree species is associated with differences in leaf carbon isotope discrimination and functional traits

Bonal, D., C. Born, C. Brechet, S. Coste, E. Marcon, J. C. Roggy et J. M. Guehl (2007). « The successional status of tropical rainforest tree species is associated with differences in leaf carbon isotope discrimination and functional traits ». In : *Annals of Forest Science* 64.2, p. 169–176.

The successional status of tropical rainforest tree species is associated with differences in leaf carbon isotope discrimination and functional traits

Damien BONAL^{a*}, Céline BORN^a, Claude BRECHET^b, Sabrina COSTE, Eric MARCON^a,
Jean-Christophe ROGGY^a, Jean-Marc GUEHL^b

^a INRA Kourou, UMR Écologie des Forêts de Guyane, BP 709, 97387 Kourou Cedex, Guyane, France

^b INRA Nancy, UMR Écologie et Écophysologie, 54280 Champenoux, France

(Received 24 April 2006; accepted 30 June 2006)

Abstract – We characterised the among species variability in leaf gas exchange and morphological traits under controlled conditions of seedlings of 22 tropical rainforest canopy species to understand the origin of the variability in leaf carbon isotope discrimination (Δ) among species with different growth and dynamic characteristics (successional gradient). Our results first suggest that these species pursue a consistent strategy in terms of Δ throughout their ontogeny (juveniles grown here versus canopy adult trees from the natural forest). Second, leaf Δ was negatively correlated with WUE and N, and positively correlated with g_s , but among species differences in Δ were mainly explained by differences in WUE. Finally, species belonging to different successional groups display distinct leaf functional and morphological traits. We confirmed that fast growing early successional species maximise carbon assimilation with high stomatal conductance. In contrast, fast and slow growing late successional species are both characterised by low carbon assimilation values, but by distinct stomatal conductance and leaf morphological features. Along the successional gradient, these differences result in much lower Δ for the intermediate species (i.e. fast growing late successional) as compared to the two other groups.

¹³C / functional diversity / leaf gas exchange / species grouping / tropical rainforest

Résumé – Le statut successional des espèces de la forêt tropicale humide est associé à des différences de discrimination isotopique du carbone et de traits fonctionnels foliaires. Nous avons caractérisé la variabilité interspécifique des échanges gazeux et des traits morphologiques foliaires en conditions environnementales contrôlées de jeunes plants de 22 espèces d'arbres de la canopée en forêt tropicale humide afin de comprendre l'origine de la variabilité de la discrimination isotopique du carbone foliaire (Δ) observée entre ces espèces présentant des caractéristiques de croissance et de dynamique distinctes (groupes successionnels). Nous montrons premièrement que les espèces tropicales possèdent une stratégie très conservée de Δ au cours de leur ontogénie (juvéniles élevés ici versus arbres adultes de la canopée en forêt naturelle). Deuxièmement, Δ était négativement corrélée à WUE et N, et positivement à g_s , mais les différences de Δ entre espèces sont principalement expliquées par des différences de WUE. Enfin, nous montrons que les espèces appartenant à des groupes successionnels distincts présentent des traits fonctionnels et morphologiques foliaires distincts. Nous confirmons que les espèces à croissance rapide qui s'installent en premier au cours de la succession écologique (FE) maximisent A avec de fortes conductances stomatiques. Les espèces climax (qui s'installent en second dans la succession écologique), à croissance rapide (FL) ou à croissance faible (SL), présentent des valeurs de A identiques, mais des valeurs de g_s ainsi que des caractéristiques morphologiques foliaires distinctes. Dans la succession écologique, ces différences se traduisent par des valeurs de Δ nettement plus faibles pour les espèces intermédiaires (c'est-à-dire les espèces climax à croissance rapide) par rapport aux deux autres groupes.

¹³C / diversité fonctionnelle / échanges gazeux foliaires / groupes successionnels / forêt tropicale humide

1. INTRODUCTION

In an attempt to simplify the complexity of the tropical rainforest ecosystem and to understand the mechanisms underlying the coexistence and distribution of the numerous tree species, forest ecologists have tried to group species according to ecological traits [3, 15, 39, 50, 55, 58]. A continuum of trait values, rather than distinct classes, is usually found and the chance to find a single clustering of species based on several ecological traits is small. However, classifications combining at least two ecological axes have been proposed [22, 24, 39, 55]. These axes can be characterised by the dynamic of the forest

(tree growth, mortality and recruitment) and the morphology of trees (height and maximal diameter of the trees) and are pertinent to understand ecosystem processes [38].

Distinct leaf functional and/or morphological traits have been found among successional groups in neotropical rainforests [3, 6, 15, 16, 30, 42, 44, 49]. Comparisons between fast growing early successional species (FE) and late successional species at a whole have been widely conducted. Higher specific leaf area (SLA), leaf dry mass based nitrogen concentration values (N) or maximum photosynthetic characteristics (A) are generally found in the former ones. In contrast, comparisons within the late successional group (fast growing late successional, FL, versus slow growing late successional, SL)

* Corresponding author: damien.bonal@kourou.cirad.fr

are more scarce. Whereas no significant differences in photosynthetic capacities among these two groups were observed in a common garden experiment [10], higher values of A and N were found in FL species (defined as “intermediate”) as compared to SL ones in an in situ experiment on saplings in BCI, Panama [15]. Furthermore, sunlit leaf carbon isotope discrimination (Δ) of adult trees was lower in FL species as compared to SL and FE species in different tree communities in French Guiana [7, 26], underlying an original non-linear distribution (modal distribution) of this trait along the successional gradient. Carbon isotope discrimination (Δ) – roughly the difference in carbon isotope composition ($\Delta^{13}\text{C}$) between the carbon source for photosynthesis (i.e. atmospheric CO_2) and the photosynthetic products (i.e. leaf material) – is a convenient measure of long-term intercellular CO_2 concentration [19–21]. It constitutes an indicator of the set point for leaf gas exchange regulation, reflecting leaf-level water-use efficiency (WUE) and overall trade-offs between carbon gain and transpirational water loss [13]. The physiological basis for these differences in Δ among successional groups has not yet been elucidated.

While screening tree communities in the tropical rainforest of French Guiana, variations in sunlit leaf Δ of up to 6.5‰ have been found within 1-ha stands [7]. This variation reflects a threefold variation in WUE among species. WUE is defined as the ratio of net carbon assimilation rate (A) to stomatal conductance for water vapour (g_s). Yet it remains unclear whether estimates of Δ -derived WUE are related to differences in A or g_s , or merely reflect differences in the trade-off between these variables [13, 21]. A large variability in such leaf gas exchange characteristics has been observed among juveniles of tropical rainforest species, with up to a four-fold range among species within the same environment [3, 6, 9, 15–17, 29, 30, 32, 37, 54]. Furthermore, based on observations on seven species growing under homogenous conditions in monospecific plantations, it has been suggested that differences in Δ among tropical rainforest tree species were mainly related to differences in g_s [26]. Species with low Δ values presented low g_s values, while A values were intermediate. We question here whether species belonging to different successional groups and differing in Δ would display different leaf gas exchange (A , g_s , WUE) and leaf morphological traits. In particular, we address the following questions:

- Are differences among species in Δ consistent over ontogenetic stages (juvenile versus adult)?
- Is the among species variability in Δ in tropical rainforests related to differences in leaf gas exchange (A , g_s , or merely the ratio $A/g_s = \text{WUE}$) and with other related leaf traits?
- Are species belonging to different successional groups (FE, FL, SL) characterised by different leaf morphological and/or functional characteristics, particularly Δ , under common environmental conditions?

2. MATERIALS AND METHODS

2.1. Plant material

In situ comparative leaf gas exchange measurements on numerous adult trees under similar environmental conditions are not fea-

sible. Then, in this study, potted seedlings of 22 abundant canopy tree species of the rainforest of French Guiana were grown in a glasshouse under common environmental conditions. The 22 focal species represented a broad range of in situ leaf Δ ([7], and Bonal, unpublished data) and were classified into three successional groups with regard to growth and dynamic characteristics according to Favrichon [22, 23]. This grouping is based on a statistical analysis of tree dynamic (growth, mortality and recruitment) and morphological and dendrometric (height and maximal diameter of the trees) variables at the adult and sub-adult stage and a PCA analysis leading to two axes related to potential size and heliophily. The three groups are fast growing early successional (FE), fast growing late successional (FL), and slow growing late successional (SL), with mean canopy tree annual diameter increments equal to 0.27 ± 0.04 , 0.24 ± 0.03 , and $0.13 \pm 0.01 \text{ cm year}^{-1}$, respectively [22, 23]. SL species regenerate and develop in the understory and retain low diameter growth rates even once installed in the canopy. In contrast, FL species display higher growth rates in the large diameter classes than in the small diameter ones. For species not included in these analyses, the successional group was assigned based on Béna [4]. No such information was found for *Eriotheca*.

Fifty seeds per species were collected from within a 10-m radius of each of at least five adult trees per species in the Paracou forest, French Guiana ($5^\circ 16' \text{ N}$, $52^\circ 55' \text{ W}$) in spring 2000, except for *Cecropia* and *Talisia* for which seeds were collected in the surroundings of Kourou, French Guiana. Randomly selected germinated seedlings (15–20 per species) were kept for 18 months in 5.3-L black polyethylene containers filled with a homogenised forest soil in a shaded understory site (about 16% full sun). In September 2001, 10–12 seedlings per species were transplanted into 20-L plastic pots and randomly distributed in a glasshouse in Kourou, French Guiana. One layer of neutral shade-cloth was used to reduce light levels to about 25% of full sun (maximum PAR $\approx 500 \mu\text{mol m}^{-2} \text{ s}^{-1}$). Pots were filled with a mixture of sand (30%) and an A-horizon soil (70%) from the Paracou forest. Plantlets were grown for eight months, during which they were watered every two days in order to match the amount of water observed to have been lost through evapotranspiration and then to maintain plants at field capacity ($\approx 0.25 \text{ m}^3 \text{ m}^{-3}$). Soil water content in the pots was recorded monthly using a TDR Trime FM2 (Imko, Ettlingen, Germany). Pots were fertilised every second month (5 g complete fertiliser per pot, 12/12/17/2 N/P/K/Mg) and treated with a commercial insecticide (Cuberol: 5% rotenone).

2.2. Leaf gas exchange, leaf and plant traits

In April 2002, leaf gas exchange measurements were conducted on 7–11 plants per species using a portable photosynthesis system (CIRAS1, PP-Systems, Hoddesdon, UK) operating in open mode and fitted with a Parkinson leaf cuvette. Gas exchange measurements were conducted on two leaves per plant under the following non-limiting environmental conditions [10]: $[\text{CO}_2] = 360 \text{ ppm}$; PAR = $670 \pm 20 \mu\text{mol m}^{-2} \text{ s}^{-1}$; vapour pressure deficit = $1.2 \pm 0.4 \text{ kPa}$; air temperature = $30.0 \pm 2.1^\circ \text{C}$. Equations of Caemmerer and Farquhar [8] were used to calculate net carbon assimilation rate on a leaf area (A , $\mu\text{mol m}^{-2} \text{ s}^{-1}$) or mass (A_m , $\text{mmol g}^{-1} \text{ s}^{-2}$) basis, stomatal conductance for water vapour (g_s , $\text{mol m}^{-2} \text{ s}^{-1}$) and intrinsic water-use efficiency ($\text{WUE} = A/g_s$). After the gas exchange measurements, 10 to 15 mature and fully expanded leaves per plant were collected in order to characterise specific leaf area (SLA, $\text{cm}^2 \text{ g}^{-1}$) and leaf thickness (LT, μm) and calculate leaf density ($\text{LD} = 1 / (\text{SLA}$

\times LT), g cm^{-3}). The leaves were then dried for 48 h at 70.0 °C and finely ground.

2.3. Carbon, nitrogen and isotope analyses

A sub-sample of 10^{-3} g of dry leaf powder was analysed for total carbon (C, %) and nitrogen (N, mg g^{-1}) concentration (ThermoQuest-NA-1500-NCS, Carlo Erba, Italy) at the stable isotope facility of INRA Nancy, France. Leaf carbon isotope composition ($\delta^{13}\text{C}$) for each plant was estimated on the same sub-sample using an isotope ratio mass spectrometer (Delta-S Finnigan Mat, Bremen, Germany). Glasshouse air carbon isotope composition ($\delta_a = -7.85\text{‰}$) was estimated using leaf carbon isotope composition ($\delta^{13}\text{C}$) of corn grown for four months during the acclimation phase in the glasshouse [35]. Leaf carbon isotope discrimination (Δ) was calculated as:

$$\Delta = \frac{\delta_a - \delta^{13}\text{C}}{1 + 0.001 \times \delta^{13}\text{C}} \quad (1)$$

Δ is theoretically related with WUE ($= A/g_s$) through the following equation [19]:

$$\Delta = b - \frac{1.6 \times (b - a) \times \text{WUE}}{C_a} \quad (2)$$

where C_a is the CO_2 concentration in air, and a and b are fractionation factors during the diffusion of CO_2 through the stomata and during photosynthetic carboxylation, respectively, that are assumed here to be constant for all species.

2.4. Ontogenetic test

To compare leaf Δ of seedlings grown in the glasshouse (original data of this study) with that of sunlit leaves of mature canopy trees, we compiled Δ data from two 1-ha stands (Saint-Elie [7], and Bafog, Bonal, unpublished data) in French Guiana. These two stands comprised 3–9 mature and dominant trees characterised for leaf Δ for 14 species out of the 22 studied species in the glasshouse (Tab. I). For more details on data collecting protocol, see Bonal et al. [7].

2.5. Statistical analyses

Statistical analyses were performed using SAS programs (SAS-Institute, Cary, USA). Differences among species for the different parameters were tested using ANOVA in the general linear regression model (GLM) procedure. Correlations between the different parameters were tested using Pearson's correlation coefficients. The relationship between Δ and WUE was tested using a general linear regression model. As both sets of juvenile and adult Δ data were subjected to measurements errors, a model II regression analysis was performed to test for any relationship between glasshouse and forest values among the 14 species. Statistical differences among successional groups were tested using an ANOVA and post-hoc Duncan's multiple range tests.

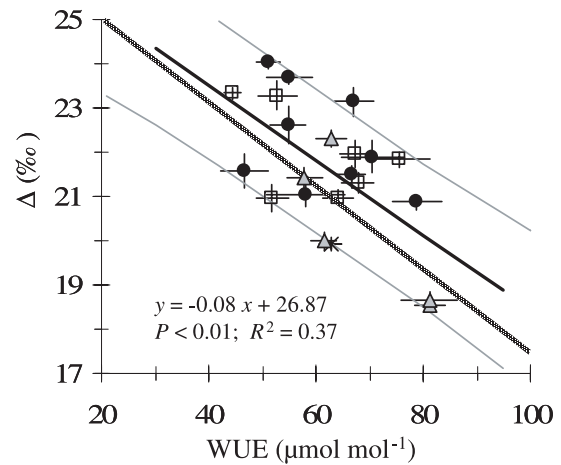


Figure 1. The relationship between leaf carbon isotope discrimination (Δ) and intrinsic water-use efficiency ($\text{WUE} = A/g_s$) for potted seedlings of 22 tropical rainforest species grown in non-limiting environmental conditions in a glasshouse ($n = 7$ to 11 plants per species). The theoretical relationship is represented by the dotted line ($\Delta = -0.095 \times \text{WUE} + 27.00$) [21]. Vertical and horizontal bars denote ± 1 standard error of the species mean. The grey lines represent the upper and lower confidence interval limits (95%) of the relationship between Δ and WUE. The Symbols correspond to the successional groups: white squares correspond to fast growing early successional species (FE); grey triangles correspond to fast growing late successional species (FL); black dots correspond to slow growing late successional species (SL); the star corresponds to an undetermined species.

3. RESULTS

Mean species Δ varied from 18.6 to 24.0‰, with low within species variability (Fig. 1). WUE, SLA and N varied over a twofold range (Fig. 1, Tab. I). A , g_s and LD varied over a threefold range.

For the significant statistical correlations obtained among measured parameters (Tab. II), we underline that A or A_m were significantly correlated with all parameters except Δ and WUE. Δ was negatively correlated with WUE and g_s and positively with N. WUE and g_s were strongly negatively correlated.

The relationship between Δ and WUE was highly significant (Fig. 1). This relationship was not statistically different from the one described by the theory since the slope and the y -intercept of the theoretical line (-0.095 and 27.00 , respectively) fell in the confidence interval (95%) of the slope (-0.084 ± 0.024) and the y -intercept (26.87 ± 1.56) of the relationship between Δ and WUE.

There was a strong positive and linear relationship between Δ of sunlit leaves of dominant canopy trees and the leaves of potted seedlings of the same species grown in the glasshouse ($P < 0.01$) (Fig. 2).

