



Universidade de Aveiro Departamento de Biologia
2017

**LAURA INÊS
HENRIQUES
ANTÃO**

**EFEITOS DE ESCALA ECOLÓGICA EM
PADRÕES DE BIODIVERSIDADE**

**EFFECTS OF ECOLOGICAL SCALING ON
BIODIVERSITY PATTERNS**



**LAURA INÊS
HENRIQUES
ANTÃO**

**EFEITOS DE ESCALA ECOLÓGICA EM
PADRÕES DE BIODIVERSIDADE**

**EFFECTS OF ECOLOGICAL SCALING ON
BIODIVERSITY PATTERNS**

Tese apresentada à Universidade de Aveiro, em regime de co-tutela com a Universidade de St Andrews, para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Biologia e Ecologia das Alterações Globais, realizada sob a orientação científica da Doutora Maria Dornelas da *School of Biology, University of St Andrews*, Professor Amadeu Soares do Departamento de Biologia da Universidade de Aveiro, e Professora Anne Magurran da *School of Biology, University of St Andrews*.

Thesis presented to the University of Aveiro, in co-tutelle with the University of St Andrews, in order to fulfil the necessary requirements to obtain the degree of Doctor of Biology and Ecology of Global Change, carried out under the scientific supervision of Doctor Maria Dornelas from the School of Biology, University of St Andrews, Professor Amadeu Soares from the Biology Department, Universidade de Aveiro, and Professor Anne Magurran from the School of Biology, University of St Andrews.

Apoio financeiro do POCTI no âmbito do III Quadro Comunitário de Apoio.

Apoio financeiro da FCT e do FSE no âmbito do III Quadro Comunitário de Apoio.

o júri

presidente

Prof. Doutor Luís Filipe Pinheiro de Castro
professor catedrático da Universidade de Aveiro

Prof. Doutor Robert J. Whittaker
professor catedrático da Universidade de Oxford, Reino Unido

Doutor Luís António da Silva Borda de Água
investigador convidado da Universidade do Porto

Doutor Ulisses Manuel de Miranda Azeiteiro
professor associado com agregação da Universidade de Aveiro

Doutora Maria Azeredo de Dornelas
professora associada da Universidade de St Andrews, Reino Unido

agradecimentos

I cannot thank Maria Dornelas enough. With tireless support, invaluable advice, and untiring trust in me, Maria pushed me always forward and always to be better. I could never have hoped for a better supervisor or a better model. I am so grateful and proud to be her student. A huge thank you to Anne Magurran, for her wise guidance and support, and for an amazing sense of showing where the questions lie. I am also grateful to Amadeu Soares, for taking a chance on me and allowing me to start this project.

A huge thank you to Sean Connolly and Brian McGill, for being willing to discuss my results with me and contributing to my understanding of biodiversity patterns. I am very grateful to all the scientists and participants involved in data collection and data provision, and to all the data providers and their funders for making their data available. I also want to thank the online repositories OBIS, Ecological Data Wiki and GBIF. Without their effort and hard work, I would not have been able to use so many different high quality data in my analyses.

A massive thank you to all the friends who have relentlessly supported me: Alessandra, Inês, Sara, Maria, Miguel (B.), Maria João, Esme, Viviana, Susana (X.), Miguel (N.) and Susana (P.). In a way or another, closer or further away, your support, patience and kindness have helped me during this time. A more general thank you to all the people who were kind and generous to me through this long time (extended family and friends), who were willing to listen to me and to share their experiences, including all the other PhD students in the HMB – it helped to keep me grounded and sane. I also want to thank (and include anyone I might have not mention above) everyone who has encouraged me in this adventure, and finally also everyone who showed any interest in knowing “How is the PhD going?” Although, a “no-no” question for PhD students in the process, a showing of concern and interest for everyone else in the real world.

I want to say a massive special thank you to my Mother, for being the best mom, for always being there and for her unwavering belief in me. I also want to thank my Grandmother; their support and care have always accompanied and comforted me.

I have no words to thank Nuno, my love, my best friend, best husband, most supportive, most patient and most understanding companion on this long journey. You always believe in me, most importantly you believe in me when I don't. Thank you for dreaming with me, or better said, for making me dream and look beyond. This thesis is for Nuno and for my Mother; I would not be here without them.

Finally, I am most grateful to Universidade de Aveiro and to Fundação para a Ciência e Tecnologia, Portugal, for financial support during the PhD.

palavras-chave

Biodiversidade; Estrutura das comunidades; Escala espacial; Padrões de diversidade; Distribuição das abundâncias relativas das espécies (SADs); Macroecologia; Lognormal; Logseries; Multimodalidade; Amplitude taxonômica; Heterogeneidade ecológica; β diversidade; Composição específica; Dissimilaridade; *Turnover* espacial; *Nestedness* (Aninhamento); Dependência da escala; Agregação intraespecífica; Agregação espacial de espécies; Teorias unificadas

resumo

A biodiversidade é determinada por uma miríade de processos complexos que actuam a escalas diferentes. Face às actuais taxas de perda e alteração da biodiversidade, é vital melhorar a nossa compreensão da estrutura subjacente das comunidades ecológicas. Esta tese focou-se na análise de *Species Abundance Distributions* (SAD; Distribuição das abundâncias relativas das espécies), enquanto medida sintética de biodiversidade e da estrutura das comunidades, e de padrões de Beta (β) diversidade, enquanto medida de descrição da variação espacial na composição específica das comunidades. Os efeitos de escala nestes dois padrões de biodiversidade foram sistematicamente avaliados, analisando uma grande variedade de comunidades, incluindo diferentes taxa e habitats, dos reinos terrestre, marinho e água doce. O conhecimento sobre as propriedades de escala dos padrões de abundância e de composição específica das comunidades deve ser totalmente integrado na investigação da biodiversidade, no sentido de a podermos compreender melhor, bem como aos processos que a sustentam, desde escalas locais à escala global.

As SADs descrevem a abundância relativa das espécies presentes numa comunidade. Apesar de serem tipicamente descritas por distribuições unimodais, como a logseries ou a lognormal, SADs empíricas podem também exibir várias modas. No entanto, a existência de múltiplas modas em SADs tem sido largamente ignorada, sendo normalmente assumida como um padrão raro ou atribuído a erros de amostragem. Desta forma, a frequência de multimodalidade em SADs é desconhecida, bem como os factores que podem levar à sua ocorrência. Nesta análise, efectuei a primeira avaliação empírica global da frequência de multimodalidade, analisando várias comunidades de diferentes taxa, habitats e extensões espaciais. Usando um método melhorado que combina dois critérios de selecção de modelos, estimei (conservadoramente) que cerca de 15% das comunidades analisadas eram multimodais com grande suporte. Além disso, demonstrei que a multimodalidade é mais comum em comunidades com maior extensão espacial e com maior diversidade taxonômica (isto é, comunidades filogeneticamente mais diversas, uma vez que a diversidade taxonômica foi medida como o número de famílias). Estes resultados sugerem uma ligação entre SADs multimodais e heterogeneidade ecológica, aqui amplamente definida para incorporar a variabilidade espacial, ambiental, taxonômica e funcional dos sistemas ecológicos.

Ainda não possuímos uma compreensão empírica de como a escala espacial afecta a forma das SADs. Nesta análise, estabeleci um gradiente de escala espacial abrangendo várias ordens de magnitude, começando por decompor a extensão espacial total de várias comunidades em secções menores. Este gradiente foi usado para realizar uma análise exploratória de como a forma das SADs é afectada pela área amostrada, riqueza específica, abundância total e diversidade taxonómica. Mudanças claras na forma das SADs podem fornecer informações sobre mecanismos ecológicos e espaciais relevantes que afectam a estrutura das comunidades. Esta análise demonstrou um efeito claro da área, riqueza específica e diversidade taxonómica na forma das SADs, enquanto que a abundância total não exibiu um efeito direccionado. Estes resultados apoiam as conclusões da análise anterior, mostrando uma maior prevalência de SADs multimodais para áreas maiores e para comunidades mais diversas taxonomicamente. Adicionalmente, estes resultados sugerem que os padrões de agregação espacial das espécies influenciam a forma das SADs ao longo do gradiente espacial. Por outro lado, esta análise identificou diferenças sistemáticas relativamente às previsões de duas importantes teorias macroecológicas para as SAD a escalas diferentes, especificamente o facto de a logseries apenas ter sido seleccionada para escalas menores e quando a riqueza específica e o número de famílias eram proporcionalmente muito menores do que para a extensão total.

A β diversidade quantifica a variação na composição específica entre locais. Apesar de ser um componente fundamental da biodiversidade, conhecimento sobre a variação das suas propriedades com a escala espacial ainda é escasso. Nesta análise, testei se dois tipos conceptuais de β diversidade apresentam variação sistemática com a escala, considerando também explicitamente os dois componentes de β diversidade: *turnover* e *nestedness* (aninhamento) – substituição de espécies vs diferenças na riqueza específica entre locais, respectivamente. Efectuei a primeira análise empírica de padrões de escala de β diversidade para diferentes taxa, revelando que as curvas de escala são notavelmente consistentes para as comunidades analisadas. A β diversidade total e a componente de *turnover* exibem um declínio segundo uma *power law* com o logaritmo da área, enquanto a componente de *nestedness* é basicamente insensível às mudanças de escala. Relativamente à análise do declínio da similaridade com a distância geográfica, enquanto a área amostrada afectou significativamente os valores de dissimilaridade total, as taxas de mudança na similaridade foram consistentes para grandes variações entre áreas amostradas. Finalmente, em ambas as análises, o *turnover* foi o principal contribuinte para as diferenças composicionais. Estes resultados sugerem que as espécies estão espacialmente agregadas ao longo das escalas espaciais analisadas (de locais a regionais). Adicionalmente, os resultados ilustram que mudanças substanciais na estrutura das comunidades podem ocorrer, apesar de a riqueza específica permanecer relativamente estável.

A análise sistemática e abrangente de SADs e de padrões de similaridade nesta tese identificou a escala espacial, a heterogeneidade ecológica e padrões de agregação espacial das espécies como componentes críticos subjacentes aos resultados encontrados. Esta investigação expandiu as escalas às quais tanto teorias que derivam SAD, como estudos de similaridade têm sido desenvolvidos e testados (desde *plots* locais a continentes). Estes resultados identificaram claros desvios face a duas importantes teorias macroecológicas para SAD a diferentes escalas. Adicionalmente, os resultados gerais desta tese indicam claramente que teorias unificadas da biodiversidade (ou assumindo um conjunto mínimo de pressupostos sintéticos) não são capazes de, por um lado, acomodar a variabilidade na forma das SADs a escalas espaciais diferentes aqui reportada, e, por outro lado, reproduzir totalmente os padrões de similaridade a todas as escalas espaciais. A incorporação de pressupostos mais realistas, ou a imposição de pressupostos dependentes da escala, pode revelar-se uma linha de investigação produtiva para as propriedades de escala das SADs e de padrões de similaridade, permitindo derivar novas previsões e melhorar a capacidade dos modelos teóricos em incorporar a variabilidade nos padrões de abundância e de similaridade a várias escalas.

keywords

Biodiversity; Community structure; Spatial scaling; Diversity patterns; Species abundance distribution; Macroecology; Lognormal; Logseries; Multimodality; Taxonomic breadth; Ecological heterogeneity; β diversity; Species composition; Dissimilarity; Spatial turnover; Nestedness; Scale dependence; Intraspecific aggregation; Species aggregation; Unified theories

abstract

Biodiversity is determined by a myriad of complex processes acting at different scales. Given the current rates of biodiversity loss and change, it is of paramount importance that we improve our understanding of the underlying structure of ecological communities. In this thesis, I focused on Species Abundance Distributions (SAD), as a synthetic measure of biodiversity and community structure, and on Beta (β) diversity patterns, as a description of the spatial variation of species composition. I systematically assessed the effect of scale on both these patterns, analysing a broad range of community data, including different taxa and habitats, from the terrestrial, marine and freshwater realms. Knowledge of the scaling properties of abundance and compositional patterns must be fully integrated in biodiversity research if we are to understand biodiversity and the processes underpinning it, from local to global scales.