There was a significant group effect on all studied traits (Fig. 3) except LD ($P = 0.23$). FE species had highest values in A_m , A , g_s , N and SLA. There was no difference in A_m or A

Table I. Mean (\pm SE, $n = 7 - 11$) of specific leaf area (SLA, $\text{cm}^2 \text{g}^{-1}$), leaf density (LD, g cm^{-3}), leaf carbon concentration (C, %), and nitrogen concentration on a dry mass basis (N, mg g^{-1}) of the 22 studied species. Species are sorted according to successional groups (FE = fast growing early successional, FL = fast growing late successional, SL = slow growing late successional, Und. = Undetermined). * Symbol indicates that this species was included in the juvenile versus adult comparison. Groups' effect was tested using an ANOVA and post-hoc Duncan's multiple range test ($P < 0.05$). Species named according to Boggan et al. [5].

Species name	Taxonomic family	Group	SLA $\text{cm}^2 \text{g}^{-1}$	LD g cm^{-3}	C %	N mg g^{-1}
<i>Bagassa guianensis</i> J.B. Aublet	Moraceae	FE	255.6 \pm 48.5	0.18 \pm 0.02	46.3 \pm 1.8	24.9 \pm 1.7
* <i>Carapa procera</i> A.P. De Candolle	Meliaceae	FE	158.6 \pm 6.6	0.30 \pm 0.01	48.2 \pm 0.2	17.9 \pm 0.6
* <i>Cecropia obtusa</i> Trécul	Moraceae	FE	215.1 \pm 9.6	0.19 \pm 0.01	47.2 \pm 0.4	21.1 \pm 1.5
<i>Hymenaea courbaril</i> Linnaeus	Caesalpiniaceae	FE	199.8 \pm 32.0	0.34 \pm 0.04	48.6 \pm 0.6	24.8 \pm 0.9
<i>Tabebuia insignis</i> Sandwith	Bignoniaceae	FE	138.9 \pm 5.9	0.31 \pm 0.02	48.4 \pm 0.3	21.4 \pm 1.3
* <i>Virola michelii</i> Heckel	Myristicaceae	FE	129.1 \pm 5.5	0.45 \pm 0.01	49.8 \pm 0.3	18.5 \pm 0.6
<i>Virola surinamensis</i> Warburg	Myristicaceae	FE	142.9 \pm 5.5	0.33 \pm 0.02	51.6 \pm 0.3	23.1 \pm 0.5
* <i>Dicorynia guianensis</i> G.J. Amshoff	Caesalpiniaceae	FL	188.5 \pm 10.4	0.27 \pm 0.02	51.1 \pm 0.3	19.8 \pm 0.9
* <i>Eperua falcata</i> J.B. Aublet	Caesalpiniaceae	FL	140.6 \pm 5.0	0.43 \pm 0.00	50.9 \pm 0.4	23.7 \pm 0.9
* <i>Eperua grandiflora</i> Benth	Caesalpiniaceae	FL	117.0 \pm 13.2	0.44 \pm 0.00	49.6 \pm 0.3	20.3 \pm 1.2
* <i>Sextonia rubra</i>	Lauraceae	FL	107.1 \pm 6.2	0.36 \pm 0.01	50.1 \pm 0.4	11.5 \pm 0.6
<i>Talisia furfuracea</i> Sandwith	Sapindaceae	FL	208.6 \pm 9.3	0.34 \pm 0.04	48.4 \pm 0.3	21.8 \pm 0.7
<i>Amanoa guianensis</i> J.B. Aublet	Euphorbiaceae	SL	112.6 \pm 6.0	0.35 \pm 0.02	47.4 \pm 1.2	15.0 \pm 1.1
* <i>Aspidosperma album</i> R. Benoist	Apocynaceae	SL	142.9 \pm 7.7	0.31 \pm 0.02	49.0 \pm 0.8	14.4 \pm 1.4
* <i>Eschweilera sagotiana</i> Miers	Lecythidaceae	SL	101.0 \pm 5.1	0.36 \pm 0.03	49.3 \pm 0.3	15.6 \pm 0.9
* <i>Lecythis persistens</i> Sagot	Lecythidaceae	SL	107.0 \pm 7.5	0.37 \pm 0.03	50.9 \pm 0.2	15.7 \pm 0.8
* <i>Licania heteromorpha</i> Benth	Chrysobalanaceae	SL	102.0 \pm 7.5	0.41 \pm 0.03	48.5 \pm 0.7	10.4 \pm 0.8
* <i>Manilkara bidentata</i> A.J. Chevalier	Sapotaceae	SL	94.9 \pm 8.5	0.32 \pm 0.02	47.6 \pm 1.2	11.7 \pm 0.9
<i>Poraqueiba guianensis</i> J.B. Aublet	Icacinaceae	SL	100.4 \pm 5.5	0.37 \pm 0.02	49.2 \pm 0.3	10.7 \pm 0.5
* <i>Protium subseratum</i> Engler	Burseraceae	SL	154.8 \pm 10.9	0.35 \pm 0.03	47.5 \pm 0.2	14.4 \pm 0.5
* <i>Vouacapoua americana</i> J.B. Aublet	Caesalpiniaceae	SL	161.8 \pm 4.4	0.45 \pm 0.01	51.5 \pm 0.4	18.6 \pm 0.6
<i>Eriotheca</i> sp.	Bombacaceae	Und.	98.2 \pm 5.9	0.25 \pm 0.01	46.0 \pm 0.3	15.9 \pm 0.7
ANOVA-Test for group effect			$P < 0.001$	$P = 0.230$	$P < 0.001$	$P < 0.001$

between FL and SL species. In contrast, N, C and SLA were significantly higher in FL species as compared to FE ones, and g_s was lower in FL species as compared to FE ones. FL species displayed significantly higher WUE and lower Δ as compared to FE and SL species, which did not differ.

4. DISCUSSION

4.1. Functional diversity

This study confirmed the high variability of leaf morphological [46, 56] and functional [6, 15, 17, 29, 30, 32, 46] traits among tropical rainforest species. In common environmental conditions, we observed differences on a two-to-three-fold range among species in leaf gas exchange, WUE, and leaf morphological characteristics, whereas the within species variability of these traits under homogenous conditions remains low (Tab. I, Fig. 1).

The relationship between time-integrated Δ and instantaneous WUE at the species level ($\Delta = -0.084 \times \text{WUE} + 26.87$; $P < 0.01$; $R^2 = 0.37$) was in agreement with the two-step carbon isotope discrimination model [19] ($\Delta = -0.095 \times \text{WUE} + 27.00$) (Fig. 1). The relationship for some species slightly deviated from the theoretical line. As already discussed by other authors [19, 41], the difference in time-integration scale between the two variables could be one major cause for this discrepancy. Whereas WUE was derived from spot measurements of instantaneous leaf gas exchange, Δ was estimated based on leaf carbon isotope composition that are integrated over the lifetime of the analysed leaf tissue (9 to 24 months). Other causes related to differences in morphology and structure of the leaves of these species (leaf density, specific leaf area, Tab. I) must be considered as well [20, 28, 40]. These differences could lead to distinct values of internal conductance to diffusion of CO_2 from the intercellular air spaces to the sites of carboxylation and then to internal isotopic fractionation factors differing among species [18]. Furthermore, recent papers

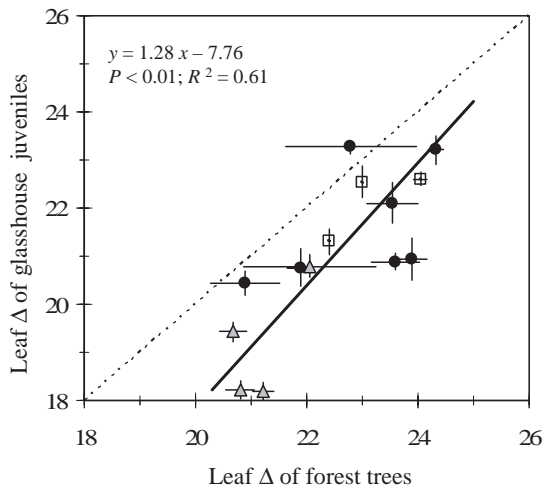


Figure 2. The relationship (Model II regression) between mean species leaf Δ values ($n = 14$) of seedlings grown in the glasshouse and of canopy trees from two forest stands in French Guiana. For the seedlings, species mean is based on 7–11 plants. For the adult trees, species mean is based on 3–9 trees. Vertical and horizontal bars denote ± 1 standard error of the mean. Symbols correspond to the successional groups: white squares correspond to fast growing early successional species (FE); grey triangles correspond to fast growing late successional species (FL); black dots correspond to slow growing late successional species (SL).

Table II. Pearson correlation coefficients among morphological and functional parameters for the 22 studied species: leaf-area based rates of net carbon assimilation (A , $\mu\text{mol m}^{-2} \text{s}^{-1}$), leaf-mass based rates of net carbon assimilation (A_m , $\mu\text{mol s}^{-1} \text{g}^{-1}$), stomatal conductance for water vapour (g_s , $\text{mol m}^{-2} \text{s}^{-1}$), intrinsic water-use efficiency ($\text{WUE} = A/g_s$), leaf carbon isotope discrimination during photosynthesis (Δ), leaf carbon concentration (C , %), leaf nitrogen concentration on a dry mass basis (N , mg g^{-1}), specific leaf area (SLA , $\text{cm}^2 \text{g}^{-1}$), and leaf density (LD , g cm^{-3}). Numbers in bold indicate significant correlation at $P = 0.05$.

	A_m	g_s	WUE	Δ	C	N	SLA	LD
A	0.63	0.79	0.07	-0.03	-0.23	0.36	0.23	-0.41
A_m		0.62	-0.17	0.23	-0.24	0.49	0.79	-0.58
g_s			-0.49	0.24	-0.29	0.11	0.27	-0.47
WUE				-0.51	0.19	0.37	-0.10	0.29
Δ					-0.06	-0.31	0.11	0.04
C						0.12	-0.07	0.37
N							0.51	-0.05
SLA								-0.53

pointed to variable isotopic fractionations during dark respiration (e.g. [51]) or post-photosynthetic fractionations [1]. These fractionations, that are not included in the simple model of discrimination described in [19] and used here, could differ among species and contribute to the shift with the theoretical line.

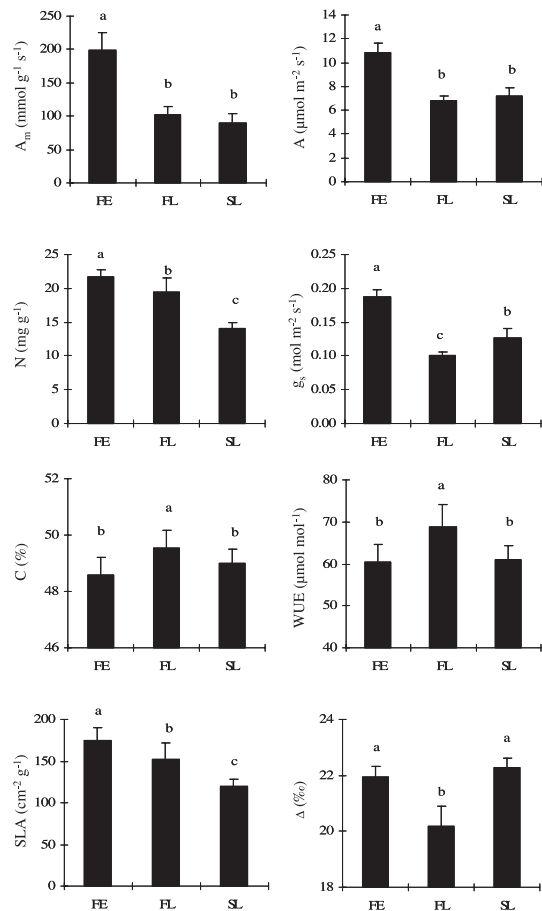


Figure 3. The mean values (± 1 SE) of mass-based carbon net assimilation (A_m), leaf area-based carbon net assimilation (A), leaf nitrogen concentration on a dry mass basis (N), stomatal conductance for water vapour (g_s), leaf carbon concentration (C), intrinsic water-use efficiency ($\text{WUE} = A/g_s$), specific leaf area (SLA), and leaf carbon isotope discrimination (Δ) for each successional groups (FE = fast growing early successional, FL = fast growing late successional, SL = slow growing late successional) for potted seedlings of 22 species of the rainforest of French Guiana grown in a glasshouse under non-limiting environmental conditions. For a given trait, means followed by the same letter are not significantly different at $P = 0.05$ (Duncan's multiple range test).

When grown in pots under homogeneous environmental conditions, the 22 focal species almost covered the entire gradient in Δ observed in 1-ha stands of natural forest for mature trees (5.4) [7, 26]. Furthermore, a significant correlation between Δ of 2-year old seedlings and canopy trees was found ($\Delta_{\text{juvenile}} = 1.28 \times \Delta_{\text{adult}} - 7.76$; $P < 0.01$; $R^2 = 0.61$, Fig. 2) and then an overall consistent species ranking. Several studies demonstrated ontogenetic shifts for functional traits such as biomass allocation, SLA or A for tropical [36, 43, 53] or temperate (e.g. [11]) species, whereas no such shift was observed for hydraulic traits for Patagonian conifers [27]. Our results support the statement that among species differences in

the metabolic intrinsic set point for the trade-off between carbon and water fluxes at the leaf level, as reflected by Δ [13], are maintained over the lifetime of these species. Except for one species, leaf Δ of glasshouse seedlings were lower (1.5‰ on average) than values of adult trees (Fig. 2), which can be interpreted as a higher WUE for juveniles as compared to trees (see Eq. (1)). These differences can be mainly explained by the differences in environmental conditions (i.e. light, air temperature and humidity, CO₂ concentration) encountered by the two considered sets of values. These conditions are recognised to strongly influence Δ (see synthesis in [14]). Furthermore, isotope fractionations come into play during carbohydrate storage and translocation, or when leaf tissues are partially constructed from stored carbohydrate reserves, as for adult trees [52].

4.2. Association of traits

Previous studies on trade-offs among leaf traits concentrated mainly on traits related to carbon assimilation, leaf structure or chemical concentrations [45, 46, 56, 57]. Leaf carbon traits of tropical rainforest seedlings studied here were in agreement with general trade-offs since carbon assimilation increases with increasing N and SLA, and decreases with increasing C and LD (Tab. II). It has been suggested that including additional traits related to water relations (e.g. leaf conductance to water vapour) in these approaches may improve our understanding of plant function [48, 57]. We present here evidence that other associations of traits at the leaf level do exist among tropical rainforest species with regard to carbon isotope discrimination and stomatal regulation (Tab. II). Higher Δ are associated with lower WUE and g_s , and higher N. Furthermore, higher A or A_m are strongly associated with g_s . These data suggest that biophysical and natural selection constraints result in the exclusion of some combinations between transpirational fluxes and leaf structure or carbon concentrations. Including these functional traits in larger meta-analyses on leaf economics may therefore allow improvements of the interpretation of functional differences among tropical rainforest tree species.

In order to understand the origin of the strong variability in Δ among species, we compared the leaf gas exchange and morphological traits of these species. Differences among species in Δ were not accounted for by differences in photosynthetic characteristics, but rather by differences in g_s and WUE (Tab. II, Fig. 1), in agreement with previous observations on tropical rainforest species [6, 26]. However, although the relationship between Δ and g_s was statistically significant (Tab. II), g_s explained only a very small part of the variability in Δ among species ($R^2 = 0.08$). In contrast, WUE explained 37% of this variability (Fig. 1). Therefore, we suggest that it is the compromise between carbon fluxes and stomatal conductance for water vapour, i.e. WUE, rather than A or g_s , that is the main determinant for the differences among tropical rainforest species in Δ (Fig. 1). High Δ in some species are associated with low WUE, and vice versa. Such comparative studies on a large number of tropical rainforest species

have not yet been conducted. Nevertheless, correlations between Δ and A or g_s have been observed in other forested ecosystems, with contrasting patterns. Differences in Δ among evergreen species from Japanese warm-temperate forests were not caused by the variation in g_s , but mainly by the difference in long-term photosynthetic capacity [28]. The variability in WUE among Caribbean hybrid Pine clones was also mainly attributed to differences in photosynthetic capacity rather than in stomatal conductance [59].

4.3. Differences among groups

Our results pointed to significant differences in Δ , leaf gas exchange, and leaf characteristics among seedlings of species belonging to different successional groups (Fig. 3) that are consistent with other studies (e.g. [3, 15, 30, 32]).

Fast growing early successional species (FS) tend to maximise carbon assimilation rates (highest A) with stomata widely open (highest g_s) for high transpiration rates. This strategy is associated with higher SLA, N, A_m , A or g_s values than those of the late successional species, and high Δ (21.9) and low WUE (60.5 $\mu\text{mol mol}^{-1}$). Furthermore, we present evidence for the existence of two distinct groups within the late successional species (fast growing versus slow growing) with regard to leaf functional and morphological traits, supporting previous studies under natural conditions [15]. The two groups are characterised by non-significantly different carbon assimilation rates (A_m or A), but FL species have lower g_s and higher SLA, C and N than SL species (Fig. 3). Furthermore, the strategy of the FL species in the trade-off between A and g_s seems to be associated with much lower Δ (20.2‰) values than SL ones (22.3‰). The observed differences among these groups in Δ seem to be environmentally stable since our results for seedlings grown in a glasshouse are in agreement with previous studies on sunlit leaves of forest trees under natural conditions [7, 26]. These functional differences seem to be consistent with the suggested adaptations of tropical rainforest species to the numerous and complex ecological niches related to light and water acquisition [47].

In contrast with most leaf functional traits, we demonstrated here the absence of a linear distribution in Δ , WUE and C along the successional gradient (Fig. 3). Despite large differences in leaf gas exchange and morphological characteristics, FE and SL species display non-significantly different Δ , WUE and C values, whereas species with the lowest Δ (highest WUE) have an intermediate position along the successional gradient. The lowest Δ (i.e. highest WUE and low g_s) observed for these species might be associated with a better ability to compete under high water vapour pressure deficit [2, 34] such as in the upper canopy layer, and/or under low soil water conditions [12, 31, 33, 41] such as during the seasonal dry season encountered in French Guiana [25]. Whether this peculiar 'V shaped' pattern of relationship between Δ and the position within the successional gradient of species holds in other forest biomes is a worthy question.

Acknowledgements: The authors wish to express their deep thanks to Jean-Yves Goret and Elli Lentilus, INRA Kourou, for their precious contribution to this experiment, to Chris Baraloto who provided the seedlings, and Tancrede Almeras for his help in the data analysis. We also thank two anonymous reviewers who greatly enhanced the first manuscript.

REFERENCES

- [1] Badeck F. W., Tcherkez G., Nogue S., Piel C., Ghashghaie J., Post-photosynthetic fractionation of stable carbon isotopes between plant organs – a widespread phenomenon, *Rapid. Commun. Mass Spectrom.* 19 (2005) 1381–1391.
- [2] Barbour M.M., Farquhar G.D., Relative humidity- and ABA-induced variation in carbon and oxygen isotope ratios of cotton leaves, *Plant Cell Environ.* 23 (2000) 473–485.
- [3] Bazzaz F.A., Pickett S.T.A., Physiological ecology of a tropical succession: a comparative review, *Annu. Rev. Ecol. Syst.* 11 (1980) 287–310.
- [4] Béna P., *Essences forestières de Guyane*, Imprimerie Nationale, 1960.
- [5] Boggan J., Funk V., Kelloff C., Hoff M., Cremers G., Feuillet C., Checklist of the plants of the Guianas: Guiana, Surinam, French Guiana, Georgetown, Guyana, 1997.
- [6] Bonal D., Barigah T.S., Granier A., Guehl J.-M., Late stage canopy tree species with extremely low $\delta^{13}\text{C}$ and high stomatal sensitivity to seasonal soil drought in the tropical rainforest of French Guiana, *Plant Cell Environ.* 23 (2000) 445–459.
- [7] Bonal D., Sabatier D., Montpied P., Tremeaux D., Guehl J.-M., Interspecific variability of $\delta^{13}\text{C}$ among canopy trees in rainforests of French Guiana: Functional groups and canopy integration, *Oecologia* 124 (2000) 454–468.
- [8] Caemmerer S. von, Farquhar G.D., Some relationships between the biochemistry of photosynthesis and the gas exchange rates of leaves, *Planta* 153 (1981) 376–387.
- [9] Cao K.F., Water relations and gas exchange of tropical saplings during a prolonged drought in a Bornean heath forest, with reference to root architecture, *J. Trop. Ecol.* 16 (2000) 101–116.
- [10] Coste S., Roggy J.-C., Imbert P., Born C., Bonal D., Dreyer E., Leaf photosynthetic traits of 14 tropical rain forest species in relation to leaf nitrogen concentration and shade tolerance, *Tree Physiol.* 25 (2005) 1127–1137.
- [11] Covone F., Gratani L., Age-related physiological and structural traits of chestnut coppices at the Castelli Romani park (Italy), *Ann. For. Sci.* 63 (2006) 239–247.
- [12] Cowan I.R., Economics of carbon fixation in higher plants, in: Givnish T.J. (Ed.), *On the economy of plant form and function*, Cambridge University Press, Cambridge, 1986, pp. 133–170.
- [13] Ehleringer J.R., Gas exchange implications of isotopic variation in arid-land plants, in: Smith J.A.C., Griffiths H. (Eds.), *Water deficits. plant responses from cell to community*, Environmental Plant Biology Series, Lancaster, UK, 1993, pp. 265–284.
- [14] Ehleringer J.R., Hall A.E., Farquhar G.D., *Stable isotopes and plant-carbon water relations*, Academic Press Inc., 1993.
- [15] Ellis A.R., Hubbell S.P., Potvin C., In situ field measurements of photosynthetic rates of tropical tree species: a test of the functional group hypothesis, *Can. J. Bot.* 78 (2000) 1336–1347.
- [16] Ellsworth D.S., Reich P.B., Photosynthesis and leaf nitrogen in five Amazonian tree species during early secondary succession, *Ecology* 77 (1996) 581–594.
- [17] Engelbrecht B.M.J., Wright S.J., De Steven D., Survival and eco-physiology of tree seedlings during El Niño drought in a tropical moist forest in Panama, *J. Trop. Ecol.* 18 (2002) 569–579.
- [18] Evans J.R., Sharkey T.D., Berry J.A., Farquhar G.D., Carbon isotope discrimination measured concurrently with gas exchange to investigate CO_2 diffusion in leaves of higher plants, *Aust. J. Plant Physiol.* 13 (1986) 281–292.
- [19] Farquhar G.D., O’Leary M.H., Berry J.A., On the relationship between carbon isotope discrimination and the intercellular carbon dioxide concentration in leaves, *Aust. J. Plant Physiol.* 9 (1982) 121–137.
- [20] Farquhar G.D., Ehleringer J.R., Hubick K.T., Carbon isotope discrimination and photosynthesis, *Annu. Rev. Plant Physiol.* 40 (1989) 503–537.
- [21] Farquhar G.D., Hubick K.T., Condon A.G., Richards R.A., Carbon isotope fractionation and plant water-use efficiency, in: Rundel P.W., Ehleringer J.R., Nagy K.A. (Eds.), *Stable isotopes in ecological research*, Springer-Verlag, New York, 1989, pp. 21–40.
- [22] Favrichon V., Classification des espèces arborées en groupes fonctionnels en vue de la réalisation d’un modèle de dynamique de peuplement en forêt Guyanaise, *Rev. Ecol. (Terre Vie)* 49 (1994) 379–403.
- [23] Favrichon V., Modèle matriciel déterministe en temps discret : application à l’étude de la dynamique d’un peuplement forestier tropical humide (Guyane française), Ph.D. thesis, Université Claude Bernard-Lyon I, 1995.
- [24] Favrichon V., Apports d’un modèle démographique plurispécifique pour l’étude des relations diversité / dynamique en forêt tropicale guyanaise, *Ann. For. Sci.* 55 (1998) 655–669.
- [25] Guehl J.-M., Dynamique de l’eau dans le sol en forêt tropicale humide guyanaise. Influence de la couverture pédologique, *Annu. For. Sci.* 41 (1984) 195–236.
- [26] Guehl J.-M., Bonal D., Ferhi A., Barigah T.S., Farquhar G.D., Granier A., Community-level diversity of carbon-water relations in rainforest trees, in: Gourlet-Fleury S., Laroussini O., Guehl J.-M. (Eds.), *Ecology and management of a neotropical rainforest*, Paracou (French Guiana), Elsevier, Paris, 2004, pp. 65–84.
- [27] Gyenge J.E., Fernández M.E., Salda G.D., Schlichter T., Leaf and whole-plant water relations of the Patagonian conifer *Austrocedrus chilensis* (D. Don) Pic. Ser. et Bizzarri: implications on its drought resistance capacity, *Ann. For. Sci.* 62 (2005) 297–302.
- [28] Hanba Y.T., Wada E., Osaki M., Nakamura T., Growth and $\delta^{13}\text{C}$ responses to increasing atmospheric carbon dioxide concentrations for several crop species, *Isotopes Environ. Health Stud.* 32 (1996) 41–54.
- [29] Hogan K.P., Smith A.P., Samaniego M., Gas exchange in six tropical semi-deciduous forest canopy tree species during the wet and dry seasons, *Biotropica* 27 (1995) 324–333.
- [30] Huc R., Ferhi A., Guehl J.-M., Pioneer and late stage tropical rainforest tree species (French Guyana) growing under common conditions differ in leaf gas exchange regulation, carbon isotope discrimination and leaf water potential, *Oecologia* 99 (1994) 297–305.
- [31] Jones H.G., Drought tolerance and water-use efficiency, in: Smith J.A.C., Griffiths H. (Eds.), *Water deficits. plant responses from cell to community*, Environmental Plant Biology Series, Lancaster, UK, 1993, pp. 193–203.
- [32] Kitajima K., Relative importance of photosynthetic traits and allocation patterns as correlates of seedling shade tolerance of 13 tropical trees, *Oecologia* 98 (1994) 419–428.
- [33] Le Roux D., Stock W.D., Bond W.J., Maphanga D., Dry mass allocation, water use efficiency and $\delta^{13}\text{C}$ in clones of *Eucalyptus grandis*, *E. grandis* × *camaldulensis* and *E. grandis* × *nitens* grown under two irrigation regimes, *Tree Physiol.* 16 (1996) 497–502.
- [34] Madhavan S., Treichel I.W., O’Leary M.H., Effects of relative humidity on carbon isotope fractionation in plants, *Bot. Acta* 104 (1991) 292–294.
- [35] Marino B.D., Mc Elroy M.B., Isotopic composition of atmospheric CO_2 inferred from carbon in C_4 plant cellulose, *Nature* 349 (1991) 127–131.

- [36] McConnaughay K.D.M., Coleman J.S., Biomass allocation in plants: ontogeny or optimality? A test along three resource gradients, *Ecology* 80 (1999) 2581–2593.
- [37] Meinzer F.C., Goldstein G., Holbrook N.M., Jackson P., Cavelier J., Stomatal and environmental control of transpiration in a lowland tropical forest tree, *Plant Cell Environ.* 16 (1993) 429–436.
- [38] Naeem S., Wright J.P., Disentangling biodiversity effects on ecosystem functioning: deriving solutions to a seemingly insurmountable problem, *Ecol. Lett.* 6 (2003) 567–579.
- [39] Oldeman R.A.A., van Dijk J., Diagnosis of the temperament of tropical rain forest trees. Rain forest regeneration and management, in: Gomez-Pompa A., Whitmore T.C., Hadley M. (Eds.), *Rain forest regeneration and management*, Unesco, Paris, 1991, pp. 21–65.
- [40] Parkhurst D.F., Diffusion of CO₂ and other gases inside leaves, *New Phytol.* 126 (1994) 449–479.
- [41] Picon C., Guehl J.-M., Ferhi A., Leaf gas exchange and carbon isotope discrimination responses to drought in a drought-avoiding (*Pinus pinaster*) and a drought-tolerant (*Quercus petraea*) species under present and elevated atmospheric CO₂ concentration, *Plant Cell Environ.* 19 (1996) 182–190.
- [42] Poorter L., Kwant R., Hernández R., Medina E., Werger M.J.A., Leaf optical properties in Venezuelan cloud forest trees, *Tree Physiol.* 20 (2000) 519–526.
- [43] Reich A., Holbrook N.M., Ewel J.J., Developmental and physiological correlates of leaf size in *Hyeronima alchorneoides* (Euphorbiaceae), *Am. J. Bot.* 91 (2004) 582–589.
- [44] Reich P.B., Walters M.B., Ellsworth D.S., Uhl C., Photosynthesis-nitrogen relations in Amazonian tree species. I: Patterns among species and communities, *Oecologia* 97 (1994) 62–72.
- [45] Reich P.B., Walters M.B., Ellsworth D.S., From tropics to tundra: Global convergence in plant functioning, *Proc. Nat. Ac. Sci.* 94 (1997) 13730–13734.
- [46] Reich P.B., Ellsworth D.S., Walters M.B., Vose J.M., Gresham C., Volin J.C., Bowman W.D., Generality of leaf trait relationships: a test across six biomes, *Ecology* 80 (1999) 1955–1969.
- [47] Ricklefs R.E., Environmental heterogeneity and plant species diversity: a hypothesis, *Amer. Nat.* 111 (1977) 377–381.
- [48] Shipley B., Lechowicz M.J., Wright I., Reich P.B., Fundamental tradeoffs generating the worldwide leaf economics spectrum, *Ecology* 87 (2006) 535–541.
- [49] Strauss-Debenedetti S., Bazzaz F.A., Photosynthetic characteristics of tropical trees along successional gradients, in: Mulkey S.S., Chazdon R.L., Smith P.A. (Eds.), *Tropical Forest Plant Ecophysiology*, Chapman and Hall, New York, 1996, pp. 162–186.
- [50] Swaine M.D., Whitmore T.C., On the definition of ecological species groups in tropical rain forests, *Vegetatio* 75 (1988) 81–86.
- [51] Tcherkez G., Nogués S., Bleton J., Cornic G., Badeck F., Ghashghaie J., Metabolic origin of carbon isotope composition of leaf dark-respired CO₂ in French bean, *Plant Physiol.* 131 (2003) 237–244.
- [52] Terwilliger V.J., Kitajima K., Le Roux-Swarthout D.J., Mulkey S.S., Wright S.J., Intrinsic water-use efficiency and heterotrophic investment in tropical leaf growth of two neotropical pioneer tree species as estimated from delta C-13 values, *New Phytol.* 152 (2001) 267–281.
- [53] Thomas S.C., Bazzaz F.A., Asymptotic height as a predictor of photosynthetic characteristics in Malaysian rain forest trees, *Ecology* 80 (1999) 1607–1622.
- [54] Tobin M.F., Lopez O.R., Kursar T.A., Responses of tropical understory plants to a severe drought: tolerance and avoidance of water stress, *Biotropica* 31 (1999) 570–578.
- [55] Turner I., *The Ecology of trees in the tropical rain forest*, Cambridge Tropical Biology Series, University Press, Cambridge, 2001.
- [56] Wright I.J., Reich P.B., Westoby M., Ackerly D.D., Baruch Z., Bongers F., Cavender-Bares J., Chapin T., Cornelissen J.H.C., Diemer M., Flexas J., Garnier E., Groom P.K., Gulias J., Hikosaka K., Lamont B.B., Lee T., Lee W., Lusk C., Midgley J.J., Navas M.-L., Niinemets U., Oleksyn J., Osada N., Poorter H., Poot P., Prior L., Pyankov V.I., Roumet C., Thomas S.C., Tjoelker M.G., Veneklaas E.J., Villar R., The worldwide leaf economics spectrum, *Nature* 428 (2004) 821–827.
- [57] Wright I.J., Reich P.B., Cornelissen J.H.C., Falster D.S., Garnier E., Hikosaka K., Lamont B.B., Lee W., Oleksyn J., Osada N., Poorter H., Villar R., Warton D.I., Westoby M., Assessing the generality of global leaf trait relationships, *New Phytol.* 166 (2005) 485–496.
- [58] Wright S.J., Plant diversity in tropical forests: a review of mechanisms of species coexistence, *Oecologia* 130 (2002) 1–14.
- [59] Xu Z.H., Saffigna P.G., Farquhar G.D., Simpson J.A., Haines R.J., Walker S., Osborne D.O., Guinto D., Carbon isotope discrimination and oxygen isotope composition in clones of the F1 hybrid between slash pine and Caribbean pine in relation to growth, water-use efficiency and foliar nutrient concentration, *Tree Physiol.* 20 (2000) 1209–1218.

APPENDIX O

A trait database for Guianan rain forest trees permits intra- and inter-specific contrasts

Ollivier, M., C. Baraloto et E. Marcon (2007). « A trait database for Guianan rain forest trees permits intra- and inter-specific contrasts ». In : *Annals of Forest Science* 64, p. 781–786.