SADs depict the relative abundance of the species present in a community. Although typically described by unimodal logseries or lognormal distributions, empirical SADs can also exhibit multiple modes. However, the existence of multiple modes in SADs has largely been overlooked, assumed to be due to sampling errors or a rare pattern. Thus, we do not know how prevalent multimodality is, nor do we have an understanding of the factors leading to this pattern. Here, I provided the first global empirical assessment of the prevalence of multimodality across a wide range of taxa, habitats and spatial extents. I employed an improved method combining two model selection tools, and (conservatively) estimated that ~15% of the communities were multimodal with strong support. Furthermore, I showed that the pattern is more common for communities at broader spatial scales and with greater taxonomic diversity (i.e. more phylogenetically diverse communities, since taxonomic diversity was measured as number of families). This suggests a link between multimodality and ecological heterogeneity, broadly defined to incorporate the spatial, environmental, taxonomic and functional variability of ecological systems.

Empirical understanding of how spatial scale affects SAD shape is still lacking. Here, I established a gradient in spatial scale spanning several orders of magnitude by decomposing the total extent of several datasets into smaller subsets. I performed an exploratory analysis of how SAD shape is affected by area sampled, species richness, total abundance and taxonomic diversity. Clear shifts in SAD shape can provide information about relevant ecological and spatial mechanisms affecting community structure. There was a clear effect of area, species richness and taxonomic diversity in determining SAD shape, while total abundance did not exhibit any directional effect. The results

supported the findings of the previous analysis, with a higher prevalence of multimodal SADs for larger areas and for more taxonomically diverse communities, while also suggesting that species spatial aggregation patterns can be linked to SAD shape. On the other hand, there was a systematic departure from the predictions of two important macroecological theories for SAD across scales, specifically regarding logseries distributions being selected only for smaller scales and when species richness and number of families were proportionally much smaller than the total extent.

β diversity quantifies the variation in species composition between sites. Although a fundamental component of biodiversity, its spatial scaling properties are still poorly understood. Here, I tested if two conceptual types of β diversity showed systematic variation with scale, while also explicitly accounting for the two β diversity components, turnover and nestedness (species replacement vs species richness differences). I provided the first empirical analysis of β diversity scaling patterns for different taxa, revealing remarkably consistent scaling curves. Total β diversity and turnover exhibit a power law decay with log area, while nestedness is largely insensitive to scale changes. For the distance decay of similarity analysis, while area sampled affected the overall dissimilarity values, rates of similarity were consistent across large variations in sampled area. Finally, in both these analyses, turnover was the main contributor to compositional change. These results suggest that species are spatially aggregated across spatial scales (from local to regional scales), while also illustrating that substantial change in community structure might occur, despite species richness remaining relatively stable.

This systematic and comprehensive analysis of SAD and community similarity patterns highlighted spatial scale, ecological heterogeneity and species spatial aggregation patterns as critical components underlying the results found. This work expanded the range of scales at which both theories deriving SAD and community similarity studies have been developed and tested (from local plots to continents). The results here showed strong departures from two important macroecological theories for SAD at different scales. In addition, the overall findings in this thesis clearly indicate that unified theories of biodiversity (or assuming a set of synthetic minimal assumptions) are unable to accommodate the variability in SADs shape across spatial scales reported here, and cannot fully reproduce community similarity patterns across scales. Incorporating more realistic assumptions, or imposing scale dependent assumptions, may prove to be a fruitful avenue for ecological research regarding the scaling properties of SAD and community similarity patterns. This will allow deriving new predictions and improving the ability of theoretical models to incorporate the variability in abundance and similarity patterns across scales.

Table of Contents

1. General Introduction	1
1.1 Biodiversity patterns	1
1.2 Matters of scale	3
1.3 Species Abundance Distributions.....	5
1.4 Compositional similarity (β diversity)	8
1.5 Thesis overview.....	11
2. Collecting Data.....	13
3. Multimodality in Species Abundance Distributions – improving the detection method	17
3.1 Introduction	17
3.2 Methods.....	20
3.2.1 Multimodal functions	20
3.2.2 Model Selection.....	22
3.2.3 Simulation Study	24
3.2.4 Parametric Bootstrap	25
3.3 Results	26
3.3.1 Simulation study.....	26
3.3.2 Parametric Bootstrap	27
3.4 Discussion	34
4. Multimodality in Species Abundance Distributions – empirical analyses.....	35
4.1 Methods.....	35
4.2 Results	37

4.3	Discussion	45
5.	Species Abundance Distributions across scales	49
5.1	Introduction	49
5.2	Methods	53
5.2.1	Empirical Data	53
5.2.2	Implementing the scale gradient	53
5.2.3	Model fitting and analysis	55
5.3	Results	57
5.4	Discussion	69
6.	Multiscale spatial patterns of β diversity	77
6.1	Introduction	77
6.2	Methods	80
6.2.1	β diversity scaling curves	82
6.2.2	Distance Decay of Similarity	82
6.3	Results	84
6.4	Discussion	90
7.	General Discussion	95
7.1	Multimodality and SAD shape across scales	97
7.2	β diversity and spatial scale	101
7.3	Contribution to Macroecology theory	103
7.4	Conclusions	106
8.	References	107

Appendix I.....	123
Appendix II.....	137
Appendix III.....	155
Appendix IV.....	161
Appendix V.....	165
Appendix VI.....	175
Appendix VII.....	185

List of Figures

Figure 2.1 Datasets contributed to BioTIME	15
Figure 3.1 Examples of random sampled communities for a logseries and mixtures of one, two and three lognormal Poisson distributions (1PLN, 2PLN and 3PLN, respectively).....	21
Figure 3.2 Logseries simulation results.....	29
Figure 3.3 1PLN simulation results.....	30
Figure 3.4 2PLN simulation results.....	31
Figure 3.5 3PLN simulation results.....	32
Figure 3.6 Likelihood Ratio test frequency distributions for 1PLN simulated communities.....	33
Figure 4.1 Sampling locations of the 117 empirical SADs and the model selected as best fit	39
Figure 4.2 Species abundance distributions of the multimodal empirical datasets	40
Figure 4.3 Model selection frequency <i>versus</i> spatial extent and taxonomic breadth	42
Figure 5.1 Schematic representation of the scale gradient	54
Figure 5.2 Effect of area sampled, species richness, total number of individuals and number of families on the best model selected.....	58
Figure 5.3 Comparison of empirical SADs and fitted models across the scale gradient.....	63
Figure 5.4 Scaling relationships with area for α diversity metrics	68
Figure 6.1 Schematic representation of the scale gradient for the β diversity analyses	81
Figure 6.2 β diversity scaling curves.....	86
Figure 6.3 Distance decay relationships with geographic distance	86

Figure V.1 Non-multimodal empirical species abundance distributions	165
Figure VI.1 Total extent SADs of the empirical datasets analysed in chapter 5.....	175
Figure VI.2 Comparison of empirical SADs and fitted models.....	176
Figure VII.1 Median β_{SOR} values across all the splitting trials	185
Figure VII.2 Variability of estimated DDS linear model slopes across the splitting trials.....	185

List of Tables

Table 3.1 Overall false positive and false negative frequencies for the simulation study	28
Table 4.1 Parametric bootstrap results for the empirical SADs	43
Table 4.2 Binomial and multinomial model results for the prevalence of multimodality.....	44
Table 5.1 Empirical data used in SAD and β diversity scaling analyses.....	56
Table 6.1 Nonlinear model fitting results for β diversity scaling curves.....	89
Table I.1 Empirical datasets information for the multimodality analysis	123
Table VI.1 Linear model fitting results for the Shannon index.....	184
Table VII.1 Distance decay of similarity linear model results.....	188
Table VII.2 Comparison of the DDS slopes and intercepts between scaling levels	190

List of Appendices

- I. Empirical datasets information.
- II. Data sources references and acknowledgements.
- III. Log-likelihood functions for mixtures of 1, 2 and 3 Poisson lognormal distributions (R code).
- IV. (A) Simulation study code. (B) Likelihood ratio test code.
- V. Non-multimodal empirical species abundance distributions.
- VI. Supplementary figures and tables from chapter 5.
- VII. Supplementary figures and tables from chapter 6.

1. General Introduction

1.1 Biodiversity patterns

Biodiversity is a multifaceted concept, encompassing the diversity of life at different scales. The Convention for Biological Diversity (1992) defined "Biological Diversity" as "*the variability among living organisms from all sources including, inter alia, terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems.*" More concisely, biodiversity can be simply defined as "the variety and abundance of organisms in a given place and time" (Magurran, 2005a). Understanding the processes underpinning biodiversity has long been, and remains one of the fundamental goals of ecology (Hutchinson, 1959; MacArthur, 1972; Whittaker, 1972; Magurran, 2004). Given current rates of biodiversity loss and of biodiversity change, and the extent of human impacts on global systems leading to the suggestion of defining a new geological age – the *Anthropocene*, there is an urgent need to better understand ecological systems, in order to best manage and conserve biodiversity (Butchart *et al.*, 2010; Magurran & Dornelas, 2010; Pereira *et al.*, 2010, 2012).

Ecological systems are extremely complex and dynamic, showing high variability in space and time. A community, defined as a set of species co-occurring in space and time (Fauth *et al.*, 1996), appears to be a logical unit to address questions regarding spatial and temporal variation of biodiversity (Magurran, 2004). Information about community structure incorporates both the number of species present and their relative abundances, i.e. species richness and species evenness, respectively. A plethora of diversity indices to characterize the structure of communities exist, with the overall goal to summarize the ecological complexity into univariate metrics (Magurran, 2004). However, because different metrics differ in the weight placed on richness and evenness components, different "measurements" of biodiversity can be obtained (Magurran, 2004, 2005a). Alternatively, analysing the distribution of species abundances of the (sampled) community retains more detailed information, thus providing a more integrated assessment of community structure (Magurran, 2004; McGill *et al.*, 2007). On the other hand, some practical issues may arise with this community concept, as taxonomic, spatial and temporal extent can vary greatly (Magurran, 2004, 2005a; McGill, 2011). For

instance, the proportion of rare species will depend on the definition of a specific community, the size of the area sampled, and the period of time surveyed (Magurran, 2004, 2005b). In addition, patterns of species diversity will also strongly depend on the spatial structure of occurrence of the species present in a community. Finally, the interactions between local and regional processes (e.g. metacommunity dynamics – local communities connected via dispersal) are also determinant in shaping biodiversity patterns at different scales (Hubbell, 2001; Leibold *et al.*, 2004; Ricklefs, 2008).

Despite the incredible diversity of ecological systems, some ecological patterns are so pervasive they have become ecological laws. Species-Area Relationships (SAR), describing the increase in number of species with area sampled, Species Abundance Distributions (SAD), which describe the uneven distribution of individuals among species, and the distance decay of compositional similarity between communities are among the most studied patterns in ecology, given their potential insights into community structure and patterns of diversity (Williams, 1943; Preston, 1948; MacArthur, 1972; Rosenzweig, 1995; Nekola & White, 1999; Hubbell, 2001; McGill *et al.*, 2007). Such general patterns support the idea that basic processes structuring ecological communities underpin these large-scale emergent patterns. Macroecology searches for general ecological patterns and processes at large spatial and temporal scales and across taxonomic groups. Macroecological research stems from recognizing that, on one hand, local processes alone are not able to fully explain the patterns of abundance and distribution of species, and that on the other hand, processes operating at larger scales also affect local communities (Brown & Maurer, 1989; Brown, 1995; Gaston & Blackburn, 2000).

1.2 Matters of scale

Both biodiversity patterns and the mechanisms driving them are inherently scale dependent (Wiens, 1989; Levin, 1992; Rosenzweig, 1995; Leibold *et al.*, 2004; McGill *et al.*, 2015). Different processes determine the distribution and abundance of species, and act upon ecological communities differently at different scales (MacArthur, 1972; Ricklefs, 1987; McGill, 2010a). Climate underpins the distribution of species on global to biogeographic realms scales, while both dispersal limitation and local environmental conditions determine which species can reach and establish in some areas, while others cannot. Species interactions, such as competition and predation act on a finer scale, with demographic implications for local populations (McGill, 2010a). As a consequence of these factors, representing both dispersal and niche processes, there is striking variation of community composition across space and time (MacArthur, 1972; Rosenzweig, 1995). Moreover, the anthropogenic drivers of biodiversity change are also scale dependent (Halpern *et al.*, 2015; Venter *et al.*, 2016).