A trait database for Guianan rain forest trees permits intra- and inter-specific contrasts

Mariwenn OLLIVIER^a, Christopher BARALOTO^{b*}, Eric MARCON^a

^a AgroParisTech – ENGREF, Unité Mixte de Recherches Écologie des Forêts de Guyane, Kourou, France

^b INRA, Unité Mixte de Recherches Écologie des Forêts de Guyane, Kourou, France

(Received 11 November 2006; accepted 13 March 2007)

Abstract – We present a plant trait database covering autecology for rain forest trees of French Guiana. The database comprises more than thirty traits including autecology (e.g., habitat associations and reproductive phenology), wood structure (e.g., density and tension characteristics) and physiology at the whole plant (e.g., carbon and nitrogen isotopes) and leaf level (e.g., specific leaf area, photosynthetic capacity). The current database describes traits for about nine hundred species from three hundred genera in one hundred families. For more than sixty species, data on twelve morphological and ecophysiological traits are provided for individual plants under different environmental conditions and at different ontogenetic stages. The database is thus unique in permitting intraspecific analyses, such as the effects of ontogenetic stages or environmental conditions on trait values and their relationships.

plant traits / tropical forest / French Guiana / functional groups / plasticity / ontogeny

Résumé – Une base de données sur l'autécologie des arbres de la forêt tropicale de Guyane française. Nous présentons une base de données sur l'autécologie des arbres de la forêt tropicale de Guyane française. La base contient des données sur plus de trente traits concernant l'autécologie (par exemple, les préférences d'habitat et la phénologie reproductive), la structure du bois (par exemple, la densité et les caractéristiques du bois de tension) et la physiologie aux niveaux de la plante entière (par exemple, les isotopes du carbone et de l'azote) et de la feuille (par exemple, la surface spécifique ou la capacité photosynthétique). Dans son état actuel, la base décrit les traits d'environ neuf cents espèces de trois cents genres dans cent familles. Pour plus de soixante espèces, des données sur douze traits morphologiques et écophysologiques sont fournis au niveau individuel pour des plants dans différentes conditions environnementales à différents stades ontogéniques. Cette base de données permet donc des analyses intraspécifiques, comme les effets des stades ontogéniques ou des conditions environnementales sur les valeurs des traits et leurs relations, ce en quoi elle n'a pas d'équivalent.

traits / forêt tropicale / Guyane française / groupes fonctionnels / plasticité / ontogénie

1. INTRODUCTION

Databases compiling species traits are important tools for plant ecologists to understand patterns of species abundance and distribution at a time of rapid loss of species diversity [10, 16, 17, 23, 24]. Recent studies have underlined at least four compelling research applications for such databases. First, trait databases can help us to understand basic strategies of resource use or biomass allocation among plants. Recent compilations [10, 34, 51, 52] illustrate how data from many different sources can be combined to confirm general conclusions of plant functioning that have been suggested from local datasets. Second, trait databases permit comparisons and contrasts of species diversity and plant functional types across natural environmental gradients, both within and among systems. For example, several studies demonstrate how trait values such as high foliar nutrient content are associated with particular environmental conditions such as high annual precipitation [33, 49, 50]. Third, trait databases are being used to select focal species for experimental communities to test relationships between species diversity, functional diversity and

ecosystem function [21, 40], or to refine subsequent analyses for existing experiments [31]. More recently, a fourth objective has been underlined, to understand evolutionary patterns among trait associations, such as the origin of seed mass associations with other plant traits [27, 28].

In general, within-species analyses for continuous traits, such as leaf attributes, use a mean trait value for species, without consideration of the variability masked by that mean value. To address this gap, we propose a fifth application of trait databases of a particular construction, within-species analyses. We recognize three particular types of intra-specific variability that could influence the mean value of traits reported in most databases, noting that analyses of each of these levels of variation represent advances for the application of trait databases. First, the observed phenotype of many plant traits can be strongly influenced by genotype of individuals for which trait screening has been conducted; we refer to this as the effect of *genetic diversity*. For example, Balaguer et al. [1] found significant differences in biomass allocation patterns and foliar nutrient contents among *Quercus coccifera* seedlings from three Mediterranean ecotypes differing in isozyme patterns. A second level of intraspecific trait variability occurs based on

* Corresponding author: baraloto.c@kourou.cirad.fr

the environmental conditions under which measurements are made; we refer to this as species *plasticity*. For example, foliar traits are often reported for ‘sun leaves’, but the definition of sun may include plants grown in pots under high transmission shade cloth and those in the field under open conditions [48]. In some cases, these environmental effects can interact with genotype effects so that the observed phenotype is the result of genotype \times environment interactions; for example, in the study by Balaguer et al. [1], the three ecotype populations responded differently when grown in sun vs. shade. A third level of variation that may occur within species involves differences in trait values with plant size or developmental stage; we refer to this as *ontogenetic plasticity*. In a recent meta-analysis, for example, Thomas and Winner [47] report significant differences between saplings and adult trees of 35 tree species, for several photosynthetic traits.

In this paper, we present MARIWENN, a trait database for woody plant species of the Guiana Shield region of South America that has been constructed to permit both intra- and inter-specific contrasts. First, we describe the construction of the database and the sources of available data; in doing so, we contrast the design and potential uses of the database with those of other plant trait databases such as GLOPNET and LEDA. We then present some examples of analyses that can be conducted using the database, including the unique aspect of within species comparisons in addition to the contemporary interspecific contrasts.

2. CONTENT

We gathered plant trait data for more than nine hundred woody plant species from French Guiana, representing over three hundred genera in more than one hundred families. Many data sources appear only in the grey literature, and thus would not otherwise be easily accessible to all researchers. The first part of the database was built to be an exhaustive compilation of the results of research on general species traits. No standardization of the data was made at this step; the purpose was just to organize the data rigorously to allow users to find data sources and the methods employed. The result is a comprehensive synthesis of data covering fields from wood structure to reproductive phenology (Tab. I). The modular structure of the database allows new data to be entered as it is generated.

The second purpose of the database was to structure data of plant traits to allow multivariate analyses. Unlike the first approach, this framework requires normative rules of measure and organization of the data. Moreover, specific measures are required to structure the database. The trait list reflects the state of the art of research and may change according to demands and new discoveries (Tab. II). Unlike the GLOPNET [23] databases, MARIWENN contains trait values measured on individual plants. Each value is then linked to many other fields that permit more complex queries: details of measurements (protocol); its author (reference); the environment, described with two levels of detail (general environment such as glasshouse or canopy, and detailed environment indicating the soil or the topographic position, or light level); and the ontogenetic stage of the plant. The mean and standard deviation of the trait can be computed as requests are made, for each ontogenetic stage and each environment type. Filters are available to reduce the dataset to a chosen light

level or detailed environment. This organization allows the retention of a large number of individuals or the isolation of particular environmental conditions, as a trade-off between sample size and variability among individuals.

The recorded traits are based on those described by Cornelissen et al. [7], without limitation (Tab. II). A priority of recent research has been leaf traits, including: specific leaf area, leaf area, laminar thickness, foliar carbon, nitrogen and phosphorus contents, and photosynthetic traits. An intensive campaign of measurement is being processed to enhance the database.

The botanical database is a straight adaptation of the checklist of the plants of the Guianas [3], including, where possible, a reference to the herbarium of Cayenne (IRD). Taxonomy is detailed down to the variety or subspecies, even though the standard level of detail is the species. Vernacular names are available as supplementary information. However, we caution the use of the database as a source of cross-referencing between scientific and common names because these links often vary between regions. The sites of field and experimental studies are referenced and their main characteristics detailed for each entry.

We chose to develop the database to maximize its versatility. No data related to the studied species are excluded a priori. The geographic limit is that of the botanical database which includes the plateau of the Guianas. The present content of the database is restricted to forest trees, but data from mangroves, savannas or non ligneous vegetation will be added as future research programs provide them.

3. USING THE DATABASE

All the published data are available through the Internet on <http://ecofog.cirad.fr/Mariwenn>. Unpublished data may be available in advance upon request of a password from the corresponding author. Future work will naturally be keeping data compilation up to date and also completing the trait records at plant level. We hope to gather individual data for most of the traits of the 100 most abundant woody species in French Guiana within two years.

Data can be obtained by species (all data available for a given species) or by topic (all species available for a given subject).

The web access is particularly easy to use but does not allow complex queries. Direct access using SQL queries is possible from the local network only, for technical and security reasons. Scientific collaborations are thus the easiest way to obtain complete access to the database, and interested researchers are invited to contact the corresponding author.

3.1. Examples of intraspecific analyses

In its current state, the database allows analyses within species for variation between environmental conditions, or between ontogenetic stages (see examples suggested in Tab. III). Current collections for trait screening are following half-sibling cohorts within species and will thus permit contrasts to be made to analyze ‘genotype’ or genotype \times environment effects on trait values.

Table I. Traits that have been measured at the species level that can be used in interspecific comparisons within this database or in concert with other databases, across sites or biomes.

Trait	References
<i>Ecophysiological data</i>	
Nitrogen: Isotopic signature ($\delta^{15}\text{N}$) and leaf nitrogen concentration in various forest sites	[35–38]
Carbon: $\delta^{13}\text{C}$ values and leaf carbon concentration at several sites	[4]
Photosynthesis-related ecophysiological parameters measured in glasshouse	[8]
<i>Biomechanics</i>	
Wood density at 12% moisture	[5]
Wood durability, impregnability; durability against termites and fungi	[14, 15]
Tension wood characteristics	[41]
<i>Soil-vegetation relations</i>	
Characterization of the edaphic preferences of species	[6, 30, 45]
<i>Architecture and phenology</i>	
Seedling morphology	[2]
Architectural patterns of trees	[18–20]
Vegetative phenology	[26]
Reproductive phenology	[9, 26, 42–44]
<i>Reproduction</i>	
Seeds and fruit characteristics	[2, 6, 12, 26, 42]
Pollen dispersal	[9]
<i>Forest dynamics</i>	
Pioneer species and soil seed bank	[29]
Response groups of species for light	[11]
Height groups: position of species in the vertical structure of the forest	[6]
Horizontal spatial structure of tree species	[6]

Table II. Traits describing species morphology and physiology that have been measured for individual plants for a given ontogenetic stage and under particular controlled environmental conditions, thereby permitting intra-specific analyses of species' plasticity across different environmental gradients, or ontogenetic shifts in trait values.

Trait	Unit	Measurement
Relative growth rate (RGR)	$\text{mg g}^{-1} \text{d}^{-1}$	After Hunt [22]
Root-shoot ratio	g g^{-1}	Root biomass/shoot biomass
Specific leaf area	$\text{cm}^2 \text{g}^{-1}$	leaf area/leaf biomass
Leaf blade surface area	cm^2	LICOR 3000 meter
SPAD chlorophyll estimation	SPAD units	Minolta SPAD-502 meter
Leaf thickness	μm	Mitutoyo caliper
Leaf dry matter content	mg g^{-1}	After Garnier et al. [13]
Leaf mass ratio	g g^{-1}	Leaf area/total biomass
Net assimilation rate (A_{max})	$\mu\text{mol CO}_2 \text{cm}^{-2} \text{s}^{-1}$	CIRAS-1 System
Dark respiration rate (R_d)	$\mu\text{mol CO}_2 \text{cm}^{-2} \text{s}^{-1}$	CIRAS-1 System
Stomatal conductance (G_s)	$\mu\text{mol CO}_2 \text{cm}^{-2} \text{s}^{-1}$	CIRAS-1 System
Foliar [$\delta^{13}\text{C}$]	‰	Mass spectrometer [4]
Foliar [N]	%	CHN autoanalyzer
Foliar [P]	%	HF digest; colorimetry
Water use efficiency (WUE)	$\mu\text{mol H}_2\text{O mol}^{-1} \text{CO}_2$	A_{max}/G_s
Nitrogen efficiency index (NUE)	$\mu\text{g g}^{-1} \text{d}^{-1}$	Foliar [N]/RGR
Phosphorus efficiency index (PUE)	$\mu\text{g g}^{-1} \text{d}^{-1}$	Foliar [P]/RGR

Table III. Examples of intra-specific calculations of species' plasticity or performance response ratios, across different environmental gradients, that can be performed using the MARIWENN database.

Calculated index	Unit	Calculation
Response ratio - Light	%	RGR_{hiite}/RGR_{loite}
Response ratio - Soil moisture	%	RGR_{hiSM}/RGR_{loSM}
Response ratio - Soil nutrients	%	RGR_{hiSN}/RGR_{loSN}
Plasticity in SLA	%	range of SLA/SLA_{max}
Plasticity in WUE	%	range of WUE/WUE_{max}
Plasticity in NUE	%	range of NUE/NUE_{max}
Plasticity in root-shoot allocation	$g\ g^{-1}$	range of RS/RS_{max}
Plasticity in leaf area ratio	$cm^2\ g^{-1}$	range of LAR/LAR_{max}

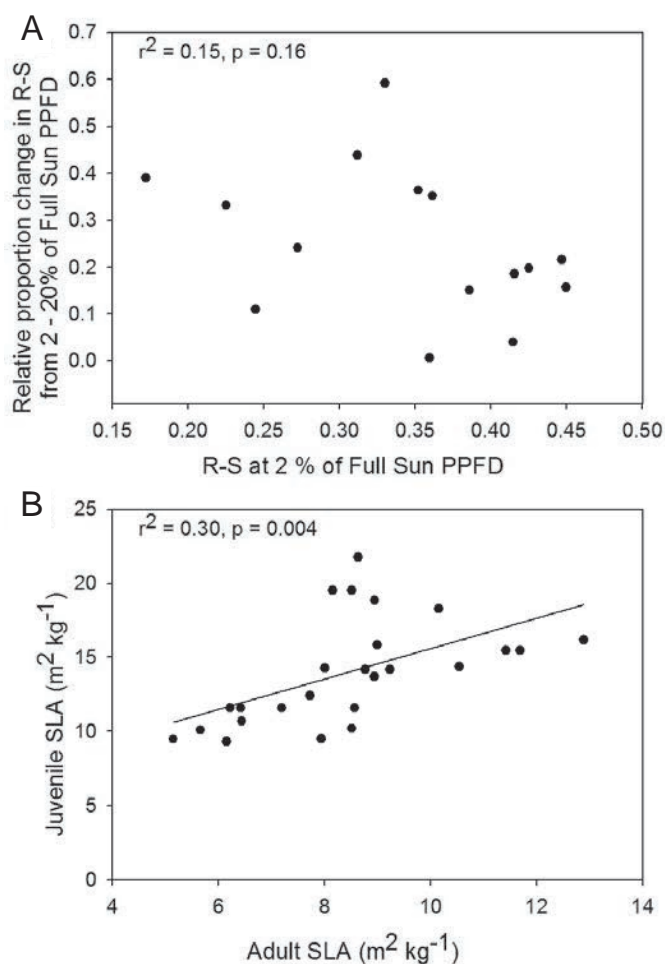


Figure 1. Examples of intra-specific analyses that can be conducted using the MARIWENN database. (A) Do species with particular mean values of a given trait exhibit greater breadth in trait values across a range of environmental conditions? In this example, we test whether species with low root-shoot ratio (R-S) have a larger range in R-S (relativized to maximum value; see Tab. II), across a light gradient varying from 2–20% of full sun. Data from C. Baraloto, unpublished. (B) Do species maintain trait values throughout developmental stages and/or size classes? In this example, we test whether mean values for SLA of sun leaves for 25 species change between juveniles and adult trees. Data from C. Baraloto and D. Bonal, unpublished.

Figure 1 illustrates two types of analyses that can be conducted using queries of the current database. The first example examines, for a given ontogenetic stage, if species-level trait breadth differs among species. In this case, the example addresses a species-level scenario for the hypothesis of Taylor and Aarssen [46] or Lortie and Aarssen [25] who suggest that a greater breadth of traits related to fitness should be exhibited by generalist species because they are exposed to selection under heterogeneous environments. If it is assumed that among tropical tree seedlings, the more specialized ecological guild is the light-demanding species, who generally have low root-shoot ratios [32], then we would predict a negative relationship between trait breadth and trait value in this case. However, no significant relationship was found for the species in the MARIWENN database (Fig. 1A).

The second example tests whether trait values, at a given environmental level (in this case leaves exposed to full sun) differ between developmental stages. Figure 1B shows a significant relationship between adult and juvenile specific leaf area (SLA). Nonetheless, a large degree of variation exists around this relationship, and many species pairs switch relative positions between stages. Moreover, as with the study of Thomas and Winner [47] or that of Roggy et al. [39], adult leaves have consistently lower SLA (or higher LMA).

3.2. Using these results to refine interspecific analyses

Each of the above examples shows how the intra-specific analyses can respond to particular research questions. In addition, we suggest that these types of analyses should serve as precursors to species-level analyses. When we find significant effects of environment or stage on mean trait values, this suggests that these factors need to be considered when conducting analyses among species. In the first example, (Fig. 1A), it is clear that the magnitude of shifts in root-shoot ratio between light environments differs among species (although not predictably based on a given trait value). This suggests that the results of multivariate analyses among species would be strongly dependent on the environmental conditions under which plants were grown for trait screening. Such variation may occur at what we have called the detailed environment, as in our example, or at what we have called the general environment.

For example, growing species in pots may influence the values of traits such as specific root length or root-shoot ratio (K. Kitajima, pers. comm.). The second example (Fig. 1B) indicates that for the 25 tropical tree species, multivariate analyses of foliar trait associations including specific leaf area (SLA, or its inverse, LMA), such as those conducted by Wright et al. [52], should control for the developmental stage of the plants measured in the database because species' values may shift rankings between stages.

Acknowledgements: We thank Jans Bakker and Jean-Christopher Roggy for valuable comments made on previous drafts of this manuscript. M. Ollivier was supported by EcoFoG Joint Research Unit and C. Baraloto acknowledges US NSF *OISE* 0301937.