There is no single adequate scale at which to describe ecological systems, thus being able to translate information from local to larger scales is one of the most relevant questions in ecology (Levin, 1992). Additionally, different organisms perceive scale differently, and the two components of scale, extent and grain, should be referenced to specific taxa (Wiens, 1989; Levin, 1992). It has long been recognized that the scale of observation affects biodiversity patterns, and should be taken into consideration when drawing conclusions about the underpinning processes that might explain the observed patterns. On the same vein, the fact that there are mismatches in some biodiversity patterns and trends can be attributed to scale (Wiens, 1989; Sax & Gaines, 2003; Pereira *et al.*, 2012; McGill *et al.*, 2015). Scaling rules provide one possible framework to describe and synthesize the patterns of species abundance and distribution in space, time and taxonomic groups. The pursuit of scaling relationships is central for ecological research, and can also be a useful approach to infer diversity patterns for scales or areas for which no data was collected or no information is available (e.g. using SARs for estimating species richness at larger areas, or species loss under habitat loss). There is a longstanding and ongoing endeavour to understand the scaling properties of biodiversity patterns (Wiens, 1989; Levin, 1992, 2000; Rosenzweig, 1995; White, 2007; Borda-de-Água *et al.*, 2012; Barton *et al.*, 2013). Numerous studies have been devoted to the scaling properties of species richness, with a thruphasic SAR emerging across scales (Williams, 1943; Rosenzweig, 1995; Harte *et al.*, 2009; Storch *et al.*, 2012). Species abundance distributions research has also been rooted in scaling issues (Fisher *et al.*, 1943; Preston, 1948). However, we still do not have a thorough

understanding of what factors determine SAD shape, and we lack empirical understanding of how SAD shape changes with scale. Finally, less attention has been dedicated to the scaling properties of compositional similarity metrics (or Beta (β) diversity), with a lack of both theoretical predictions and a general framework for describing its spatial scaling patterns (Koleff *et al.*, 2003; Gaston *et al.*, 2007; Barton *et al.*, 2013).

An integrative approach analysing different biodiversity patterns, across spatial scales and for different ecological communities, would provide more detailed information, on one hand, while assessing the generality of the findings on the other, thus allowing a better understanding of the mechanisms shaping biodiversity patterns. In this thesis, I analysed a broad range of communities, including different taxa and habitats, from the terrestrial, marine and freshwater realms, with an emphasis on the spatial and organizational scales suggested to underpin the variability of ecological patterns (Levin, 1992). I focused on Species Abundance Distributions, as a synthetic measure of biodiversity and community structure, and on β diversity patterns as a description of the spatial variation of species composition.

1.3 Species Abundance Distributions

One of the most fundamental patterns of species diversity is the uneven relative abundance of species (Fisher *et al.*, 1943; Preston, 1948; Magurran, 2004; McGill *et al.*, 2007). Species abundance distributions (SAD) describe the commonness and rarity of species in ecological communities and represent one of ecology's universal laws: most species in a community are rare, only a few are common. Plotted as a histogram of the number of species vs. number of individuals every community yields a characteristic and ubiquitous 'hollow curve' (McGill *et al.*, 2007). Empirical datasets consistently produce species abundance distributions that are hyperbolic on arithmetic scale and modal on a log-abundance scale. SADs are an important synthetic measure of biodiversity and community structure, being more informative than univariate indexes of diversity, and enabling comparisons of communities without species in common (Magurran, 2004; McGill *et al.*, 2007).

On a log-abundance scale, empirical SADs generally exhibit "logseries-like" or "lognormal-like" shapes. Hence, these classical distributions have been central to species abundance modelling (Magurran, 2004; McGill *et al.*, 2007). The two distributions differ especially in the proportion of rare species: high for the logseries, with a mode occurring for the singletons species (Fisher *et al.*, 1943), and lower for the lognormal, with a mode for species with intermediate abundances (Preston, 1948). Fisher *et al.*'s (1943) logseries is one of the first attempts to describe the relationship between the number of species and the respective number of individuals mathematically. The lognormal distribution was proposed by Preston (1948) and has been particularly prominent as a SAD model. Preston plotted species abundances in a log₂ scale, conveying the intuitive approach of doubling classes of abundance that he called 'octaves'. May (1975) proposed the lognormal distribution as a statistical expectation of the central limit theorem, i.e. SADs result from a random multiplicative process acting on species abundances. Although it was at first a statistical-based model, it has since been attributed biological explanations, particularly related to niche apportionment models (Magurran, 2004). Sugihara (1980) suggested that the lognormal is a consequence of species within a community sequentially dividing niche space, and later, Engen & Lande (1996) derived the lognormal distribution by modelling stochastic heterogeneous population dynamics.

SADs have played a major role in the development of theories of biodiversity and biogeography (Hubbell, 2001; McGill *et al.*, 2007; Harte *et al.*, 2008). Niche theory is the classic theoretical

explanation for the observed patterns in species abundance distributions (Hutchinson, 1959; MacArthur, 1960; Sugihara, 1980; Tokeshi, 1990). It is assumed that the abundance of a species somehow reflects its ability to compete for limited resources. Based on this assumption, SADs help to understand the common processes that determine the structure of communities. For almost a century, a series of species abundance models have been proposed trying to explain the universal SAD hollow curve (see McGill *et al.* 2007 for an extensive list): statistical models like the geometric series (Motomura, 1932), the logarithmic series (Fisher *et al.*, 1943) and the lognormal distribution (Preston, 1948); niche partitioning models, such as MacArthur's broken-stick model (1957), Sugihara breakage model (1980) and several models from Tokeshi (1990, 1993, 1996, 1999); spatial distribution models, such as fractal distribution (Harte *et al.*, 1999), multifractal (Borda-de-Água *et al.*, 2002) or continuum theory (McGill & Collins, 2003); and neutral models proposed by Caswell (1976), Bell (2000, 2001) and Hubbell (2001). All the models are linked, at least to some degree, with ecological mechanisms, or later gained some biological meaning (Magurran, 2004). More recently, constraint-based models, as opposed to process-based models, have also been proposed (Pueyo *et al.*, 2007; Harte *et al.*, 2008). SADs have played a pivotal role in niche *vs* neutral explanations for the maintenance of biodiversity (Bell 2001, Enquist *et al.* 2002, Hubbell 2001, McGill 2003b, Volkov *et al.* 2003, Dornelas *et al.* 2006, Volkov *et al.* 2007), with the same datasets sometimes being used as empirical support for and against each model (McGill *et al.*, 2006), as comparisons were often made based on visually inspections or on poor statistical tests (McGill, 2003a; McGill *et al.*, 2006). On the other hand, a good fit is not, by itself, a strong test of mechanism – pattern does not equal process (McGill, 2003a; Magurran, 2005b).

Comparing the fit of alternative models has been a common approach to try to reveal the processes shaping the SAD (McGill, 2003a; McGill *et al.*, 2007). The rationale of investigation was to compare empirical patterns of species abundance to theoretical abundance models, with the aim of revealing how the properties of the ecological communities are reflected in the shape of the SAD. However, as more than one mechanism can produce the same pattern (Pielou, 1975; McGill, 2003a; Pueyo *et al.*, 2007), the fit of a model cannot unambiguously provide support for a given theory. As many of the theory testing focused solely on replicating the empirical distribution, this proliferation of models has not led to consensus or to the rejection of (almost) any theory (McGill, 2003a; McGill *et al.*, 2007). With the development of better statistical and methodological tools for model formulation, model fitting, goodness-of-fit testing and model selection, which increased the ability to perform a robust evaluation and selection of competing models and to make distinct and testable predictions, there has been a call for a stronger and more rigorous testing framework of SAD models (McGill,

2003a; McGill *et al.*, 2007; Connolly & Dornelas, 2011). Nonetheless, different measures of goodness-of-fit emphasize different aspects of model fitting and may give diverse ‘responses’ (McGill, 2003a; Magurran, 2005b). A systematic assessment of the ability of different model selection tools to differentiate among SAD models is still lacking. Such assessment, coupled with a model comparison framework, can help provide a more rigorous assessment of SAD studies.

SAD shape is highly sensitive to sampling effort (Fisher *et al.*, 1943; Preston, 1948). Particularly in the case of small samples, many empirical SADs are described equally well by logseries and lognormal distributions, making it hard to distinguish between the two models. Preston (1948) proposed a “veil line” to explain that since the rarest species are not observed with small samples, the left end of the distribution would be truncated, resembling a logseries, and only as the sampling effort increased would the “true” lognormal distribution be progressively unveiled. However, as demonstrated by Pielou (1977) and McGill (2003c), unveiling does not simply reveal the left-end of the distribution, but the shape of the distribution can also change. Furthermore, while the majority of intensely sampled communities seem to follow a lognormal SAD (Magurran, 2004), evidence that large samples frequently deviate from a symmetrical lognormal SAD has been reported, namely more rare species than predicted by a lognormal (Hubbell, 2001; Magurran & Henderson, 2003; McGill, 2003c), and the appearance of more than one modal class of abundance (Ugland & Gray, 1982; Gray *et al.*, 2005; Dornelas & Connolly, 2008). Broadly, these deviations from a lognormal distribution may be caused by heterogeneity within the communities. Deconstructing SADs into different ecological guilds can be an approach to identify different processes determining the assembly of each particular guild (Magurran & Henderson, 2003; Marquet *et al.*, 2004; Alonso *et al.*, 2008; Dornelas & Connolly, 2008). The variability in SADs shape when plotted on a logarithmic scale raises the question of it being caused by stochastic variation, sampling effects or genuine differences in the abundance of the underlying communities (McGill *et al.*, 2007; Connolly & Dornelas, 2011). In this thesis, I provide the first global empirical assessment of multimodality in SADs across taxa, habitats and spatial extents, and show the pattern is related to the spatial scale and taxonomic diversity of the underlying communities. I also provide a systematic evaluation of change in SAD shape across a scale gradient, assessing the effect of sampled area, species richness, total abundance and taxonomic diversity.

1.4 Compositional similarity (β diversity)

Biodiversity changes across space and time. While α diversity represents the diversity within a single site, β diversity quantifies the variation in species composition between assemblages or sites within a landscape (Whittaker, 1960). α and β jointly describe the overall diversity among all the sites in a landscape or region, i.e. γ diversity (Whittaker, 1960; Magurran, 2004). Compositional differences between sites or communities (or times) can reflect both niche processes, such as species' adaptations to different climates or habitats, as well as species dispersal limitations.

As with α diversity, a myriad of metrics to measure β diversity exist (Tuomisto, 2010a,b; Anderson *et al.*, 2011). Measures of compositional similarity can be calculated using incidence or abundance data, and range from 0 to 1, with 0 representing assemblages with no species in common, and 1 representing identical composition. Most similarity metrics also exist as distances. Loosely speaking, compositional differentiation is the opposite of community similarity (0 for identical composition, and 1 for no shared species, or complete turnover). All the metrics attempt to quantify the compositional differences among sites, and similarity should decrease as the distance between sites increases, and as the size of the areas sampled decreases (Harte & Kinzig, 1997; Nekola & White, 1999). The measurement of β diversity is also affected by the spatial scale of observation in terms of grain and extent (Wiens, 1989; Nekola & White, 1999; Keil *et al.*, 2012; Steinbauer *et al.*, 2012; Barton *et al.*, 2013). Although there has been a recent growing interest in β diversity studies (Gaston *et al.*, 2007; Anderson *et al.*, 2011), its scaling properties are still poorly understood, and there is no general framework for describing the spatial scaling properties of β diversity (Barton *et al.*, 2013).