REFERENCES

- [1] Balaguer L., Martinez-Ferri E., Valladares F., Perez-Corona M.E., Baquedano F.J., Castillo F.J., Manrique E., Population divergence in the plasticity of the response of *Quercus coccifera* to the light environment, *Funct. Ecol.* 15 (2001) 124–135.
- [2] Baraloto C., Forget P.-M., Seed size, seedling morphology, and response to deep shade and damage in neotropical rain forest trees, *Am. J. Bot.* 94 (2007) 901–911.
- [3] Boggan J., Funk V., Kelloff C., Hoff M., Cremers G., Feuillet C., Checklist of the Plants of the Guianas, 2nd ed., Museum of Natural History, Smithsonian Institution, Washington, D.C., 1997.
- [4] Bonal D., Barigah T.S., Granier A., Guehl J.-M., Late-stage canopy tree species with extremely low delta C-13 and high stomatal sensitivity to seasonal soil drought in the tropical rainforest of French Guiana, *Plant Cell Environ.* 23 (2000) 445–459.
- [5] Centre technique forestier tropical, Bois des DOM-TOM, 1989.
- [6] Collinet F., Essai de regroupement des principales espèces structurantes d'une forêt dense humide d'après leur répartition spatiale (forêt de Paracou, Guyane), Thèse de doctorat, Université Claude Bernard-Lyon I, Lyon, 1997.
- [7] Cornelissen J.H.C., Lavorel S., Garnier E., Diaz S., Buchmann N., Gurvich D.E., Reich P.B., Steege H.t., Morgan H.D., Heijden M.G.A.v.d., Pausas J.G., Poorter H., A handbook of protocols for standardised and easy measurement of plant functional traits worldwide, *Aust. J. Bot.* 51 (2003) 335–380.
- [8] Coste S., Roggy J.-C., Imbert P., Born C., Bonal D., Dreyer E., Leaf photosynthetic traits of 14 tropical rain forest species in relation to leaf nitrogen concentration and shade tolerance, *Tree Physiol.* 25 (2005) 1127–1137.
- [9] Degen B., Dendrobasis: Genetic system of tropical tree species, Silvolab, Kourou, 1999.
- [10] Diaz S., Hodgson J.G., Thompson K., Cabido M., Cornelissen J.H.C., Jalili A., Montserrat-Marti G., Grime J.P., Zarrinkamar F., Asri Y., Band S.R., Basconcelo S., Castro-Diez P., Funes G., Hamzehee B., Khoshnevi M., Perez-Harguindeguy N., Perez-Rontome M.C., Shirvany F.A., Vendramini F., Yazdani S., Abbas-Azimi R., Bogaard A., Boustani S., Charles M., Dehghan M., de Torres-Espuny L., Falczuk V., Guerrero-Campo J., Hynd A., Jones G., Kowsary E., Kazemi-Saeed F., Maestro-Martinez M., Romo-Diez A., Shaw S., Siavash B., Villar-Salvador P., Zak M.R., The plant traits that drive ecosystems: Evidence from three continents, *J. Veg. Sci.* 15 (2004) 295–304.
- [11] Favrichon V., Classification des espèces arborées en groupes fonctionnels en vue de la réalisation d'un modèle de dynamique de peuplement en forêt guyanaise, *Rev. Ecol. Terre Vie* 49 (1994) 379–403.
- [12] Forget P.-M., Dissémination et régénération naturelle de huit espèces d'arbres en forêt guyanaise, Ph.D. thesis, Université de Paris VI, 1988.
- [13] Garnier E., Shipley B., Roumet C., Laurent G., A standardized protocol for the determination of specific leaf area and leaf dry matter content, *Funct. Ecol.* 15 (2001) 688–695.
- [14] Gérard J., Narboni P., Une base de données sur les propriétés technologiques des bois, *Bois For. Trop.* 248 (1996) 65–70.
- [15] Gérard J., Edi Kouassi A., Daigremont C., Détienné P., Fouquet D., Vernay M., Synthèse sur les caractéristiques technologiques de référence des principaux bois commerciaux africains, Série FORAFRI, CIRAD-Forêt, 1998.
- [16] Grime J.P., Declining plant diversity: empty niches or functional shifts? *J. Veg. Sci.* 13 (2002) 457–460.
- [17] Grime J.P., Thomson K., Hunt R., Hodgson J.G., Cornelissen J.H.C., Rorison I.H., Hendry G.A.F., Ashenden T.W., Askew A.P., Band S.R., Booth R.E., Bossard C.C., Campbell B.D., Cooper J.E.L., Davison A.W., Gupta P.L., Hall W., Hand D.W., Hannah M.A., Hillier S.H., Hodgkinson D.J., Jalili A., Liu Z., Mackey J.M.L., Matthews N., Mowforth M.A., Neal A.M., Reader R.J., Reiling K., Ross-Fraser W., Spencer R.E., Sutton F., Tasker D.E., Thorpe P.C., Whitehouse J., Integrated screening validates primary axes of specialisation in plants, *Oikos* 79 (1997) 259–281.
- [18] Hallé F., Oldeman R.A.A., Essai sur l'architecture et la dynamique de croissance des arbres tropicaux, 1970.
- [19] Hallé F., Oldeman R.A.A., Tomlinson P.B., Tropical trees and forests – an architectural analysis, 1978.
- [20] Heuret P., Analyse et modélisation de séquences d'événements botaniques : application à la compréhension des processus de croissance, de ramification et de floraison, Ph.D. thesis, Université de Nancy I, 2002.
- [21] Hooper D.U., Chapin F.S., Ewel J.J., Hector A., Inchausti P., Lavorel S., Lawton J.H., Lodge D.M., Loreau M., Naeem S., Schmid B., Setälä H., Symstad A.J., Vandermeer J., Wardle D.A., Effects of biodiversity on ecosystem functioning: A consensus of current knowledge, *Ecol. Monogr.* 75 (2005) 3–35.
- [22] Hunt R., Plant growth analysis, The Institute of Biology's Studies in Biology, Edward Arnold, London, 1978.
- [23] Knevel I.C., Bekker R.M., Bakker J.P., Kleyer M., Life-history traits of the northwest European flora: The LEDA database, *J. Veg. Sci.* 14 (2003) 611–614.
- [24] Kuhn I., Durka W., Klotz S., BiolFlor – a new plant-trait database as a tool for plant invasion ecology, *Divers. Distrib.* 10 (2004) 363–365.
- [25] Lortie C.J., Aarssen L.W., The specialization hypothesis for phenotypic plasticity in plants, *Int. J. Plant Sci.* 157 (1996) 484–487.
- [26] Loubry D., Déterminisme du comportement phénologique des arbres en forêt tropicale humide de Guyane française, Ph.D. thesis, Université Paris 6, 1994.
- [27] Moles A.T., Ackerly D.D., Webb C.O., Tweddle J.C., Dickie J.B., Westoby M., A brief history of seed size, *Science* 307 (2005) 576–580.
- [28] Moles A.T., Ackerly D.D., Webb C.O., Tweddle J.C., Dickie J.B., Pitman A.J., Westoby M., Factors that shape seed mass evolution, *Proc. Natl. Acad. Sci. USA* 102 (2005) 10540–10544.
- [29] Molino J.-F., Sabatier D., Tree diversity in tropical rain forests: a validation of the intermediate disturbance hypothesis, *Science* 294 (2001) 1702–1704.
- [30] Paget D., Étude de la diversité spatiale des écosystèmes forestiers guyanais : réflexion méthodologique et application, Ph.D. thesis, ENGREF, 1999.
- [31] Petchey O.L., Hector A., Gaston K.J., How do different measures of functional diversity perform? *Ecology* 85 (2004) 847–857.
- [32] Poorter L., Growth responses of 15 rain-forest tree species to a light gradient: the relative importance of morphological and physiological traits, *Funct. Ecol.* 13 (1999) 396–410.
- [33] Reich P.B., Oleksyn J., Global patterns of plant leaf N and P in relation to temperature and latitude, *Proc. Natl. Acad. Sci. USA* 101 (2004) 11001–11006.

- [34] Reich P.B., Ellsworth D.S., Walters M.B., Vose J.M., Gresham C., Volin J.C., Bowman W.D., Generality of leaf trait relationships: A test across six biomes, *Ecology* 80 (1999) 1955–1969.
- [35] Roggy J.-C., Contribution des symbioses fixatrices d'azote à la stabilité de l'écosystème forestier tropical guyanais, Thèse Université C. Bernard Lyon I, 1998.
- [36] Roggy J.-C., Prévost M.-F., Nitrogen-fixing legumes and silvigenesis in a rain forest in French Guiana: a taxonomic and ecological approach, *New Phytol.* 144 (1999) 283–294.
- [37] Roggy J.-C., Prévost M.-F., Garbaye J., Domenach A.-M., Nitrogen cycling in the tropical rain forest of French Guiana: comparison of two sites with contrasting soil types using delta-15N, *J. Trop. Ecol.* 15 (1999) 1–22.
- [38] Roggy J.-C., Prévost M.-F., Gourbiere F., Casabianca H., Garbaye J., Domenach A.-M., Leaf natural 15N abundance and total N concentration as potential indicators of plant N nutrition in legumes and pioneer species in a rain forest of French Guiana, *Oecologia* 120 (1999) 171–182.
- [39] Roggy J.-C., Nicolini E., Imbert P., Caraglio Y., Bosc A., Heuret P., Links between tree structure and functional leaf traits in the tropical forest tree *Dicorynia guianensis* Amshoff (Caesalpinaceae), *Ann. For. Sci.* 62 (2005) 553–564.
- [40] Roscher C., Schumacher J., Baade J., Wilcke W., Gleixner G., Weisser W.W., Schmid B., Schulze E.D., The role of biodiversity for element cycling and trophic interactions: an experimental approach in a grassland community, *Basic Appl. Ecol.* 5 (2004) 107–121.
- [41] Ruelle J., Anatomie comparative bois normal/bois de réaction et observation des relations structure/propriétés du bois de six espèces d'angiospermes de forêt tropicale humide et de trois espèces de gymnospermes de forêt tempérée, Nancy, 2003.
- [42] Sabatier D., Fructification et dissémination en forêt guyanaise : l'exemple de quelques espèces ligneuses, Université des Sciences et Techniques du Languedoc, Montpellier, 1983.
- [43] Sabatier D., Saisonnalité et déterminisme du pic de fructification en forêt guyanaise, *Rev. Ecol. Terre Vie* 40 (1985) 289–320.
- [44] Sabatier D., Puig H., Phénologie et saisonnalité de la floraison et de la fructification en forêt dense guyanaise, Muséum National d'Histoire Naturelle, Paris, 1982.
- [45] Sabatier D., Grimaldi M., Prévost M.-F., Guillaume J., Godron M., Dosso M., Curmi P., The influence of soil cover organization on the floristic and structural heterogeneity of a Guianan rain forest, *Plant Ecol.* 131 (1997) 81–108.
- [46] Taylor D.R., Aarssen L.W., An interpretation of phenotypic plasticity in *Agropyron repens*, *Am. J. Bot.* 75 (1988) 401–413.
- [47] Thomas S.C., Winner W.E., Photosynthetic differences between saplings and adult trees: an integration of field results by meta-analysis, *Tree Physiol.* 22 (2002) 117–127.
- [48] Whitmore T.C., A review of some aspects of tropical rain forest seedling ecology with suggestions for further inquiry, in: Swaine M.D. (Ed.), *The ecology of tropical forest tree seedlings*, UNESCO/Parthenon Publishing, Paris 1996, pp. 3–39.
- [49] Wright I.J., Westoby M., Leaves at low versus high rainfall: co-ordination of structure, lifespan and physiology, *New Phytol.* 155 (2002) 403–416.
- [50] Wright I.J., Reich P.B., Westoby M., Strategy shifts in leaf physiology, structure and nutrient content between species of high- and low-rainfall and high- and low-nutrient habitats, *Funct. Ecol.* 15 (2001) 423–434.
- [51] Wright I.J., Reich P.B., Cornelissen J.H.C., Falster D.S., Garnier E., Hikosaka K., Lamont B.B., Lee W., Oleksyn J., Osada N., Poorter H., Villar R., Warton D.I., Westoby M., Assessing the generality of global leaf trait relationships, *New Phytol.* 166 (2005) 485–496.
- [52] Wright I.J., Reich P.B., Westoby M., Ackerly D.D., Baruch Z., Bongers F., Cavender-Bares J., Chapin T., Cornelissen J.H.C., Diemer M., Flexas J., Garnier E., Groom P.K., Gulias J., Hikosaka K., Lamont B.B., Lee T., Lee W., Lusk C., Midgley J.J., Navas M.-L., Niinemets U., Oleksyn J., Osada N., Poorter H., Poot P., Prior L., Pyankov V.I., Roumet C., Thomas S.C., Tjoelker M.G., Veneklaas E.J., Villar R., The worldwide leaf economics spectrum, *Nature*. 428 (2004) 821–827.

APPENDIX P

Evaluating the geographic concentration of industries using distance-based methods

Marcon, E. et F. Puech (2003). « Evaluating the geographic concentration of industries using distance-based methods ». In : *Journal of Economic Geography* 3.4, p. 409–428.

Evaluating the geographic concentration of industries using distance-based methods

*Eric Marcon** and *Florence Puech***

Abstract

We propose new methods for evaluating the spatial distribution of firms. To assess whether firms are concentrated or dispersed, economists have traditionally used indices that analyse the heterogeneity of a spatial structure at a single geographic level. We introduce distance-based methods, Besag's *L* function (derived from Ripley's *K* function) and Diggle and Chetwynd's *D* function to describe *simultaneously* spatial distribution at different geographical scales. Our empirical applications consider the distribution of French manufacturing firms in the Paris area and in France generally. For some geographic levels, results show significant concentration or dispersion of firms according to their sector of activity.

Keywords: agglomeration, clustering, geographic concentration, location of firms

JEL classification: C40, C60, L60, R12

Date submitted: 20 March 2002 **Date accepted:** 10 March 2003

1. Introduction

In recent years, the role of economic activity location and the spatial concentration of manufacturing industry have played a more prominent role in economics. Although the subject of the location of production is not new (Marshall, 1920; Lösch, 1940), there has been a renewed interest in spatial economics owing to recent works and models in the new economic geography (for a survey of this literature, see Jayet et al., 1996; Duranton, 1997; Ottaviano and Puga, 1998).

The spatial distribution of economic activity is not homogeneous and empirical examples of geographic agglomeration of manufacturing are readily observed, including the US manufacturing belts and the carpet industry in Dalton (Krugman, 1991).

To assess the geographic distribution of activity in a given territory, economists have traditionally used concentration indices such as those defined by Herfindahl, Gini, or Ellison and Glaeser (1997). However, these methods evaluate the heterogeneity of the spatial structure *at a single geographic level*. In other words, concentration is generally measured at an administrative scale (for instance at the national, regional or state level). Obviously, since spatial distribution may differ according to the observation level, it would be better to measure each of these levels at the same time. This calls for the use

* ENGREF, BP 316, 97310 Kourou, French Guyana.

email <Marcon@netcourrier.com>

** Corresponding author at: TEAM, University of Paris I–CNRS, 106–112 Boulevard de l'Hôpital, 75647 Paris Cedex 13, France.

email <Puech@univ-paris1.fr>

of a new approach, based on distances, that can describe spatial distribution at different geographic levels *simultaneously*.

Consequently, in this paper, we propose distance-based methods for evaluating the spatial distribution of firms in a given territory. Unlike other fields such as forestry (see for instance Pélissier and Goreaud, 2001), distance-based methods are quite new in economics, and as far as we know, only three existing articles use such an approach (Barff, 1987; Sweeney and Feser, 1998; Duranton and Overman, 2002).

Our article is organized as follows. In Section 2, we provide reasons for the use of new geographic concentration measures in economics. In Section 3, we explain the methodology employed, i.e., Ripley's K function (Ripley 1976, 1977). In practice, we use Besag's L function (Besag, 1977), which is generally preferred because its results are simpler to interpret. In Section 4, we introduce Diggle and Chetwynd's D function to account for the heterogeneity of space (Diggle and Chetwynd, 1991). Finally, we compare our mathematical methods to that proposed by Duranton and Overman (2002).

2. Why we need a new method for measuring the geographic concentration of firms

To analyse the spatial distribution of firms, economists have traditionally used concentration indices to assess whether there is agglomeration or dispersion of firms in a given territory.

Many concentration measures exist (Valeyre, 1993; Houdebine, 1999), but economists generally use three main tools to gauge geographic concentration: Herfindahl, Gini or Ellison and Glaeser measures.

In the context of a single industry, the Herfindahl index corresponds to the sum of regions' squared weights (with each weight being the number of employees in region i divided by the number of employees in all regions, denoted by z_i). Consequently, if we consider a territory divided into n areas indexed by i , the Herfindahl index is equal to $H = \sum_i z_i^2$. The result is in the interval $[1/n; 1]$. If all activity is located in one area, concentration is at its maximum and the Herfindahl index is equal to 1. Conversely, if there is a perfect distribution of activity in the territory, the Herfindahl is equal to $1/n$. This index is simple to compute but only assesses spatial concentration because it does not take into account the distribution of all economic activity. In other words, it does not compare a sector of activity's concentration to that of other sectors but to spatial homogeneity. Gini's index fills this gap.

Traditionally, the most frequently used index for measuring the spatial concentration of economic activity is the Gini coefficient; see, for example, Krugman (1991). Variants have been used by, for example, Brülhart and Torstensson (1996), Amiti (1997), Haaland et al. (1999), Midelfart-Knarvik et al. (2000), and Brülhart (2001).

The Gini index is constructed as follows. Consider a sector of activity s and a territory divided into regions. For each region, we calculate the ratio of the share of total employment in sector s to the share of total employment in all sectors of activity (in the considered region). We then rank the ratios to construct the location curve (namely the Lorenz curve). For this, respecting the order of magnitude of the ratios, we represent the regions' cumulative shares of total employment in sector s (on the vertical axis) and the regions' cumulative shares of total employment in all sectors of activity (on the horizontal axis). The ratio of the area between the resulting curve and the first bisector to the area under the first bisector is the Gini coefficient.