β diversity, defined as compositional heterogeneity among sites, can be associated with different concepts and reflect different aspects of compositional similarity (Koleff *et al.*, 2003; Baselga, 2010; Tuomisto, 2010a; Anderson *et al.*, 2011). On one hand, two conceptual types of β diversity can be defined: directional variation along a gradient and non-directional variation (Anderson *et al.*, 2011). In the first type, β diversity represents differences in composition between sampling units along a spatial, temporal or environmental gradient. It can be quantified as the rate of compositional change, e.g. the distance decay of similarity (DDS), one of the most widely used descriptions of spatial compositional variation. DDS describes how species composition between two sites varies with the geographic distance between them (Nekola & White, 1999; Morlon *et al.*, 2008). Distance decay is

usually analysed by regressing pairwise measures of similarity between sites against pairwise distances, where the slope represents the relative change in compositional similarity through geographic space (Nekola & White, 1999). Comparisons using pairwise dissimilarity between sites represent the inverse of DDS.

The second conceptual type of β diversity quantifies non-directional variation in community composition among a set of sample units within a given spatial (or temporal) extent (Anderson *et al.*, 2011). Multiple-site measures have been developed to quantify the overall heterogeneity among sites, providing a more adequate assessment of compositional similarity than averaging pairwise comparisons when the number of sites >2 (Baselga *et al.*, 2007; Diserud & Ødegaard, 2007; Baselga, 2013). Because this approach is non-directional, it can provide information about variation in species composition among sites at different spatial scales. One further approach is to use the slope of Species-Area Relationships (SAR) as a measurement of species spatial turnover (Harte & Kinzig, 1997); however this is only applicable for the spatial scales at which the SAR follows a power law, and only accounts for species being added as area increases (Harte *et al.*, 1999; Lennon *et al.*, 2001; McGlenn & Hurlbert, 2012). SARs are a universal macroecological pattern, showing systematic variation with scale: a triphasic curve on a log-log scale, with steeper increases in species richness at both small and large spatial scales (Williams, 1943; Rosenzweig, 1995; Storch *et al.*, 2012).

In addition, compositional differences between sites can originate from two different processes, and hence β diversity can be partitioned into two components, turnover and nestedness. This important aspect of β diversity has been recognized before (Harrison *et al.*, 1992; Lennon *et al.*, 2001; Koleff *et al.*, 2003), and has seen recent developments. Specifically, Baselga (2010) provides a partition framework that separates the two components for the two abovementioned types of β diversity. The turnover component represents the replacement of species between sites, whereas the nestedness-resultant component occurs due to changes in species richness between sites – the sites with fewer species are strict subsets of richer sites (Harrison *et al.*, 1992; Koleff *et al.*, 2003; Baselga, 2010). The two components are generated by fundamentally different processes, therefore quantifying their contribution across spatial scales can provide insights into the mechanisms underlying β diversity (Baselga, 2010; Svenning *et al.*, 2011). This can prove to be particularly relevant to understand the mechanisms underpinning biodiversity change in space and time (Dornelas *et al.*, 2014; Magurran *et al.*, 2015).

Here, I assessed if there is systematic variation of β diversity with scale, analysing both directional and non-directional types of β diversity. I provide the first attempt at building empirical β diversity scaling curves for different taxa, and consistently assess the behaviour of DDS across a scale gradient spanning several orders of magnitude, while partitioning both types of β diversity into the turnover and nestedness components.

1.5 Thesis overview

This thesis aimed to systematically analyse biodiversity patterns, namely SADs and community similarity metrics, across scales and over a wide range of communities from different ecosystems, with the aim of synthesizing observed patterns and inform on general principles affecting community structure. In Chapter 2, I describe the collection of available datasets of community abundance data from online repositories, and how a subset of the data I collected was further incorporated into a larger database of biodiversity time-series (BioTIME, ERC Funded Project).

In Chapters 3 and 4, I describe the analyses of multimodality in Species Abundance Distributions. Chapter 3 describes a simulation study to assess and improve the model selection method for detecting multimodality. Chapter 4 describes the application of this improved method to 117 empirical datasets. This investigation detected multimodality with strong support in ~15% of the SADs analysed, and also showed that multimodality is linked to the spatial scale and the taxonomic diversity of the underlying communities.

Chapter 5 describes the analyses of Species Abundance Distributions across a gradient in spatial scale for a smaller number of communities. I undertook an exploratory analysis of how the shape of SADs is affected by area sampled, taxonomic diversity, species richness and total abundance. This analysis showed that multimodality is indeed reflecting the structure of large scale and taxonomic diverse communities, with area, species richness and number of families strongly affecting SAD shape, while total abundance did not exhibit any directional effect. Moreover, it illustrated strong departures from the predictions of two macroecological theories for SAD across scales.

In Chapter 6, I describe the analyses of community similarity patterns across the scale gradient implemented in Chapter 5. This investigation showed remarkable consistency of β diversity scaling curves across the communities analysed. Furthermore, and for both types of β diversity analysed, turnover was the main driver of compositional change. The thesis concludes with a general discussion of the broad implications of the results for community ecology and macroecology (Chapter 7).

2. Collecting Data

This chapter describes the search and collection of available public datasets of community abundance. I will first describe the collection of data for carrying the analysis on the prevalence of multimodality in Species Abundance Distributions (Chapters 3 and 4). Secondly, I will describe how a subset of these data was incorporated into a larger database of biodiversity time-series.

Data for the Multimodality analysis

117 datasets from intensely sampled communities were collected from 3 online repositories: OBIS (Ocean Biogeographic Information System, <http://www.iobis.org/>), Ecological Data Wiki (<http://ecologicaldata.org/>) and GBIF (Global Biodiversity Information Facility, <http://www.gbif.org/>). These repositories hold worldwide and freely available datasets.

Datasets were selected according to the following criteria: 1) data consisted of samples or census of entire communities (*sensu lato*), i.e. did not exclude some taxa intentionally; 2) with a minimum of 10 species; 3) data consisted of numeric abundance (number of individuals or density), holding more than 10,000 records, and thus yielded more than 10,000 individuals; this was intended to minimize under-sampling effects; and 4) the large majority of records were identified to species level. This type of dataset ‘scanning’ was performed in order to retrieve the largest number of suitable available datasets, and any taxonomic or geographical bias is due to data availability. This strict set of criteria for selecting suitable datasets was intended mainly to avoid taxonomic resolution issues and the caveats of under-sampled communities, for which the rarest species are not represented. This is particularly important as alternative distributions proposed to describe species abundance distributions differ especially in the proportion of rare species, and hence are extremely difficult to distinguish for ‘veiled’ distributions (Magurran, 2004). Furthermore only communities with more than 10 species were selected due to the difficulty in constructing SADs with fewer species (McGill *et al.*, 2007; McGill, 2011).

Datasets were checked for duplicates, species with zero abundance and for non-organismal records, which were removed, but were otherwise used as provided. If there was any kind of experimental

manipulation, only control data were used in the analyses. Taxonomic checks were also carried out, mainly to identify typos and misspellings in species names, and to standardise notation for records identified as morphospecies or higher taxonomic levels – the majority of datasets contain some records corresponding to taxa other than species and/or unidentified records. Datasets were discarded whenever they were restricted, they seemed to be incomplete (e.g. less records than stated by provider) or did not contain abundance data, the records did not match the type of sampling or method description.

Overall, the datasets selected cover a wide range of taxa, habitats and temporal and spatial extents (unique sample event to several years, and plots to continents). The datasets comprise data from Marine, Terrestrial and Freshwater realms (see Fig. 4.1 for a map of the datasets location), including plankton, fish, invertebrates, birds, grasses and trees, and range from tropical forests, temperate lakes to marine benthos. The complete list of the datasets and data sources can be found in Appendices I and II, along with a full list of acknowledgements regarding the use of these data.

Integration with the BioTIME database

From the 117 datasets selected according to the specific abovementioned criteria, I was able to identify 44 datasets that also met the BioTIME project criteria for inclusion, namely the ones including a temporal axis. Specifically, all the datasets that were consistently sampled for a minimum of two years and for which it was possible to identify independent sampling events were incorporated into the BioTIME database. In addition, I searched for more datasets that would meet the BioTIME selection criteria, and was able to contribute 46 additional datasets. Efforts for data collection for BioTIME were undertaken by several people involved in the project independently of my contribution. However, the addition of 90 datasets to this endeavour from my part led to a significant growth of the BioTIME database (Fig. 2.1). In total, BioTIME currently holds 384 studies, containing over 12 million records from more than 600 thousand distinct geographic locations, and includes more than 43 thousand species. For the other analyses included in this thesis, appropriate datasets were selected from this larger database, taking the specific requirements of each investigation into consideration.

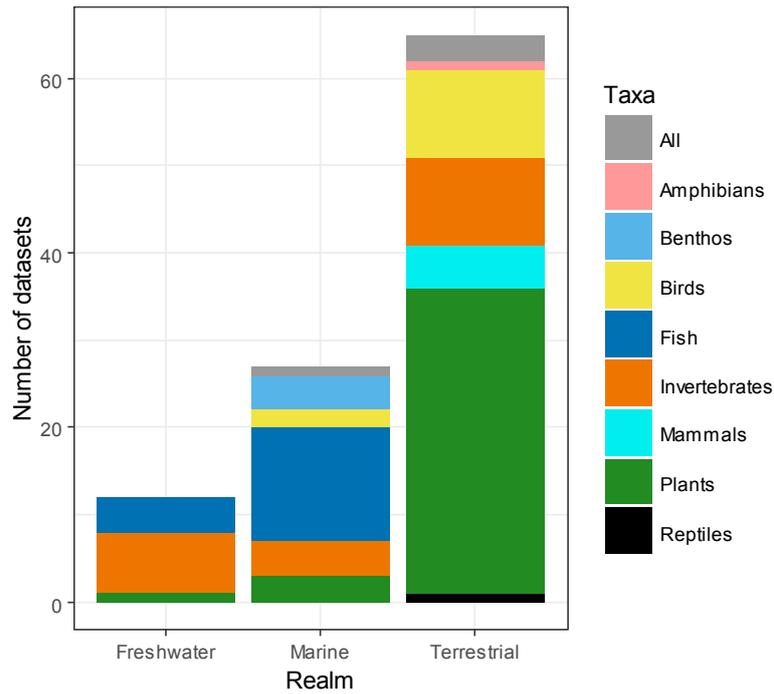


Figure 2.1 Datasets contributed to BioTIME, illustrating the taxon and realm of the community data.

Note: A static release of the BioTIME database is in press in the form of a peer-reviewed article: **BioTIME: a database of biodiversity time series for the Anthropocene**. Maria Dornelas, Laura H. Antão, Faye Moyes, Amanda E. Bates, Anne E. Magurran, *et al.* (200+ authors) (*Global Ecology and Biogeography*).

3. Multimodality in Species Abundance Distributions – improving the detection method

Note: The work and results presented in Chapters 3 and 4 were published in the form of a peer-reviewed article: **Prevalence of multimodal species abundance distributions is linked to spatial and taxonomic breadth** (2017). Laura H. Antão, Sean R. Connolly, Anne E. Magurran, Amadeu Soares & Maria Dornelas. *Global Ecology and Biogeography*, 26: 203–215. DOI: 10.1111/geb.12532.

3.1 Introduction

Explaining the patterns of commonness and rarity of species is fundamental for understating how ecological communities are structured and maintained. Species Abundance Distributions (SADs) depict the relative abundance of the species present in a community and describe one of the most fundamental patterns of species diversity – most communities contain many rare and only a few common species (McGill *et al.*, 2007). Empirical datasets consistently produce species abundance distributions that are quasi-hyperbolic on an arithmetic scale – the ubiquitous ‘hollow curve’. However, on a logarithmic scale of abundance, SADs exhibit more variability, with species abundance distributions alternately exhibiting no internal mode - most species occur at the lowest abundance class (i.e. as singletons), one internal mode, or multiple internal modes. Despite several decades of study with dozens of different models proposed to explain SADs (McGill *et al.*, 2007), there is still no consensus about what drives variation in SADs shape, nor how it might be connected to factors structuring ecological communities (Fisher *et al.*, 1943; Preston, 1948; Magurran & Henderson, 2003; McGill, 2003c; Green & Plotkin, 2007; Dornelas *et al.*, 2009). The extent to which current biodiversity theories are able to accommodate and explain such variation in SAD shape is a critical criterion to their evaluation and application (McGill *et al.*, 2007).

The two distributions recurrently proposed to describe SADs are the logseries (Fisher *et al.*, 1943) and the lognormal (Preston, 1948) (Fig. 3.1). While many intensely sampled communities seem to follow a lognormal distribution (Magurran, 2004), it has become increasingly clear that empirical SADs often deviate from a lognormal by having more than one internal mode (Ugland & Gray, 1982; Gray *et al.*, 2005; Dornelas & Connolly, 2008). Multimodality is seldom reported and its implications little explored (McGill *et al.*, 2007), with some notable, but dispersed, exceptions. Ugland & Gray (1982) proposed three lognormal distributions, corresponding to rare, intermediate abundant and common species, to describe non-equilibrium marine benthic communities. Magurran & Henderson (2003) ‘deconstructed’ an estuarine fish community into two groups - ‘core’ and ‘occasional’, based on species persistence and habitat preferences, where the first group was better fit by a lognormal, while the ‘occasional’ group of rare species followed a logseries distribution. Gray *et al.* (2005) showed that a mixture of two lognormal distributions provided a good fit to a marine benthos and a tropical tree data, again separating the species into ‘abundant’ and ‘rare’.

In the first statistical analysis comparing the fit of distributions with varying numbers of modes, Dornelas & Connolly (2008) showed that the SAD of an intensely sampled coral community was multimodal. However, the different modes could not be explained by a mixture of species associated with different habitats, and were only partially explained by different spatial aggregation. Recently, Matthews *et al.* (2014), using the same methodology for an arthropod community, showed that multimodal distributions performed better for many of the samples analysed, and that grouping ecologically different species leads to multimodality, with the rarest species mode containing a higher proportion of satellite, introduced and species better adapted to other habitats. However, the effect of dispersal ability was unclear, and a body size niche axis was unrelated to the multimodal patterns. The commonality among these studies is that they indicate that multimodality is linked to ecological heterogeneity, broadly defined as groups of species with different ecological or functional characteristics. This suggests that multimodality should have higher prevalence among communities with higher ecological heterogeneity. The concept of ecological heterogeneity proposed here is deliberately broad, and is intended to incorporate the spatial, environmental, taxonomic and functional aspects of ecological systems, rather than simply accounting for the number of species or of functional groups.

The prevalence of multimodality in empirical SADs is as yet unknown. In a recent theoretical study, Barabás *et al.* (2013) reported that stochastic versions of both resource partitioning and neutral

models can produce multimodal SADs with a 50% prevalence. The authors argue that in nature, individual realizations are likely to differ from the mean predicted pattern due to stochastic processes. On the other hand, the authors also disputed that the Emergent Neutrality model proposed by Vergnon *et al.* (2012) is the only theoretical model able to produce multimodal SADs. Apart from the Emergent Neutrality Theory, no other theoretical framework predicts that SADs can exhibit multiple modes. Also, other than the abovementioned studies that specifically tried to address multiple modes, multimodality is usually overlooked, assumed to be due to sampling errors or a rare pattern. Thus, assessing the prevalence of multimodality in empirical datasets is warranted to establish the generality of the pattern, as well as help elucidate how it can be related to different ecological explanations.

The main goals of this analysis were: 1) to improve the method of multimodality detection, to be able to confidently detect multiple modes in empirical SADs; 2) to undertake a global empirical assessment of the prevalence of multimodality for a wide range of communities. This represented the first assessment of the prevalence of multiple modes in SADs; and 3) to test the hypothesis that more heterogeneous communities are more likely to exhibit multimodality.

3.2 Methods

3.2.1 Multimodal functions

For the lognormal distribution, assuming a Poisson random sampling process from the real community and independence of species' abundances, the log-likelihood function commonly used is the 'zero-truncated' form. This conditions the probability of a species having abundance r on the species being present in the sample (Connolly & Dornelas, 2011; Connolly & Thibaut, 2012). The probability distribution for a mixture of g PLNs (Φ_g) was calculated as:

$$\Phi_g(r) = \sum_{n=1}^g \rho_n \cdot \phi_n(r)$$

where r is each abundance value, ρ_n is the proportion of species belonging to each distribution n , and $\sum_{n=1}^g \rho_n = 1$.

The logseries distribution has only one parameter (α), while 1PLN has two (μ and σ - the mean and standard deviation of log abundances), 2PLN has five parameters ($\mu_1, \mu_2, \sigma_1, \sigma_2$ and ρ - the mean and standard deviation of log abundances of each distribution in the mixture, and ρ - the probability of a species belonging to the first distribution), and 3PLN has eight parameters ($\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3, \rho_1$ and ρ_2 - the mean and standard deviation of log abundances of each distribution in the mixture, ρ_1 and ρ_2 - the probabilities of a species belonging to the first or to the second distribution, respectively) (Fig. 3.1).

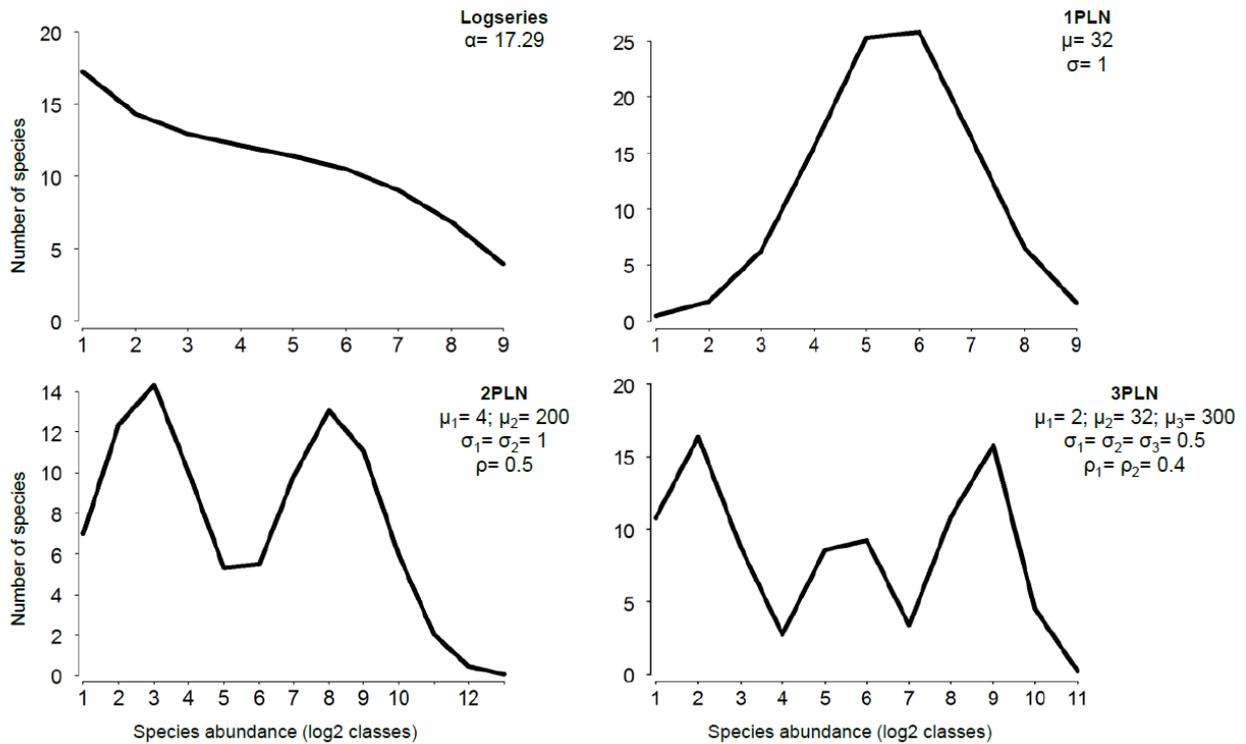


Figure 3.1 Examples of random sampled communities for a logseries, a single lognormal Poisson (1PLN), and mixtures of two and three lognormal Poisson distributions (2PLN and 3PLN, respectively). For the logseries distribution, the single parameter is Fisher's α . For the PLN models, μ and σ are the mean and standard deviation of log-abundance for one of the underlying lognormal community abundance distributions (one pair of parameters for each mode), and ρ_n is the probability that a species comes from distribution n . The parameters used to generate the random sampled abundance data for each model are shown, and the species pool size was $S = 100$ (the code to generate the 2PLN example can be found in Appendix IV).

3.2.2 Model Selection

To test whether models with more than one mode accurately reflect the abundance distributions of the underlying communities maximum likelihood methods were used to explicitly compare the fit of mixtures of 1, 2 and 3 Poisson Lognormal distributions (1PLN, 2PLN and 3PLN, respectively) (Pielou, 1969; Bulmer, 1974); a logseries distribution was also included (Fig. 3.1). All the calculations were performed in the software R (R Core Team, 2017). Functions to fit the PLN mixtures and to calculate maximum likelihood estimates (MLE) were adapted from Dornelas & Connolly (2008) but using the `dpoilog()` function from `poilog` package (Grøtan & Engen, 2008); the log-likelihood functions are otherwise similar and best fit parameters were found by minimizing the negative log-likelihood. These functions are available in Appendix III. Parameter estimation was performed using the R optimization routine `nlm` and parameter searches were initialized from multiple starting points due to the possibility of several local maxima for more complex distributions (Dornelas & Connolly, 2008; Connolly & Dornelas, 2011).

Model comparison was performed under a multi-model information-theoretic framework (Burnham & Anderson, 2002), using the second order Akaike's information criterion for small sample sizes (AIC_c , Burnham & Anderson, 2002) and Bayesian information criterion (BIC, Schwarz 1978). AIC_c was used throughout as it converges to AIC when sample size is large (Burnham & Anderson, 2002, 2004). AIC and BIC are model selection tools that provide quantitative relative support for alternative hypotheses, while finding a compromise between goodness of fit and model complexity. AIC tends to overestimate the number of distributions in mixture models, while BIC tends to underestimate them (McLachlan & Peel, 2000; Henson *et al.*, 2007). Hence, the performance of these two model selection criteria was evaluated with a simulation study (see section 3.2.3).

Model performance was evaluated in slightly different ways in the empirical and simulation studies. For the analysis of the empirical data (Chapter 4), relative support for the models was calculated as ΔAIC_c , which is the difference between the AIC_c of each model, and the lowest AIC_c in the model set (see section 4.1). Differences larger than 2 indicate substantial evidence against the model with the higher AIC_c (or BIC) (Burnham & Anderson, 2002). However, for the simulation study, the “*true model*” (the model used to generate the simulated data) is known. Therefore, AIC differences were calculated relative to this *true model*, a quantity which was termed AICdiff. Specifically, AICdiff is the AIC_c of the *true model*, minus the smallest AIC_c of the remaining models. This quantity is

negative whenever the *true* model is the best fitting model (the one with the lowest AIC score). Conversely, if one or more of the alternative models actually fits better than the *true* model does, then AICdiff will be positive. Note that AICdiff=0 does not indicate the best fitting model. An analogous quantity was calculated for BIC for the simulation study.

3.2.3 Simulation Study

Because the *PLN-mixture* method has only been applied to specific datasets (Dornelas & Connolly, 2008; Vergnon *et al.*, 2012; Matthews *et al.*, 2014), a simulation study was conducted to assess how it performed under a broad range of parameter combinations. Specifically, the simulation study was performed to determine which conditions would lead the *PLN-mixture* method to select a model with the wrong number of modes. For instance, when simulating data from a 2PLN mixture – where the underlying community has two modes - the position of the second mode was fixed, while the first mode travelled so that the distance between the modes decreased. This allowed testing when an assemblage that is actually a mixture of two lognormal distributions is mistakenly better fit by a single PLN, or by a logseries. Similarly, for 3PLN mixtures the three modes were positioned increasingly closer together, leading the simulated data to become increasingly indistinguishable from two-mode or one-mode distributions.

A *false positive* was defined as simulated samples where a multimodal model was selected with high confidence when the *true* model generating the sample was not multimodal; and a *false negative* as simulated samples where the *true* model was multimodal but for which a ‘non-multimodal’ model was selected. A range of species richness and parameter values for the four alternative abundance distributions models was used to generate simulated count data. The spectrum of parameters used was designed to cover a realistic range for species abundance data (Connolly & Thibaut, 2012), and to provide a quantitative picture of whether and when the method fails to select the true number of underlying modes.