In the case of a perfectly homogeneous distribution, the shares of total employment in sector s across all regions are equal to the share of employment in all sectors. The Gini coefficient is equal to zero because the location curve and the 45-degree line coincide. At the other extreme, the more the spatial distribution in sector s is concentrated, the further the Lorenz curve is from the bisector, and the closer the Gini coefficient is to 1.

The Gini coefficient does not take into account industrial concentration (the size of firms). For a clearer view of concentration, we should consider the fact that an industry may be concentrated because:

- a small number of firms have a large number of employees in a given area; or
- spatial clustering is the result of numerous small-sized firms being located in the same geographic location (Devereux et al., 1999).

A new index, proposed by Ellison and Glaeser in 1997, incorporates this distinction.

The concentration index of Ellison and Glaeser (Ellison and Glaeser, 1997) calculates the deviation of the observed geographic concentration of an industry from the concentration that would result from firms locating independently and randomly (the authors use the metaphor of the 'dartboard approach'). In the 'random-location model', an industry is considered to be concentrated if the observed concentration differs significantly from a random distribution of firms. Ellison and Glaeser define a gross geographic concentration index, G , which is calculated for a given industry as the squared sum of the differences (in each region) between the share of employment (s_i) in the considered sector and total employment (x_i) in all manufacturing sectors; i.e. $G = \sum_i (s_i - x_i)^2$. Ellison and Glaeser show that the expected value of G for a completely random distribution of firms, denoted by $E(G)$, is $E(G) = (1 - \sum_i x_i^2)H$, where $(1 - \sum_i x_i^2)$ measures the economic activity across locations and H is the Herfindahl index.

Hence, we obtain the following expression for the Ellison and Glaeser index (denoted by γ):

$$\gamma = (G - (1 - \sum_i x_i^2)H) / ((1 - \sum_i x_i^2)(1 - H)).$$

Testing the statistical significance of the index indicates whether a sector's distribution of activity across locations is significantly concentrated or dispersed. This index has been applied at the national level for the US (Ellison and Glaeser, 1997; and Rosenthal and Strange, 2001), the UK (Devereux et al., 1999), Spain (Callejón, 1997) and France (Maurel and Sédillot, 1999; Houdebine, 1999), and has been applied in the international context to foreign direct investment (Mucchielli and Puech, 2001; Head et al., 2002).

Having surveyed existing indices, it is worth noting that these measures only describe the location of economic activity on a *single scale*. In other words, an arbitrary geographic level of clusters (say regions) is chosen and the concentration index is computed for a set of these clusters (the country). However, to describe spatial patterns more accurately, we need to explain the spatial structure at different scales *at the same time*, not just according to administrative or arbitrary geographic scales. To illustrate this, consider the spatial distribution of firms below. The global area under consideration is divided into four areas.

The first territory contains 60 firms, evenly distributed. The other three areas contain five firms. The Herfindahl index is equal to 49/75. This value is to be compared with the minimum value of 1/4 (when each area has the same number of firms) to measure concentration at the level under consideration. The first territory has the wide

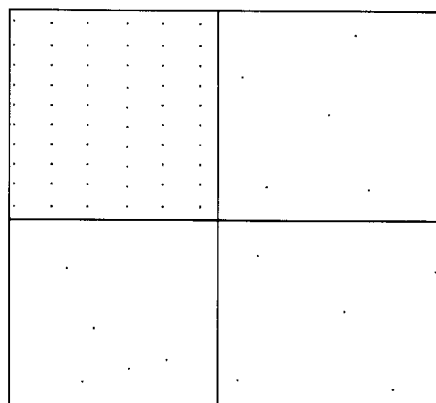


Figure 1. Arbitrary distribution of firms.

majority of firms, and their distribution is perfectly even. Cluster-based methods do not detect concentration at scales other than the one under consideration, and in our example, failed to detect an obvious dispersion at the intra-regional level. Distance-based methods such as Besag's L function, which is derived from Ripley's K function, reveal the true extent of spatial concentration. An advantage of distance-based methods is that they circumvent the *scale problem* described by Arbia (1989) as a manifestation of the 'Modifiable Areal Unit Problem' (MAUP). That is, conclusions from the evaluation of the concentration of a zoning area may depend on the scale chosen.

3. A distance-based measure: Ripley's K function

In this section, we summarize some point-pattern properties in order to introduce properly Ripley's K function. We also present Besag's L function, which is generally preferred because its results are simpler to interpret. We then discuss mathematical difficulties such as edge-effect problems. Finally, we explain how to construct confidence intervals.

3.1. Point patterns, the Poisson process and Ripley's K function

3.1.1. Intuitive presentation

Consider firms located in an area. Suppose that this area is equally attractive to all firms, so we can expect the same density of firms across the entire area. We count each firm's number of neighbors (other firms) within a given distance. We then calculate the *average* number of neighbors of *every* firm at *each* distance.

The benchmark is known as *complete spatial randomness* (CSR). In this case, firms not only locate at any place with the same probability (constant density), they also locate independently of each other. Now suppose that the location of firms is not random but is dependent on the location of other firms. If proximity to other firms is attractive to a firm, that firm will have more neighbors around it than if it were to locate randomly. On the contrary, finding fewer neighbors than expected implies that firms try to locate away from each other.

To determine whether the distribution of firms is significantly different from CSR, we use a mathematical function called L .

In what follows, we use the following terms:

- We consider a distribution to be *homogeneous* when the expected density of firms is constant everywhere in the territory.
- A distribution may be *completely random*, or *independent*, if each firm locates randomly and independently of the others. In this case, it is a Poisson distribution.
- Firms may attract each other, which leads us to observe aggregates. We discuss *concentration* or *agglomeration*. The observed density will be greater in the aggregates, which may appear to contradict the hypothesis of homogeneity. The mathematical tool used to overcome this apparent paradox was developed by Ripley (1977) and is presented in Appendix 1.
- Firms may repulse each other, which leads to *dispersion*: maximum dispersion is given by perfect regularity.

3.1.2. Mathematical definitions

A rigorous presentation of Ripley's K function can be found in Appendix 1. Alternatively, Diggle (1983), Upton and Fingleton (1985), and Cressie (1993), provide a full explanation of this function.

The K function was proposed by B.D. Ripley (Ripley, 1976, 1977). It was slightly modified by Besag, 1977, in the form of the L function, which is commonly used for simplicity of interpretation. Ripley's K function describes the spatial distribution of a set of points. We use λ to denote the average density of points. The density is considered to be constant. For each point i of the subplot under consideration, supposing a completely random distribution, the expected number of points in a circle of radius r is $\lambda\pi r^2$. Points located inside the circle around a firm are its *neighbors*. $K(r)$ is defined as the average number of neighbors divided by λ . Therefore, CSR leads to $K(r) = \pi r^2$. This value is used as a benchmark.

A practical limitation of Ripley's K function is the need to compare any value to πr^2 . Besag, 1977, normalized the function to obtain a benchmark of zero:

$$L(r) = \sqrt{\frac{K(r)}{\pi}} - r \quad (1)$$

$L(r) > 0$ indicates that the observed distribution is geographically concentrated, while $L(r) < 0$ implies dispersion. Consequently, computing L for a wide range of radius values and comparing with values obtained by the null hypothesis of a random distribution identifies significant concentration or significant dispersion at different geographic scales.

In summary, the K and L functions measure concentration by counting each firm's average number of neighbors within a circle of a given radius. We calculate the average number of firms for each radius (for instance every kilometer) and then check whether the actual distribution of firms differs significantly from a random one. For this, we generate a confidence interval by simulating a large number of independent homogeneous spatial distributions with the same number of plots and the same density using the Monte Carlo method.

3.2. Practical computation

3.2.1. The problem of 'edge effects' associated with the K function

The computation of L involves counting each point's number of neighbors in a circle of a given radius. The number of neighbors for points close to the boundary is underestimated because a part of the circle is outside the area under study (and hence contains no points), which produces a bias in the results. Appendix 2 illustrates the edge effect. Following Besag (1977), we correct this bias by using only part of the circle's area, i.e., only the part included in the area under study (or the intersection area). The number of neighbors is corrected by a factor equal to the circle's area divided by the intersection area. Unlike most authors on this subject, we choose as a correcting factor the area rather than the perimeter of the circle. This means that our interpretation need not be mathematically restricted to a half or a third of the size of the studied area (see Diggle, 1983; Rowlingson and Diggle, 1993). Our empirical results can be interpreted for any distance. However, the wider the radius, the greater the correction of the edge effect, which causes L values to tend towards zero. Even though some computer programs compute Ripley's K function, we chose to develop our own software. The software is downloadable from the authors' website¹ or available from the authors on request. For Ripley's functions, the computation of the retained correction of border effects was explained above.

3.2.2. Construction of the confidence intervals

Since K and L distributions are unknown, their variance cannot be evaluated. To construct the confidence interval for the null hypothesis, the classical technique is to generate a large number of independent random distributions of points (the Monte Carlo method, used, for example, by Goreaud, 2000). Practically, we generate a large number of simulations of homogeneous spatial distributions with the same number of points and the same density as we have in our sample. A confidence level—say 5%—is chosen. The 5% confidence interval of K for each value of r is delimited by the outer 5% of the randomly generated values. The more simulations generated, the narrower the confidence interval. We generate 20 to 10,000 simulations depending on the computing time available.

3.2.3. Computation and issues

We follow the steps below to compute an estimator of an L function and its confidence interval. We denote the number of points by n and use L rather than \hat{L} to denote the estimator, for readability.

1. Calculate the distance between each couple of points. The number of couples is $n(n-1)/2$. Computing time consequently increases in proportion to the square of the number of points.
2. For each chosen value of the radius and for each point, count the number of neighbors (points closer than r to the reference point) and correct for edge effects. Then calculate the average number of neighbors.

1 <http://e.marcon.free.fr/Ripley> (French and English versions).

3. Generate a certain number of random distributions (from 20 to 10,000 as suggested) of the same number of points in the entire area under consideration. For each random set, calculate L .

In summary, calculating L is quite computer intensive, with computing time approximately proportional to the square of the number of points multiplied by the number of random simulations. In addition, complexity depends on the shape of the area under study. Our empirical analysis was limited to rectangular areas for two reasons. First, calculating edge effects and simulating random points inside the area requires exact knowledge of boundary coordinates. Second, the correction of edge effects on convex shapes requires more complicated algorithms.

3.2.4. Curve interpretation

An independent distribution of points leads to a flat L curve; i.e. L is equal to zero, whatever the considered distance. Figure 2 and Figure 3 are taken from Goreaud (2000). Figure 2 shows a regular distribution of points (the map on the left). The L curve (on the right) indicates negative peaks at the maximum dispersion distances. CI is the confidence interval at the 1% level, generated by using 10,000 simulations. Note that the negative peak is just before the grid size: for $r < 10$, points have no neighbors, so $L = -r$.

Figure 3 illustrates an agglomerated distribution of points. The L positive peak for $r = 8$ corresponds to the average aggregate radius.

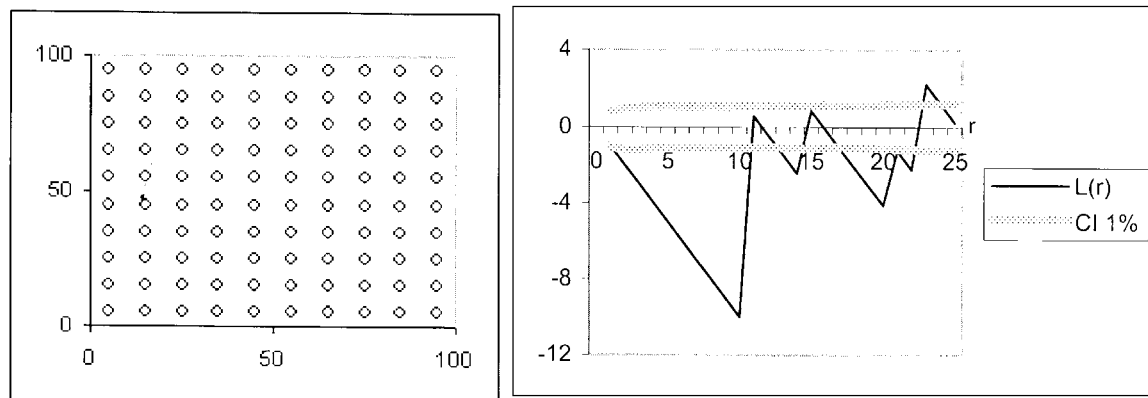


Figure 2. Regular distribution.

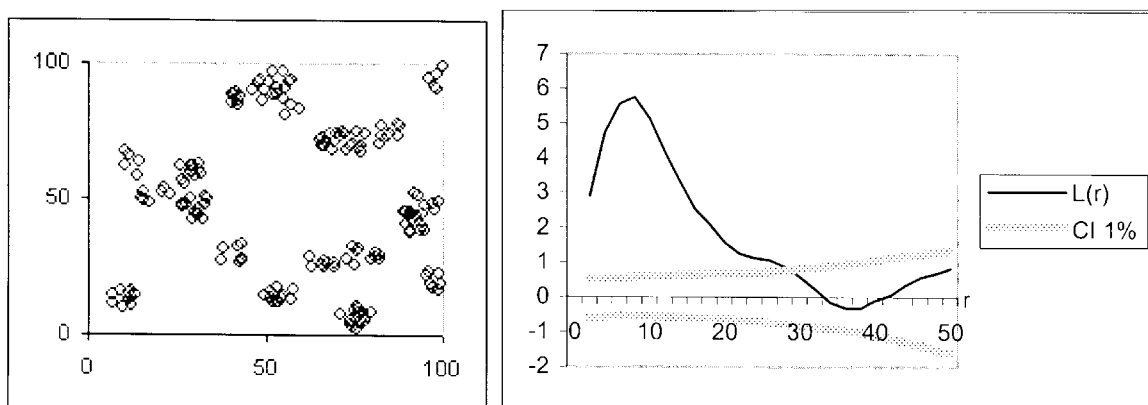


Figure 3. Concentrated distribution.

The L curve identifies the spatial structure.

- Significantly positive values indicate concentration. Peaks correspond to the aggregates' radii.
- Significantly negative values indicate dispersion. Peaks correspond to the grid size.
- A confidence interval allows the elimination of non-significant peaks.

3.3. The database and the French regions studied

Our study is based on the Annual Business Survey 'Enquête Annuelle d'Entreprise' (EAE), which records information at the business-unit level, conducted by the French Ministry of Industry. This annual survey lists *all* French manufacturing firms in France employing at least twenty workers (the food industry is not included). The spatial distribution of firms is recorded for 1996, for which year the database contains around 45,000 observations. We only use information on productive plants.

The zip code of each manufacturing firm is known. Consequently, we use a geographic database based on Lambert geographic coordinates,² in order to construct a location map of French manufacturing firms. The largest cities, such as Paris, have several zip codes, which improves precision. We have the data for more than 36,000 locations. The margin of error for any firm's location is about 2 km.

Since correcting edge effects on complicated shapes requires complex algorithms, we adopt a rectangular region in our empirical application. The region is defined by a large industrial area of 40×40 km around Paris.

Concerning sectors of activity, we chose the French Activity Nomenclature 36 (NAF) which is a nomenclature that lists the main sectors of economic activity. Manufacturing activities are classified into 14 sectors: clothing and leather (C1); printing and publishing (C2); pharmacy and perfumery (C3); home equipment (C4); motor vehicles (D0); naval, aeronautic and rail construction (E1); mechanical engineering (E2); electric and electronic products (E3); mineral products (F1); textiles (F2); wood and paper products (F3); chemical, rubber and plastics industry (F4); metallurgy and metal transformation (F5); and electrical engineering (F6).

3.4. Results

3.4.1. The Paris area

Our empirical analysis concentrates on French manufacturing industries in the Paris area. We consider Paris and its suburbs. Figure 4 shows the spatial distribution of manufacturing firms (several firms with the same zip code appear as a single point). No obvious structure is apparent. The number of business units is 5,739 covering all sectors of manufacturing activity.