The function `fisher.ecosystem()` from `untb` package (Hankin, 2007) was used to generate count data from logseries distributions. The parameter space was explored by factorially varying total number of individuals $N = (1000, 10000, 20000)$ and number of species $S = (20, 100, 200, 500)$ (this is the species pool to be sampled). To generate PLN count data the `rpoilog()` function from `poilog` package (Grøtan & Engen, 2008) was used. Parameter values of μ , σ and ρ were varied systematically, using μ values that would fall into different octaves and $\sigma = (0.5, 1, 2)$. For the 1PLN simulations μ values were increased, thus replicating the unveiling process, while keeping σ and S constant. For the 2PLN simulation, the second mode was fixed (μ_2 in octave 8), while μ_1 values increased, falling into different octaves increasingly closer to μ_2 . A similar procedure was performed for the 3PLN simulations, fixing the third mode (μ_3 in octave 9), and decreasing the distance between the modes,

first by positioning μ_2 closer to μ_3 , and then bringing the three modes to consecutive octaves. Species richness levels were set as $S = (20, 100, 500)$, representing the species pool to be sampled. All the simulated sampled communities with less than 10 species were excluded, due to the difficulty in constructing SADs with fewer species (McGill *et al.*, 2007; McGill, 2011). All the fitting routines were run on non-binned data.

A total of 162 parameter combinations were examined; for each parameter combination, 100 simulated SAD samples were generated and the alternative log-likelihood functions were fit (more details in Appendix IV). For each simulated SAD sample AICdiff was calculated as:

$AIC_{diff} = AIC_c \text{ true model} - \min(AIC_c \text{ remaining models})$ (and similarly for BIC), the *true* model being the one generating the data, not the specific parameters used. This is a slight modification of the standard calculation of ΔAIC (Burnham & Anderson, 2002), as explained before.

3.2.4 Parametric Bootstrap

Following the simulation study results, some 1PLN parameter combinations were identified where AIC_c strongly selected a more complex model than the one generating the data with a frequency of up to ~25% of the simulated samples (Fig. 3.3). To minimise the chance of a multimodal model being selected due to overfitting of the method, Likelihood Ratio tests (LRT) were additionally calculated. Likelihood Ratio tests assess if the improvement in goodness of fit of a more complex models is greater than would be expected by chance, if the simpler model were true. LRT are only applicable to nested models, so the logseries was not included in this analysis. Because the null distribution of LRT is known to occasionally deviate from a χ^2 distribution (McLachlan, 1987; McLachlan & Peel, 2000), null LRT frequency distributions from 1PLN simulated communities were generated. This allows calculating the equivalent of a p-value for the null hypothesis that the sampled data are consistent with a 1PLN distribution, thus providing an alternative assessment of whether a multimodal model provided the best fit for that parameter combination. For the simulation study, this was illustrated by comparing LRT distributions for two parameter combinations, one from the parameter space where AIC_c successfully selected 1PLN, and the other from the space where AIC_c has a higher probability of selecting a more complex model. See section 4.1 for the application of this procedure to the empirical datasets.

3.3 Results

3.3.1 Simulation study

Overall, the *PLN-mixture* method was robust to large variation in the parameters used to perform the simulations. The false positive frequency was very low, particularly for BIC where in only 1% of the cases was a multimodal model selected with high confidence as the best fit model when the true model was not multimodal, and for AIC_c it was 6% (Table 3.1). The position of the modes, species richness and particularly σ values showed strong effects in the best-fit model selection, for both AIC_c and BIC, sometimes with different directions.

In more detail, for all the parameter combinations used to generate logseries data, the average frequency of selecting the *true* model was 90% for AIC_c and 96% for BIC (Table 3.1 and Fig. 3.2). For the simulations with 1PLN as the *true* model, species richness (S) had strong and disparate effects. For AIC_c, the percentage of false positives increased with S, while for BIC the percentage of failures decreased. For highly truncated distributions (very small μ), logseries was selected as the best model, but as the sampled ‘communities’ became unveiled, 1PLN was selected. When inspecting the 1PLN simulation results in more detail, some particular parameter combinations led AIC_c to consistently and strongly favour more complex models than the one generating the data (Fig. 3.3). Increasing the mean (μ) and particularly the standard deviation (σ) caused AIC_c to increasingly overestimate the number of modes (e.g., for simulations with S=500, on average 4.3, 8.1 and 17.2% for $\sigma=0.5, 1$ and 2 , respectively), whereas BIC only very rarely selected 2PLN or 3PLN, except for S=20 when this pattern was reversed (Table 3.1 and Fig. 3.3).

The overall false negative frequency, i.e. simulations where the model generating the sampled communities was multimodal but for which a ‘non-multimodal’ model was selected as best fit, was 25% for AIC_c and 39% for BIC (Table 3.1). For 2PLN and 3PLN simulations, the true model was selected when the modes were clearly separated, for smaller σ values and for higher species richness. BIC started to select a simpler model as the distance between the modes decreased ‘earlier’ than AIC_c, which was still able to select the true model for closer modes.

Again in more detail, for the 2PLN simulations, 1PLN started to be selected as best model as the distance between the two modes decreased and as σ values increased. For some simulated communities, 3PLN was selected as best model with high confidence, particularly for AIC_c ; again this pattern was reversed for $S=20$, where only BIC selected 3PLN as best model a few times (Fig. 3.4). Increasing σ values and varying the proportion ρ , as well as lower species richness, progressively increased the frequency of 1PLN being selected as best model. For the simulations with 3PLN as the *true* model, both AIC_c and BIC successfully selected the model generating the data with high confidence when the three modes were well apart and σ was small. When the modes are 2 octaves apart, 2PLN starts to be selected, and then 1PLN when the modes were in consecutive octaves. Again σ values had a strong impact on the selection criteria, with higher σ leading to a ‘quicker’ shift from 3PLN to 2PLN and further to 1PLN being selected. Species richness also had strong effect, as 3PLN was never selected when $S=20$ and only when the modes were well apart for $S=100$ (for instance, for $S=20$, AIC_c only once selected 3PLN as the best model, and BIC only 63/800 simulated communities) (Fig. 3.5).

3.3.2 Parametric Bootstrap

When likelihood ratio tests were used in addition to AIC_c , the chance of selecting a more complex model decreased compared to when using AIC_c alone (Fig. 3.6). For the parameter space where AIC_c very rarely selected a multimodal model, the LRT distribution overlapped with the AIC_c selection pattern (Fig. 3.6 a and b). When AIC_c had a higher false positive frequency, using the LRT reduced the chance of erroneously selecting a multimodal model. Furthermore, the parametric bootstrap p-value is more conservative than the critical value from a χ^2 distribution for the latter case (Fig. 3.6 c and d). Following these results, and because the high false negative frequency for BIC suggests that it might not effectively detect multimodality, both AIC_c and PBLRT were used to analyse the empirical SADs.

Table 3.1 Overall false positive and false negative frequencies (a) and detailed results for each *true* model used to generate simulated data (b). All the frequencies indicate incorrectly selecting ‘non-multimodal’ or ‘multimodal’ models with high confidence ($AIC_{diff} / BIC_{diff} \geq 2$).

a)

True Distribution	Non-multimodal		Multimodal	
	Logseries	1PLN	2PLN	3PLN
Modes	No internal mode	1	2	3
Parameters	1	2	5	8
Parameter combinations	12	82	44	24
# Simulations	1200	8200	4400	2400
% Failures	AIC _c	3.167	6.610	23.364
	BIC	0.500	1.061	41.477

b)

True Distribution	Modes	Number of species S			
	No internal mode	500	100	20	Total
Logser	Parameter combinations	12			
	Total # Simulations	1200			
Total # Failures	AIC _c	38			
	BIC	6			
% Failures	AIC _c	3.167			
	BIC	0.500			
1PLN	1				
	Parameter combinations	29	29	24	82
# Simulated communities		2900	2900	2400	8200
	AIC _c	308	194	40	542
Total # Failures	BIC	2	8	77	87
	AIC _c	10.621	6.690	1.667	6.610
% Failures	BIC	0.069	0.276	3.208	1.061
	2PLN	2			
Parameter combinations		18	13	13	44
	Total # Simulations	1800	1300	1300	4400
Total # Failures	AIC _c	46	193	789	1028
	BIC	482	656	687	1825
% Failures	AIC _c	2.556	14.846	60.692	23.364
	BIC	26.778	50.462	52.846	41.477
3PLN	3				
	Parameter combinations	8	8	8	24
Total # Simulations		800	800	800	2400
	AIC _c	52	95	553	700
Total # Failures	BIC	115	268	443	826
	AIC _c	6.500	11.875	69.125	29.167
% Failures	BIC	14.375	33.500	55.375	34.417

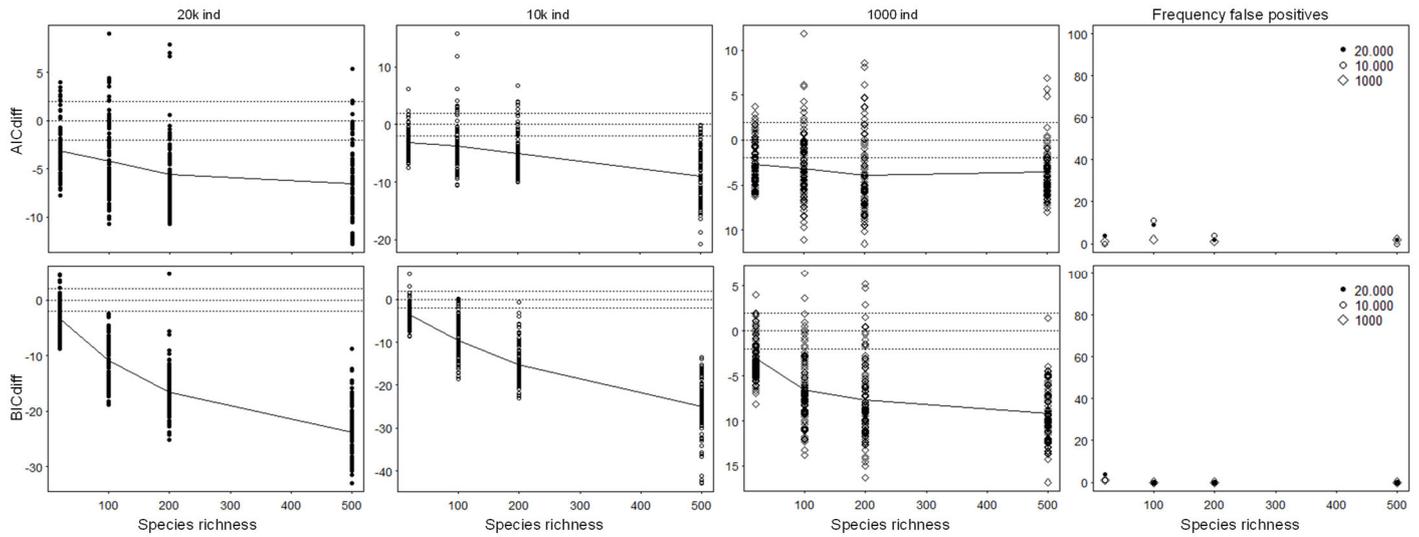


Figure 3.2 Logseries simulation results for AICdiff (top row) and BICdiff (bottom row) varying the number of species and number of individuals. The full line represents the mean AICdiff and BICdiff for the 100 simulated communities; horizontal lines for $y = -2, 0$ and 2 were added to aid visualization (AICdiff ≤ 0 correctly select *true* model; AICdiff ≥ 2 fail to select *true* model with high confidence). The last plot on the right shows the frequency of false positives out of the 100 sampled communities for each parameter combination (i.e. frequency of AIC_c or BIC selected 2PLN or 3PLN with high confidence).

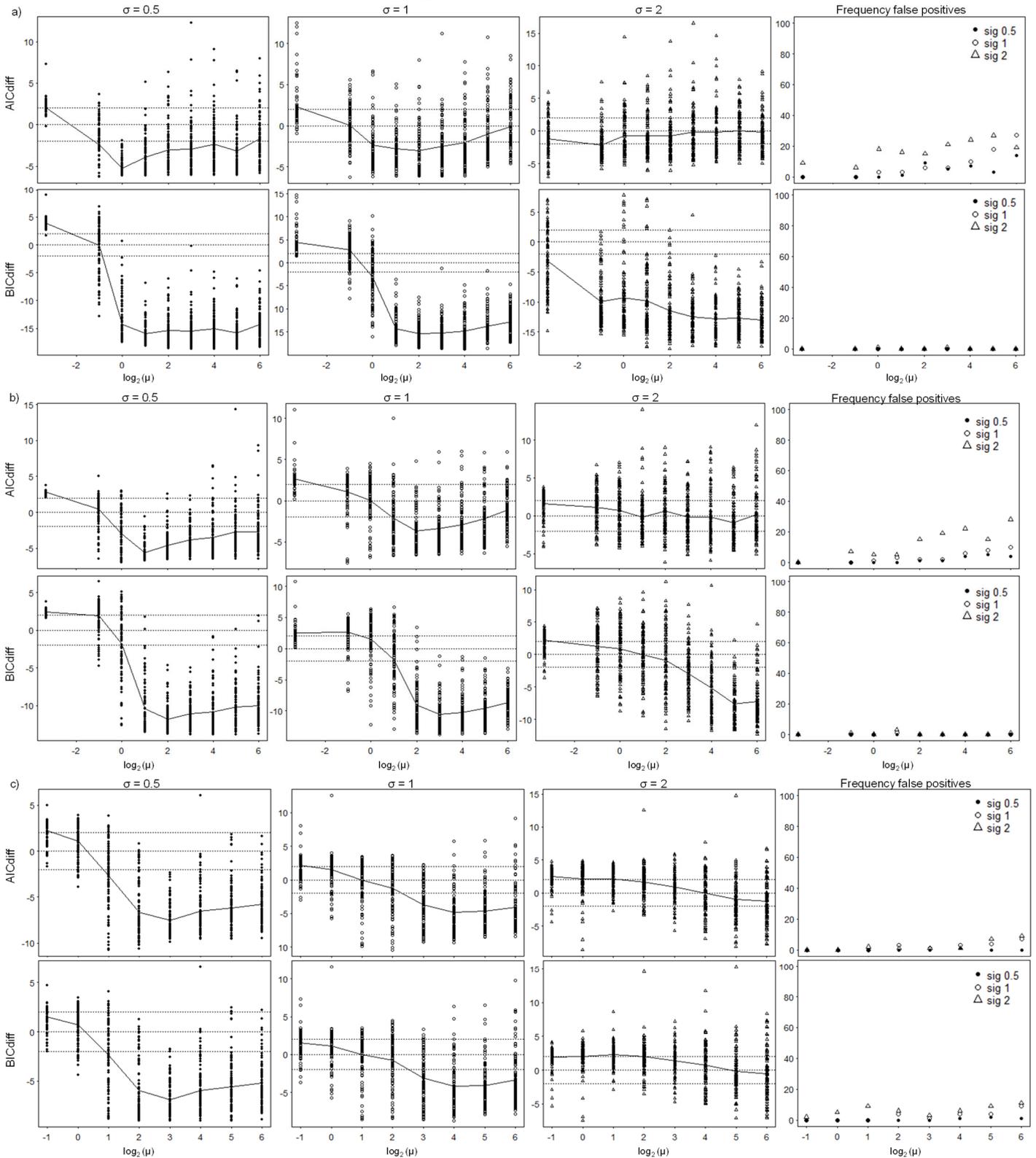


Figure 3.3 IPLN simulation results for AICdiff (top row) and BICdiff (bottom row) for $S=500$ (a), $S=100$ (b) and $S=20$ (c), varying μ and σ values (μ on a \log_2 scale). The full line represents the mean AICdiff and BICdiff for the 100 simulated communities. Horizontal lines for $y = -2, 0$ and 2 were added to aid visualization (AICdiff ≤ 0 correctly select *true* model; AICdiff ≥ 2 fail to select *true* model with high confidence). The last plot on the right shows the frequency of false positives (i.e. frequency of AIC_c or BIC selected 2PLN or 3PLN with high confidence).

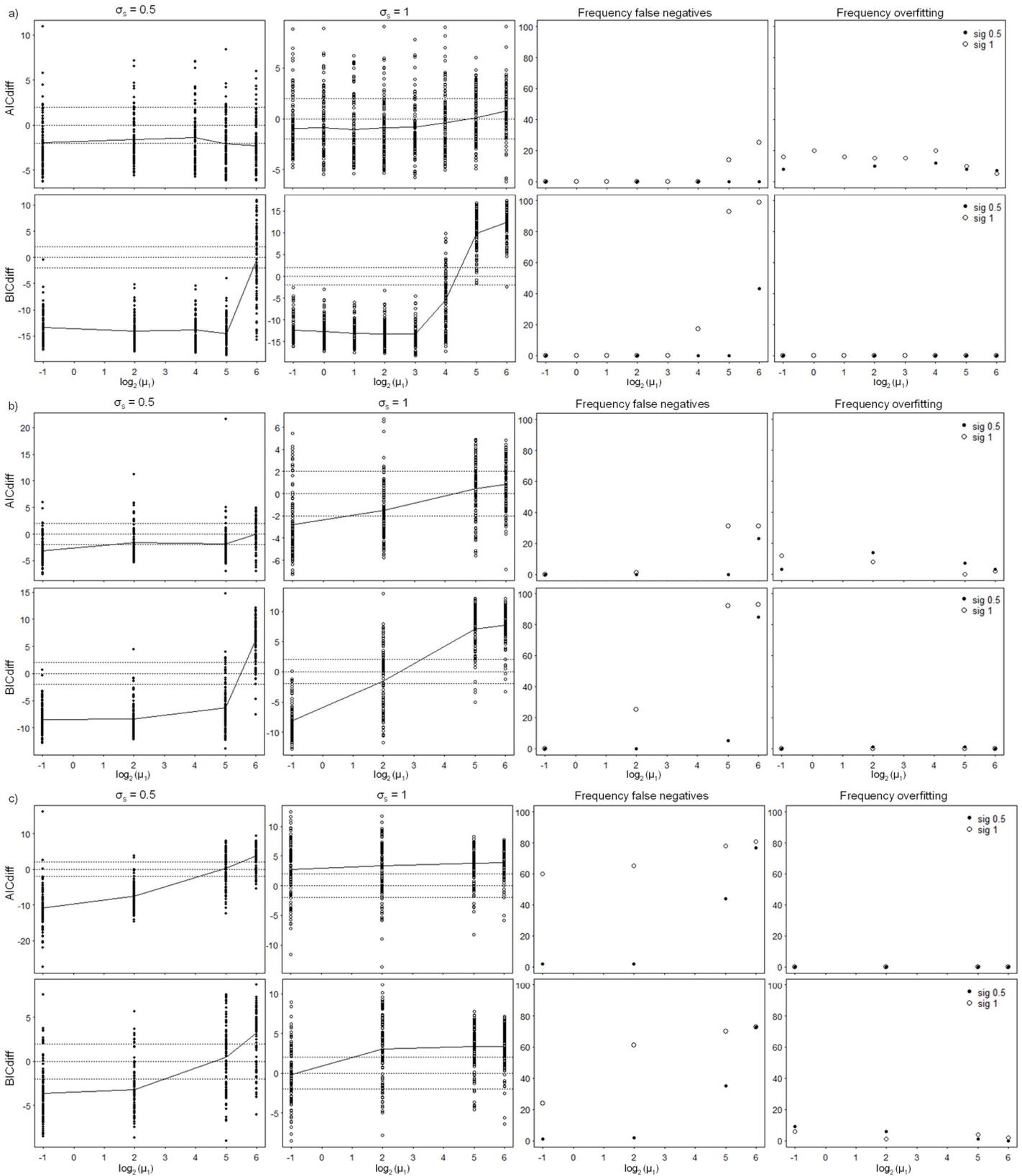


Figure 3.4 2PLN simulation results varying the position of the first mode μ_1 and σ values (μ_2 fixed, μ on a log2 scale); top row results for AIC_c and bottom row for BIC, for S=500 (a), S=100 (b) and S=20 (c). The last two plots on the right show the frequency of false negatives (i.e. frequency of AIC_c or BIC selected 1PLN or logseries with high confidence), and the frequency of overfitting (i.e. selection frequency of 3PLN with high confidence).

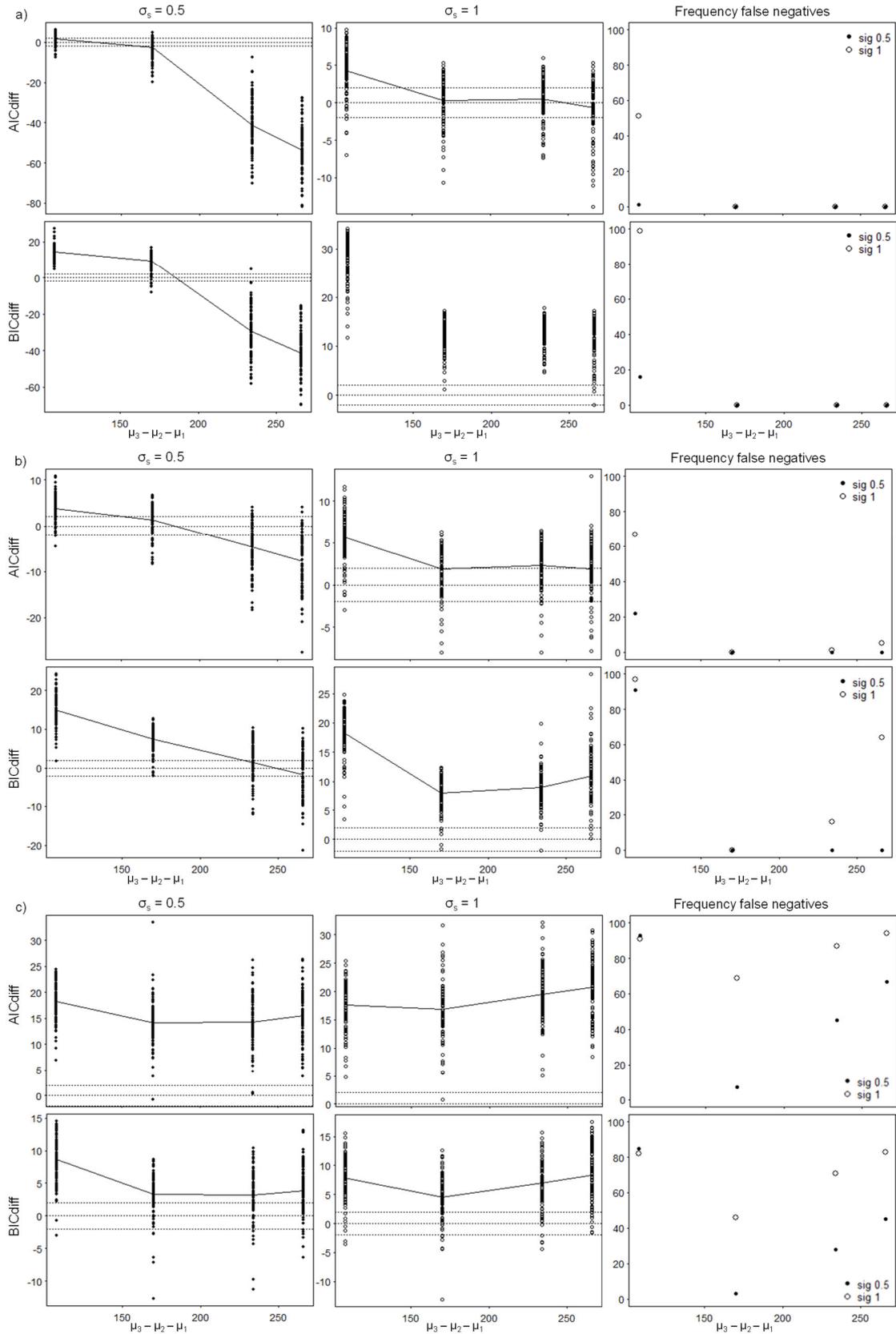


Figure 3.5 3PLN simulation results varying the position of the first and second modes (μ_3 fixed), represented as the distance between the three modes ($\mu_3 - \mu_2 - \mu_1$), and σ values; top row results for AICdiff and bottom row for BICdiff, for $S=500$ (a), $S=100$ (b) and $S=20$ (c). The last plot on the right shows the frequency of false negatives.