At 1 km steps, we calculate the L function for each sector and for manufacturing as a whole. Confidence intervals are computed at the 5% level. We used only 100 simulations because the values of the L functions are sufficiently higher than those of the confidence interval. Note that L_{\min} and L_{\max} values close to zero limit the interval in which the null hypothesis of a random distribution is valid. For each of the 14 sectors considered and for manufacturing as a whole, the associated L curves show significant concentration for all distances from 0 to 25 km. Above this value, distances resemble the linear extent of the

2 This is a French projection system, in which each zip code is defined by its coordinates in kilometers.

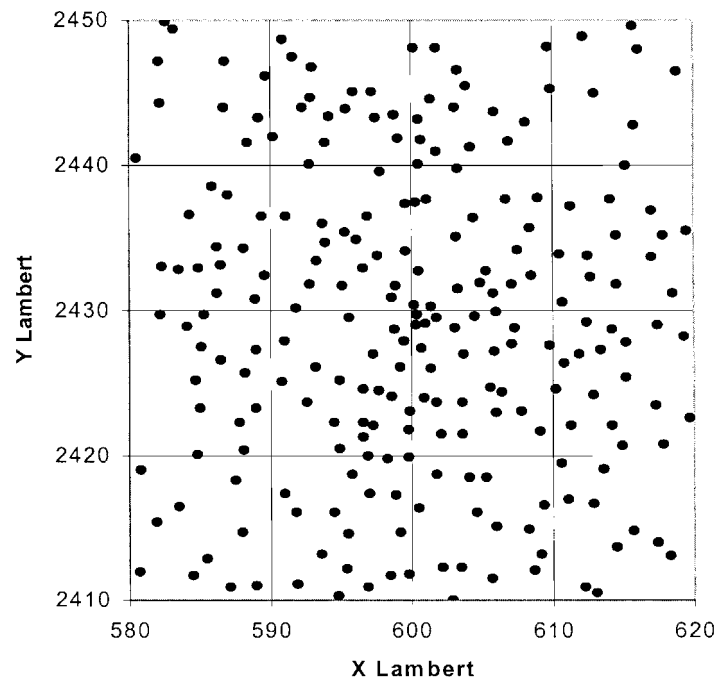


Figure 4. Paris area distribution of firms.

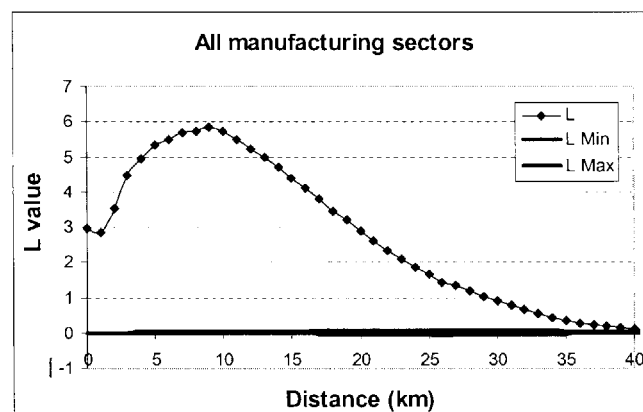


Figure 5. L function for all sectors around Paris (5,739 firms).

area: i.e. the edge-effect corrections dominate actual values and L tends to zero. In what follows, we present only three L curves and their associated confidence intervals for the time period under study: one for all manufacturing activity and two for individual sectors of activity. All other L curves are available on request from the authors.

Empirical results from Besag's function for all manufacturing activity (Figure 5), for clothing and leather (Figure 6), and for electric and electronic products (Figure 7), show that manufacturing firms are significantly concentrated. The main features of these results do not lie on the observed geographic agglomeration of firms but allow detection of differences between geographical concentration scales according to the industry considered; i.e. significant spatial concentration peaks for each sector do not occur at the same distance.

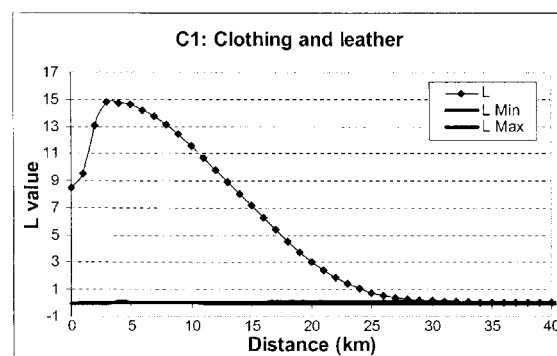


Figure 6. L function for sector C1 around Paris (1,011 firms).

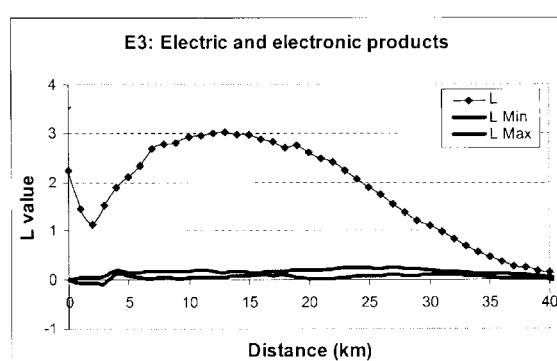


Figure 7. L function for sector E3 around Paris (490 firms).

The concentration peak for all manufacturing is at a radius of 9 km. Clothing and leather is spatially concentrated at 3 km, and the peak for electric and electronic products is at 13 km. These different concentration patterns are first described, then explained.

3.4.2. Economic stakes and involvements

We used these data to illustrate the usefulness and limitations of the tool.

Our applications gave some results to confirm the utility of a distance-based method. We simultaneously detected concentration at different scales: differences between sectors enable consideration of potentially discriminating economic determinants. We showed that all sectors are concentrated. This proves that firms do not locate randomly and that aggregates can be observed. This initial result may be considered quite trivial. A more interesting approach involves comparing the distribution of a particular set of other firms (say a sector of activity) to the distribution of all firms. This new benchmark allows space to be considered heterogeneous and enables evaluation of concentration relative to average industrial density rather than to spatial homogeneity.

4. Assuming non-homogeneous populations

By using K and L functions, we examined the hypothesis of constant density (the case of a homogeneous distribution). In this section, we reconsider this hypothesis by assuming *heterogeneity* of the distribution. This seems more realistic for large areas and areas that

include geographical features such as lakes and mountains, where firms cannot locate. Diggle and Chetwynd (1991) propose a function called the D function to take into account density variations. Sweeney and Feser (1998) first used this function in economics to investigate whether small firms were more concentrated than large ones.

4.1. Definition

We use a case-control design: a population of *cases* (firms belonging to a particular sector of activity) is compared to a reference population of *controls* (all other manufacturing firms). The distribution of firms' location is given. The null hypothesis is random labeling; i.e. a firm can belong randomly to the cases or the controls. Diggle and Chetwynd (1991) showed that, at any distance r , $K_{\text{cases}}(r) = K_{\text{controls}}(r)$, where $K_{\text{cases}}(r)$ and $K_{\text{controls}}(r)$ are Ripley's K function for the cases and the controls respectively. The D function is defined as the difference between $K_{\text{cases}}(r)$ and $K_{\text{controls}}(r)$:

$$D(r) = K_{\text{cases}}(r) - K_{\text{controls}}(r) \quad (2)$$

Whether D is positive or negative depends on whether the cases are more aggregated or more dispersed than the controls. The controls constitute a benchmark capturing spatial heterogeneity: D reveals concentration or dispersion relative to the controls. Note that neither the interpretation of D values nor the occurrence of peaks is straightforward. D will only detect the occurrence of statistically significant concentration or dispersion within a range of distances.

4.2. Practical computation

The value of D is computed in two steps. First, K is calculated for the case population (firms in the chosen sector) and then for the controls (all firms outside the chosen sector). D is the difference between these two values of K . Confidence intervals are generated by using the Monte Carlo methods. Following the method originally used by Diggle and Chetwynd (1991), we keep the firms' location unchanged and we randomly affect the actual set of sectors to existing firms. The null hypothesis is that the sector of activity is random, while the location of firms is fixed.

4.3. Empirical applications

To apply the D function, we consider two areas. In the first step, to compare results from the L and D functions, we use the same area as previously; i.e. Paris and its suburbs. In the second stage, we use a larger French rectangular area measuring 550×630 km. We refer to this area as 'France' for simplicity. Note that it was impossible to use the whole of France because of border-effect corrections.

4.3.1. The Paris area

By 1 km steps from 1 km to 20 km (above, edge effects are too important, see Diggle and Chetwynd, 1991, p. 1158), we compute the D function for the Paris area. For each sector of activity, D_{\min} and D_{\max} represent the null hypothesis of a random distribution at a confidence level of 5%, generated from 100 simulations.

Table 1. Results of the *D* function for each sector of activity in Paris area

Sector of activity	Number of firms	Significant concentration	Significant dispersion	Significant (<i>D</i> value)
C1: Clothing and leather	1,011	All distances	—	5 km (976)
C2: Printing and publishing	1,192	All distances	—	9 km (407)
C3: Pharmacy and perfumery	318	0–1 km 7–20 km	—	0 km (15) 15 km (89)
C4: Home equipment	333	All distances	—	7 km (142)
D0: Motor vehicles	96	—	1–20 km	10 km (–246)
E1: Naval, aeronautic and rail construction	57	—	1–20 km	13 km (–296)
E2: Mechanical engineering	611	—	All distances	10 km (–383)
E3: Electric and electronic products	490	—	All distances	9 km (–291)
F1: Mineral products	223	—	All distances	11 km (–281)
F2: Textiles	181	All distances	—	9 km (324)
F3: Wood and paper products	182	—	All distances	14 km (–148)
F4: Chemical, rubber and plastics industry	345	—	All distances	10 km (–240)
F5: Metallurgy and metal transformation	442	—	All distances	9 km (–329)
F6: Electrical engineering	258	—	All distances	9 km (–301)

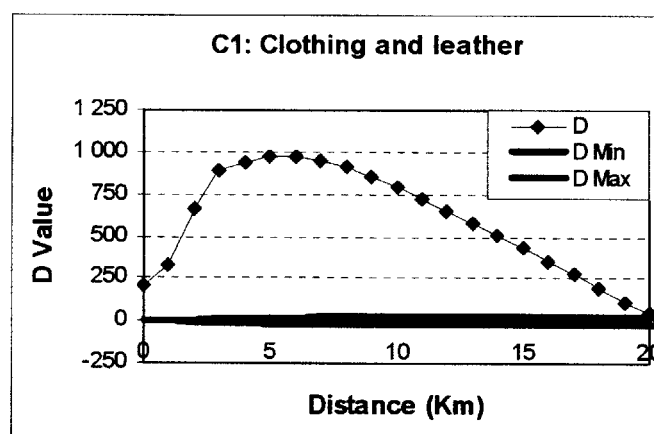
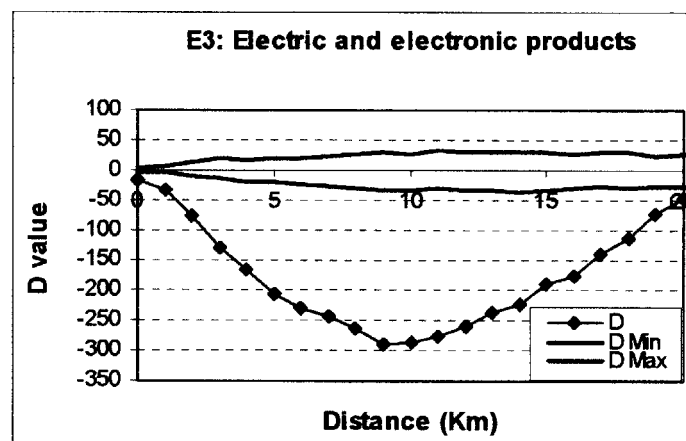
**Figure 8.** *D* function for sector C1 around Paris (1,011 firms).**Figure 9.** *D* function for sector E3 around Paris (490 firms).

Table 1 reports the D function for each sector of activity. We only present D functions for clothing and leather (Figure 8) and electric and electronic products (Figure 9). The former sector is significantly concentrated and the latter is significantly dispersed.

Computing D for a wide range of radius values detects significant concentration or significant dispersion *at different scales*. In summary, we found nine significantly dispersed sectors and five significantly concentrated sectors. Moreover, the distance at which there is significant dispersion or concentration differs according to the activity sector (see Table 1). Some sectors have only one significant peak while others have two which indicates multiple levels of concentration. Distance-based methods indicate the distance at which the maximum concentration or dispersion arises. Nevertheless, as already suggested, the value of the D function cannot be interpreted directly (unlike the L function). D values only show ‘how far’ the K function for the chosen sector is from the K function for all other manufacturing.

As expected, evaluating geographic concentration by using all manufacturing industry as a benchmark leads to different conclusions from those obtained with CSR as the benchmark. With respect to the D function, all sectors less concentrated than aggregate manufacturing are considered dispersed.

4.3.2. France

Our data cover 25,186 productive business units, i.e. around three-fifths of French productive business units registered in the EAE database. We chose to compute the D function at 1 km intervals from 1 km to 10 km, then at 5 km intervals from 10 km to 100 km and then at 100 km intervals from 100 km to 400 km. The confidence interval for the null hypothesis is constructed from 20 simulations.

Table 2 reports D functions for each sector, and Figure 10 and Figure 11 illustrate the D functions for clothing and leather (C1) and electric and electronic products (E3).

The D functions for the 14 manufacturing sectors reveal that four sectors are significantly concentrated, four sectors are dispersed, six industries are significantly concentrated or dispersed depending on the geographic scale and no sector is neither significantly concentrated nor dispersed at any distance.

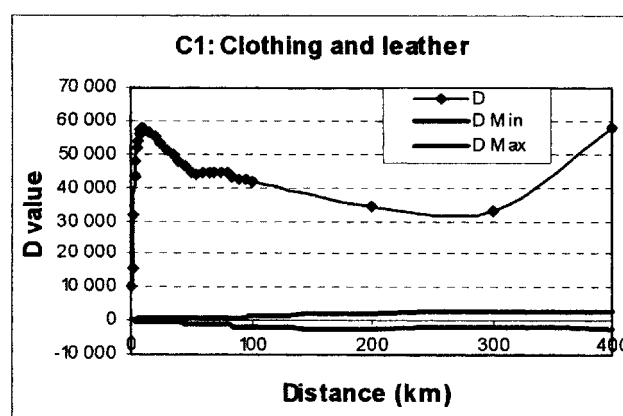
A brief overview of the empirical results show at first some important differences between sectors of activity for distances at which there is significant concentration or dispersion. For example, firms belonging to the chemical, rubber, and plastics industry (F4) are significantly dispersed below 100 km and over 250 km, whereas firms in clothing and leather (C1) are spatially concentrated whatever scale is considered. Second, comparing the two areas shows that industries do not necessarily have the same location patterns in an urban area (around Paris) as in the whole of ‘France’. For instance, consider Figure 9 and Figure 11: firms in electric and electronic products around Paris are dispersed, but elsewhere in ‘France’ they are concentrated. Motor vehicles (D0) and metallurgy and metal transformation (F5) are similar examples.

5. Comparing Ripley’s K function to Duranton and Overman’s K function

In spatial statistics, Ripley’s functions are among the frequently used measures to determine spatial concentration by analysing the distribution of points. They are currently applied in many fields such as forestry (Moeur, 1993; Pélissier and

Table 2. Results of the D function for each sector of activity in 'France'

Sector of activity	Number of firms	Significant concentration	Significant dispersion	Significant peak (D value)
C1: Clothing and leather	2,223	All distances	—	10 km (57,873) 400 km (58,013)
C2: Printing and publishing	2,604	All distances	—	300 km (73,417)
C3: Pharmacy and perfumery	748	All distances	—	200 km (114,752)
C4: Home equipment	1,524	—	15–280 km	200 km (–8,424)
D0: Motor vehicles	560	150–390 km	1–100 km	30 km (–10,118) 300 km (9,165)
E1: Naval, aeronautic and rail construction	218	30–60 km	3–10 km 120–400 km	8 km (–1,867) 45 km (6,955) 400 km (–20,442)
E2: Mechanical engineering	3,324	—	All distances	85 km (–8,381) 400 km (–15,559)
E3: Electric and electronic products	1,610	All distances	—	55 km (24,576) 100 km (24,234)
F1: Mineral products	2,394	—	All distances	200 km (–39,326)
F2: Textiles	1,110	0–8 km 55–280 km	25–35 km	4 km (2,195) 30 km (–3,104) 200 km (21,545)
F3: Wood and paper products	1,292	—	All distances	400 km (–36,368)
F4: Chemical, rubber and plastics industry	2,158	150–220 km	0–100 km 250–400 km	30 km (–12,161) 200 km (4,535) 400 km (–14,469)
F5: Metallurgy and metal transformation	2,913	300–400 km	0–230 km	35 km (–14,179) 400 km (8,454)
F6: Electrical engineering	1,037	20–330 km	0–10 km	9 km (–2,754) 200 km (19,155)

**Figure 10.** D functions for sector C1 in France (2,223 firms).

Goreaud, 2001), ecology (Cole and Syms, 1999) and geographical epidemiology (Kingham et al., 1995; Gatrell and Bailey, 1996; Jones et al., 1996). Duranton and Overman (2002) developed their own K function, apparently independently of Ripley, since Ripley is not cited. Their specification of K is quite similar to Ripley's, but noticeably different. They define $K(r)$ as the average number of points located at (rather than up to) the distance r from each firm. Like Ripley's K , the function is computed in steps. It is

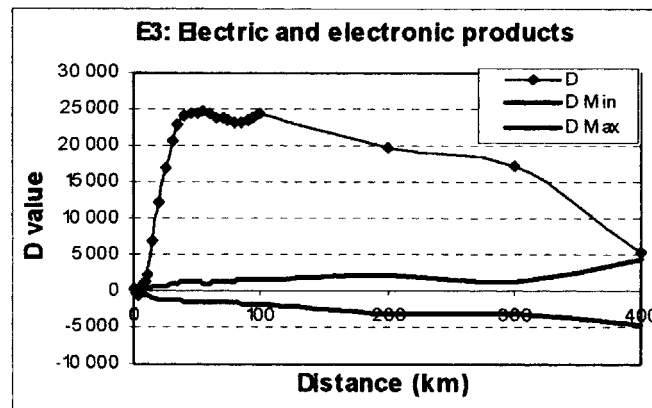


Figure 11. D function for sector E3 in France (1,610 firms).

smoothed for continuous results and normalized so that its sum between zero and the infinite equals one. We did not follow this procedure for three reasons. First, the mathematical background and experience acquired from 25 years of research (albeit rare in economics) is a decisive reason to work on improving the existing tools rather than restart from the beginning. Second, Duranton and Overman's K values cannot be interpreted owing to the specification of the function. Concentration and dispersion are detected but not quantified. Third, the construction of the function is problematic. Localization is evaluated by comparing the K function to the random distribution of an equivalent set of firms. Since both K and the benchmark sum to unity, if K exceeds the benchmark for a range of distances, it will necessarily be below the benchmark elsewhere. The authors deal with this problem by limiting the distance considered (up to 180 km for an empirical application on the whole of Great Britain). This distance range is economically pertinent. However, the unsolved question is whether measured concentration at these distances is meaningful, and whether dispersion at greater distances is simply a mathematical conjecture, or the opposite; i.e. observed concentration at distances of interest may simply be a consequence of actual dispersion at long distances. Despite these issues, Duranton and Overman's K function has valuable advantages over those of Ripley. First, firm size can be accounted for and industrial concentration controlled for. This represents the next step in the development of Ripley's functions. The second advantage is that edge-effect corrections are unnecessary, which simplifies computation considerably. The number of neighbors of points close to borders is underestimated, but for the null hypothesis also. The counterpart is that the values of the function are meaningless.

Given this brief comparison, none of the tools is yet completely satisfactory. We suggest that Ripley's functions have a much more solid background, but should be developed to incorporate at least the qualities of Duranton and Overman's K function.

6. Conclusion

This paper has introduced some new methods in economics to determine geographic concentration.

The proposed methods provide a more comprehensive approach to measuring concentration by analysing simultaneously the spatial distribution of firms at different

geographic scales. As underlined by Gatrell et al. (1996, p. 258), a priori, '*it seems sensible to use methods that preserve the original continuous setting of the data*'. In spite of this, our proposed measure does not supersede traditional indices but rather complements and enhances those already existing. Our measure complements more than substitutes existing measures for three reasons. First, Ripley's function fails to take into account the individual characteristics of firms (since every firm is considered as a point, regardless of its size), and its benchmark is not the most pertinent. Second, data requirements and computing intensity limit its practical usefulness. However, concentration measures can be improved using this method since they will not be restricted to a geographical level; i.e. the *exact* spatial concentration or dispersion scale can be determined.

Nevertheless, empirical results obtained using the proposed statistical approach pave the way for future research in geographic economics. The development of a distance-based method to assess the geographic concentration of activity is likely to be used widely in spatial economics, even though the technique requires improvement.

Acknowledgements

We wish to thank the referees for helpful comments, François Goreaud for his figures and comments, Françoise Maurel for useful suggestions, and Paul Baker for sound advice.

References

- Amiti, M. (1997) Specialisation patterns in Europe. Discussion Paper 363, Centre For Economic Performance, London.
- Arbia, G. (1989) *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Dordrecht: Kluwer.
- Barff, R.A. (1987) Industrial clustering and the organization of production: a point pattern analysis of manufacturing in Cincinnati, Ohio. *Annals of the Association of American Geographers*, 77(1): 89–103.
- Besag, J.E. (1977) Comments on Ripley's paper. *Journal of the Royal Statistical Society B*, 39(2): 193–195.
- Brühlhart, M. (2001) Evolving geographical concentration of European manufacturing industries. *Weltwirtschaftliches Archiv*, 137(2): 215–243.
- Brühlhart, M., Torstensson, J. (1996) Regional integration, scale economies an industry location in the European Union. Research Paper 1435, Centre for Economic Policy Research, London.
- Callejón, M. (1997) Concentración geográfica de la industria y economías de aglomeración. *Economía Industrial*, 0(5): 61–68.
- Cole, R.G., Syms, C. (1999) Using spatial patterns analysis to distinguish causes of mortality: an example from kelp in north-eastern New Zealand. *Journal of Ecology*, 87(6): 963–972.
- Cressie, N.A. (1993) *Statistics for Spatial Data*. New York: Wiley.
- Devereux, M.P., Griffith, R., Simpson, H. (1999) The geographic distribution of production activity in the UK. Working Paper 26/99, The Institute for Financial Studies, London.
- Diggle, P.J. (1983) *Statistical Analysis of Spatial Point Patterns*, London: Academic Press.
- Diggle, P.J., Chetwynd, A.G. (1991) Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, 47: 1155–1163.
- Duranton, G. (1997) La nouvelle économie géographique: agglomération et dispersion. *Economie et Prévision*, 131(5): 1–24.
- Duranton, G., Overman, H.G. (2002) Testing for location using micro-geographic data. Discussion Paper 3379, Centre for Economic Policy Research, London.

- Ellison, G., Glaeser, E.L. (1997) Geographic concentration in US manufacturing industries: a dartboard approach. *Journal of Political Economy*, 105(5): 889–927.
- Gatrell, A.C., Bailey, T.C. (1996) Interactive spatial data analysis in medical geography. *Social Science & Medicine*, 42(6): 843–855.
- Gatrell, A.C., Bailey, T.C., Diggle, P.J., Rowlingson, B.S. (1996) Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, 21: 256–274.
- Goreaud, F. (2000) Apports de l'analyse de la structure spatiale en forêt tempérée à l'étude de la modélisation des peuplements complexes. Ph.D. dissertation, ENGREF, Nancy.
- Haaland, J.I., Kind, H.J., Midelfart-Knarvik, K.H., Torstensson, J. (1999) What determines the economic geography of Europe? Discussion Paper 2072, Centre for Economic Policy Research, London.
- Head, K., Mayer, T., Ries, J. (2002) The geographic concentration of FDI in Asia. In J.H. Dunning and J.-L. Mucchielli (eds), *Multinational Firms: The Global and the Local Dilemma*. London: Routledge.
- Houdebine, M. (1999) Concentration géographique des activités et spécialisation des départements français. *Economie et Statistique*, 326–327(6–7): 189–204.
- Jayet, H., Puig, J.-P., Thisse, J.-F. (1996) Enjeux économiques de l'organisation du territoire. *Revue d'Economie Politique*, 106(1): 127–158.
- Jones, A.P., Langford, I.H., Bentham, G. (1996) The application of K-function analysis to the geographical distribution of road traffic accident outcomes in Norfolk, England. *Social Science & Medicine*, 42(6): 879–885.
- Kingham, S.P., Gatrell, A.C., Rowlingson, B. (1995) Testing for clustering of health events within a geographical information system framework. *Environment and Planning A*, 27(5): 809–821.
- Krugman, P. (1991) *Geography and Trade*. Cambridge, MA: MIT Press.
- Lösch, A. (1940) *The Economics of Location*. Yale: Yale University Press (translation).
- Marshall, A. (1920) *Principle of Economics*. London: Macmillan.
- Maurel, F., Sédillot, B. (1999) A measure of the geographic concentration in French manufacturing industries. *Regional Science and Urban Economics*, 29(5): 575–604.
- Midelfart-Knarvik, K.H., Overman, H.G., Redding, S.J., Venables, A.J. (2000) The location of European industry. Economic Paper 142, European Commission Brussels.
- Moeur, M. (1993) Characterizing spatial patterns of trees using stem-mapped data. *Forest Science*, 39(4): 756–775.
- Mucchielli, J.-L., Puech, F. (2001) Location of multinational firms: an application of the Ellison and Glaeser index to French firms in Europe, Miméo, TEAM University of Paris I-CNRS.
- Ottaviano, G.I.P., Puga, D. (1998) Agglomeration in the global economy: a survey of the new economic geography. *The World Economy*, 21(6): 707–731.
- Pélissier, R., Goreaud, F. (2001) A practical approach to the study of spatial structure in simple cases of heterogeneous vegetation. *Journal of Vegetation Science*, 12(1): 99–108.
- Ripley, B.D. (1976) The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13: 255–266.
- Ripley, B.D. (1977) Modelling spatial patterns. *Journal of the Royal Statistical Society B*, 39(2): 172–212.
- Rosenthal, S.S., Strange, W.C. (2001) The determinants of agglomeration. *Journal of Urban Economics*, 50(2): 191–229.
- Rowlingson, B.S., Diggle, P.J. (1993) Splan: spatial point pattern analysis code in S-Plus. *Computers & Geosciences*, 19(5): 627–655.
- Sweeney, S.H., Feser, E.J. (1998) Plant size and clustering of manufacturing activity. *Geographical Analysis*, 30(1): 45–64.
- Upton, G., Fingleton, B. (1985) *Spatial Data Analysis by Example. Volume 1. Point Pattern and Quantitative Data*. New York: Wiley.
- Valeyre, A. (1993) Mesures de dissemblance et d'inégalité interrégionales: principes, formes et propriétés. *Revue d'Economie Régionale et Urbaine*, 1: 17–53.

Appendix 1

Ripley's K function: mathematical presentation

Considering a spatial distribution of points, we define the position of each plot (firms in our study) by its coordinates (x, y) . The distribution of firms can be considered as a realization of a random process called a point pattern. Point patterns are defined by two fundamental properties. The first-order characteristic of a point pattern is its density, denoted by $\lambda(x, y)$. The density is $\lambda(x, y) = \lim_{dS \rightarrow 0} N(dS)/dS$ where dS is a small surface centered on (x, y) and $N(dS)$ is the number of points inside it. The density around (x, y) is the number of points divided by the area, when the area tends to zero. The density determines the probability, denoted by $P(dS)$, of the occurrence of at least one point in the elementary area dS around (x, y) . The value of $P(dS)$ is:

$$P(dS) = dS\lambda(x, y) \quad (A1)$$

Proof: Calculation of the first-order property of a homogeneous point process

Consider the studied domain area, denoted by D , and an elementary area, denoted by dS , as shown in Figure A1.

The number of points in the domain is denoted by N .

N/D is the average density, denoted by λ .

The number of points located in area dS follows a Poisson distribution with parameter λdS .

Thus, the probability of finding k points in this area is:

$$P(k) = e^{-\lambda dS} \frac{(\lambda dS)^k}{k!}$$

The probability of finding no points is consequently:

$$P(0) = e^{-\lambda dS}$$

We define dS as small enough for λdS to be small compared to 1. The first few terms of Taylor's expansion of the exponential function yields:

$$P(0) = 1 - \lambda dS + \frac{(\lambda dS)^2}{2!} - \frac{(\lambda dS)^3}{3!} + \dots \approx 1 - \lambda dS$$

We use the first-order approximation:

$$P(0) = 1 - \lambda dS$$

The probability of finding at least one point in area dS , denoted by $P(dS)$ is:

$$1 - P(0) = \lambda dS$$

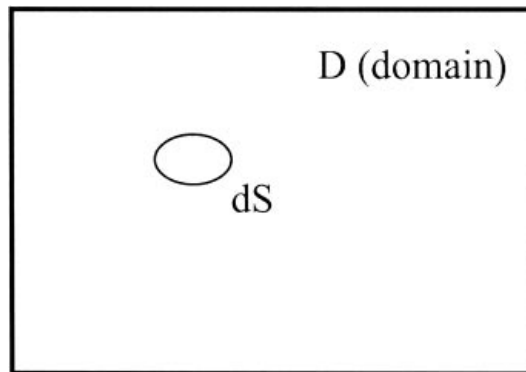


Figure A1. Elementary area.

This is the expression of the first-order characteristic of the point process:

$$P(dS) = \lambda dS$$

Notice that, for any size S :

$$\lambda S = \frac{NS}{D}$$

That is, λS is the number of points expected in the area S .

This result holds provided dS is small enough, since $dS\lambda(x, y)$ must be less than unity whatever the value of $\lambda(x, y)$. In the case of a homogeneous distribution, $\lambda(x, y)$ is constant; i.e. $P(dS) = \lambda dS$. In what follows, we consider *only* homogeneous distributions.

Note that, in the context of a random and independent distribution, the number of points in an area s is *not* constant but follows a Poisson distribution with parameter λs . A constant number of points implies a regular (not a random) distribution.

The second-order characteristic of a point pattern, denoted by $g((x_1, y_1), (x_2, y_2))$, determines the joint probability of the presence of at least one point in each elementary area centered on (x_1, y_1) and (x_2, y_2) , denoted by $P(dS_1, dS_2)$:

$$P(dS_1 dS_2) = P(dS_1)P(dS_2)g((x_1, y_1)(x_2, y_2)) \quad (\text{A2})$$

By making the assumption of isotropy, $g(\cdot)$ only depends on the distance r between area couples. Thus, we denote this by $g(r)$, which is sometimes called the pair correlation function (Cressie, 1993). In the case of an independent process, the probability $P(dS)$ of the occurrence of a point is independent of the position of other points, so $g(r)$ is equal to 1. At a given radius r , concentration is characterized by $g(r) > 1$; i.e. the joint probability of the presence of two points is greater than the product of individual probabilities. However, a regular distribution, which implies dispersion, is characterized by $g(r) < 1$.

To quantify concentration or dispersion, we consider each point and count its neighbors within the distance r . The average number of neighbors is compared to the expected number, $\lambda\pi r^2$. Points are aggregated if there are more neighbors than expected. Ripley (1977) linked this intuitive approach to $g(\cdot)$. Ripley's K function describes the spatial distribution of a point process. $K(r)$ is defined as the number of neighbors divided by the average density. For each point of the considered subplot, the expected number of points in a circle of radius r is $\lambda K(r)$. Under the assumptions of homogeneity (λ is constant) and isotropy, $g((x_1, y_1), (x_2, y_2))$ depends only on the distance between the two points (x_1, y_1) and (x_2, y_2) . Ripley (1977) established that:

$$K(r) = \int_{\rho=0}^r g(\rho) 2\pi\rho d\rho \quad (\text{A3})$$

In the case of an independent process, $g(r)$ is equal to 1, so the expected number of points in a circle of radius r is $\lambda\pi r^2$, and hence $K(r) = \pi r^2$. This value is used as a benchmark: i.e. $K(r) > \pi r^2$ indicates concentration because it is equivalent to the area-weighted average value of $g(r)$ being greater than 1.

Finally, we evaluate concentration and dispersion by counting the average number of neighbors for each point in a circle of radius r :

- in an independent distribution of points, the number of neighbors will be $\lambda\pi r^2$, so $K(r) = \pi r^2$;
- in the case of concentration, K will be greater than πr^2 ; and
- in the case of dispersion, K will be lower than πr^2 .

K is dimensionally homogeneous to a surface. Its value is the surface of the circle in which the same number of points would have been observed in an independent distribution. For instance, $K(10) = 5$ implies that as many points can be found in a circle of radius 10 as there would be in a circle of radius 15 in the case of a homogeneous point process.

Ripley's function should *not* only be understood as a mathematical formulation of an intuitive result. It allows consideration of apparent heterogeneity, such as aggregates, as the result of attraction strengths (the second-order property of the point process) within a homogeneous distribution (density, the first-order property, is constant).

Appendix 2

Edge-effect correction

To correctly evaluate the geographic concentration using K or L functions, border-effects correction should be incorporated. To illustrate this, consider the most important rectangular area we can take in France (of dimension 550×630 km). Remember that with K or L functions, boundaries of the spatial area under study should not be too uneven so that edge effects can be easily corrected.

Figure A2 shows the bias in our results without edge-effect correction (computed for all manufacturing sectors). The L values should tend to zero at long distances. This bias increases with distance. The number of points included in a circle around each point is not underestimated when the circle is small relative to the area under study. This bias persuaded us to give up studying the distribution of firms in the whole of France to work on a simple shape area instead.

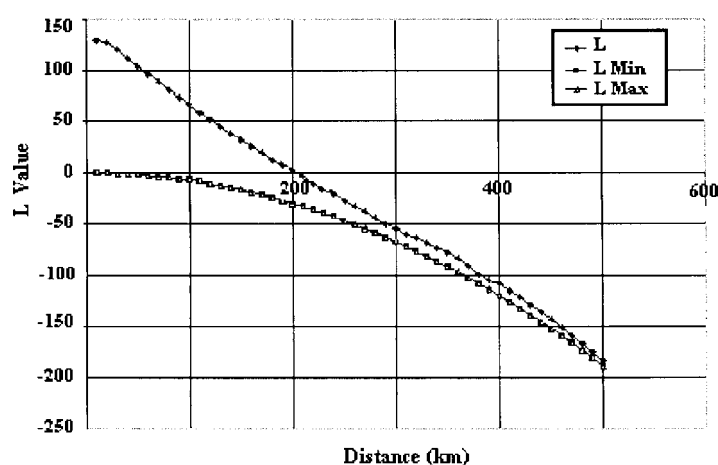


Figure A2. L function without edge-effect correction (L_{\min} and L_{\max} are confused).