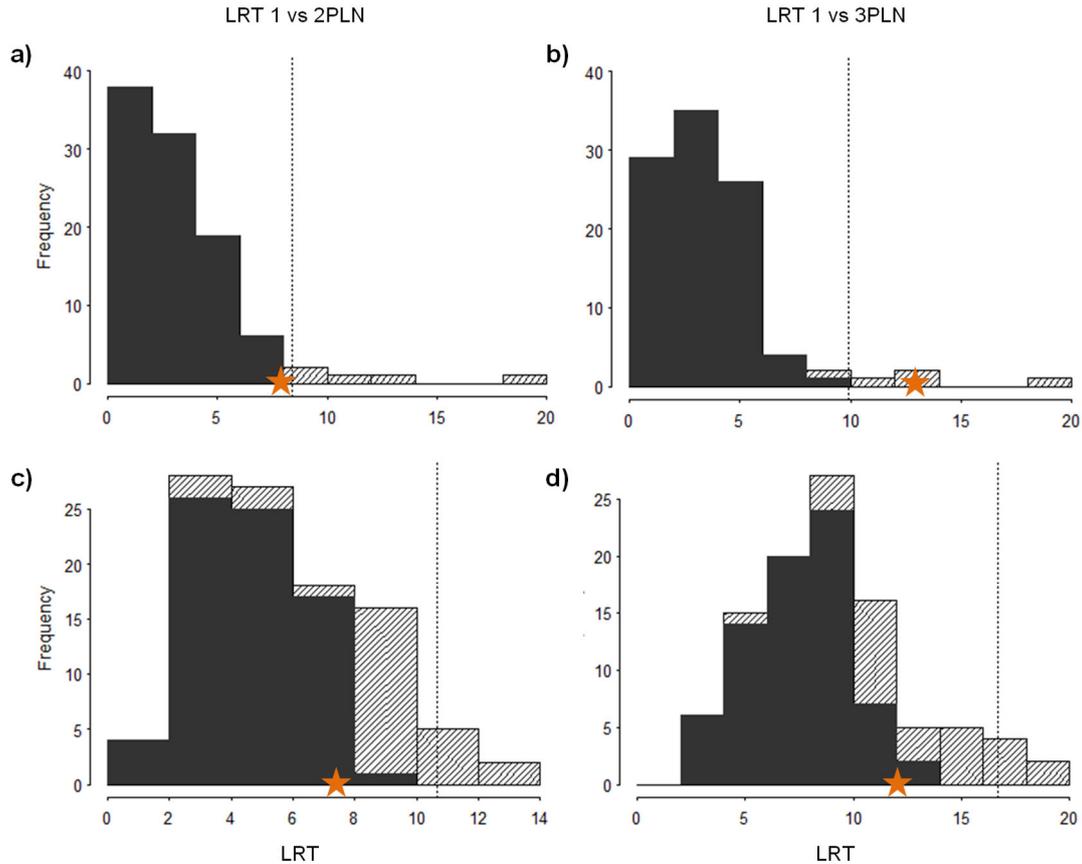


Figure 3.6 Likelihood Ratio test (LRT) frequency distributions calculated from the 1PLN simulated communities for the parameter combinations $\{\mu=8 \text{ and } \sigma=0.5\}$ (a, b) and $\{\mu=32 \text{ and } \sigma=2\}$ (c, d) (both $S=500$) – represented as black bars. The first parameter combination is from the parameter space where AIC_c successfully selected 1PLN, and the second is from the space where AIC_c has a higher probability of selecting a more complex model. Left panels show the distribution comparing 1 vs 2PLN (a and c) and right panels 1 vs 3PLN (b and d). Dotted vertical lines indicate the bootstrap p-value. Diagonal striping histograms represent the frequency of AIC_c selecting multimodality with strong support for the same set of simulated communities, showing that using the LRT distribution rather than AIC_c alone allows reducing the false positive probability. The critical value from the χ^2 distribution for $\alpha=0.05$ is also shown, represented as a star (d.f. = 3 for 1 vs 2PLN, and d.f. = 6 for 1 vs 3PLN). The bootstrap p-value is more conservative for the parameter space where AIC_c is more likely to overfit.

3.4 Discussion

The simulation study showed that the position of the modes, species richness and particularly σ values greatly affected model selection, for both AIC_c and BIC. Additionally, species richness often had contrary effects on the information criteria; this can be related to the high level of penalization exerted by AIC_c as sample size decreases (Burnham & Anderson, 2002), while the opposite happens for BIC (by definition), which can be problematic when testing for multimodality in SADs. As expected, BIC was more conservative than AIC_c , reflected both in the very low false positive frequency and particularly in the relatively high frequency of false negatives. While the former is a highly desirable feature of a selection method, the latter suggests that BIC can be insensitive to deviations in SADs indicative of multimodality.

On the other hand, although AIC_c overestimated the number of modes for some parameter combinations, for a large number of empirical SADs with estimated parameters within that space, the more parsimonious model was selected. This suggests that AIC_c is not overestimating the number of modes generally, and that model selection criteria might be affected by parameter values in a nondirectional fashion. As noted before for SADs, comparative measures of goodness of fit can often produce conflicting results (McGill, 2003a; McGill *et al.*, 2007). This study showed that additionally calculating LRT frequency distributions further reduces the probability of erroneously selecting multimodality when compared to using AIC_c alone.

4. Multimodality in Species Abundance Distributions – empirical analyses

Having extensively assessed the performance of the *PLN-mixture* detection method with the simulation study described in Chapter 3, and knowing that calculating Likelihood Ratio tests in addition to AIC_c decreases the probability of erroneously selecting a more complex model, the detection method was employed to empirical community data. The data was collected according to the criteria described in Chapter 2. This analysis represented the first empirical assessment of the prevalence of multiple modes in SADs.

4.1 Methods

For each of the 117 empirical datasets (Appendix I; see a complete list of the data sources in Appendix II) a simplified vector of abundances was obtained, corresponding to one year of sampling only (the most recent year with at least 10,000 individuals where multiple years were sampled). This was intended to prevent interannual variability from inducing multimodality *sensu* Magurran & Henderson (2003), as the focus here was in assessing the prevalence of multimodal SADs independent of a temporal effect of species abundances fluctuations among years. A map with the datasets location is shown in Fig. 4.1. The datasets were classified according to spatial extent and taxonomic breadth (Table I.1). These two variables were intended to represent different axes of ecological heterogeneity. Regarding spatial extent, as explicit estimates of extent were not available for all datasets, datasets were classified as Local when data originated from plots or sampling locations within less than 1° latitude/longitude, as Regional when data comprised larger areas (e.g. countrywide or larger biome patches), and as Continental when data spanned broader areas such as the whole eastern North American coast or Antarctica. Regarding taxonomic breadth, the number of families was used to quantify this variable. The four alternative abundance distributions models were fitted to each empirical SAD and relative support for each model was calculated as ΔAIC_c (Burnham & Anderson, 2002). BIC was not included in the empirical analysis, as per the results of the simulation study. All the fitting routines were run on non-binned data.

For all the SADs selected as multimodal by AIC_c, a parametric bootstrap likelihood ratio test was conducted (PBLRT; see Knappe & de Valpine (2012) for an example). The parametric bootstrap procedure consisted of randomly generating species abundance values from a 1PLN density function parameterized using the model's maximum likelihood estimates for that empirical dataset (Connolly *et al.*, 2009). As these analyses are very computationally intensive (Dornelas & Connolly, 2008; Connolly & Dornelas, 2011), 100 parametric bootstrap samples were generated for each dataset, using $\hat{\mu}$ and $\hat{\sigma}$ (the estimated mean and standard deviation of log-abundances, respectively) and sample size as the observed number of species, and the log-likelihood functions were fit (code available in Appendix IV). For instance, for dataset ID4, estimated parameters were $\hat{\mu}=19.21$ and $\hat{\sigma}=5.31$ (Table 4.1), and the number of species is $S=39$. Using these parameter values and S as sample size, 100 parametric bootstrap samples were generated, and the PLN mixture distributions were fitted. This procedure allowed comparing the *empirical* likelihood ratio, calculated from the empirical SAD fitting, with the frequency distribution expected under the null hypothesis that the data are actually a single PLN.

Finally, it was assessed whether the prevalence of multimodality was influenced by spatial extent and taxonomic breadth (and their interaction) using two models: first, a binomial generalised linear model (GLM) was used, aggregating 1PLN and logseries as 'non-multimodal', using the R function `glm()` with the logit link function (*binomfit* model below). Additionally, a multinomial Bayesian generalised linear model was used to assess the prevalence of multimodality, 1PLN and logseries separately. The Markov chain Monte Carlo (MCMC) estimation was performed using the R package `MCMCglmm` (Hadfield, 2010). A model with a random intercept was fitted to obtain improved parameter estimates for each level of the fixed effects (see `MCMCglmm` vignette (Hadfield, 2010) and Gelman & Hill, 2007), running 5,000,000 iterations with a burn-in of 100,000 and a thinning interval of 25 (*multinomfit* model below).

```
binomfit <- glm (multimodal/non-multimodal ~ SpatialExtent * NumberFamilies, family="binomial")
```

```
multinomfit <- MCMCglmm (MODELselected ~ -1 + trait + trait:(SpatialExtent * NumberFamilies),  
rcov=~ idh (trait):units, family= "categorical", nitt= 5 000 000, thin= 25, burnin= 100 000)
```

4.2 Results

Of the 117 empirical SADs, AIC_c selected a multimodal model for 47 SADs, 26 of which with high confidence. For many SADs, estimated 1PLN parameters fell within the parameter space for which AIC_c often selects a multimodal model with high confidence when the true distribution is unimodal (specifically with an estimated standard deviation of log abundance, $\hat{\sigma}$, of about 2). On the other hand, all the SADs selected as logseries also had estimated $\sigma \geq 2$ for the 1PLN model. This suggests that the method is not overfitting generally, but can occasionally select a more complex model. On visual inspection, none of the fitted curves seemed to be odd-looking or out of phase with the empirical SAD (Figs. 4.2 and V.1 in Appendix V), although it is possible that SADs that appear unimodal are better fit by multimodal models, and vice-versa (Matthews *et al.*, 2014).

For the SADs selected as multimodal, PBLRT supported AIC_c model selection for 17 SADs, where the empirical likelihood ratio values were higher than the bootstrap p-value from the PBLRT distribution) (Fig. 4.2 and Table 4.1). For the cases where the PBLRT results did not support multimodality, the second best model was assumed to be the best model (either logseries or 1PLN). Overall, 17 SADs were multimodal with high confidence, 1PLN was the best model for 54 and for 46 it was logseries (Table I.1). None of the datasets selected as logseries had continental spatial scale.

Regarding the effect of spatial extent and taxonomic breadth, both have a positive effect on the prevalence of multimodality (Table 4.2). For the binomial GLM, SADs with Local spatial extent were significantly less likely to be multimodal ($p=0.0073$) *vs* Continental and Regional scales, and there is a positive effect of the interaction between number of families and the Local scale ($p=0.00407$). When using the multinomial GLM, SADs with Local spatial extent were again significantly less likely to be multimodal *vs* 1PLN (Fig. 4.3; $pMCMC=0.01943$), but not at Continental and Regional scales. There is a positive effect of the interaction between number of families and the Local scale, with the proportion of multimodality *vs* 1PLN increasing as the number of families increases ($pMCMC=0.00106$). In other words, relative to 1PLN, multimodality is significantly less prevalent at Local scales and low family richness, compared to when family richness is higher or spatial extent is Regional or Continental. Conversely, logseries is less prevalent *vs* 1PLN at Continental scales ($pMCMC=0.01636$), and more prevalent at Regional and Local scales ($pMCMC=0.00923$ and $pMCMC=0.01578$, respectively; Fig. 4.3 and Table 4.2). These effects are

independent of number of families, which does not influence significantly the proportion of logseries vs 1PLN.

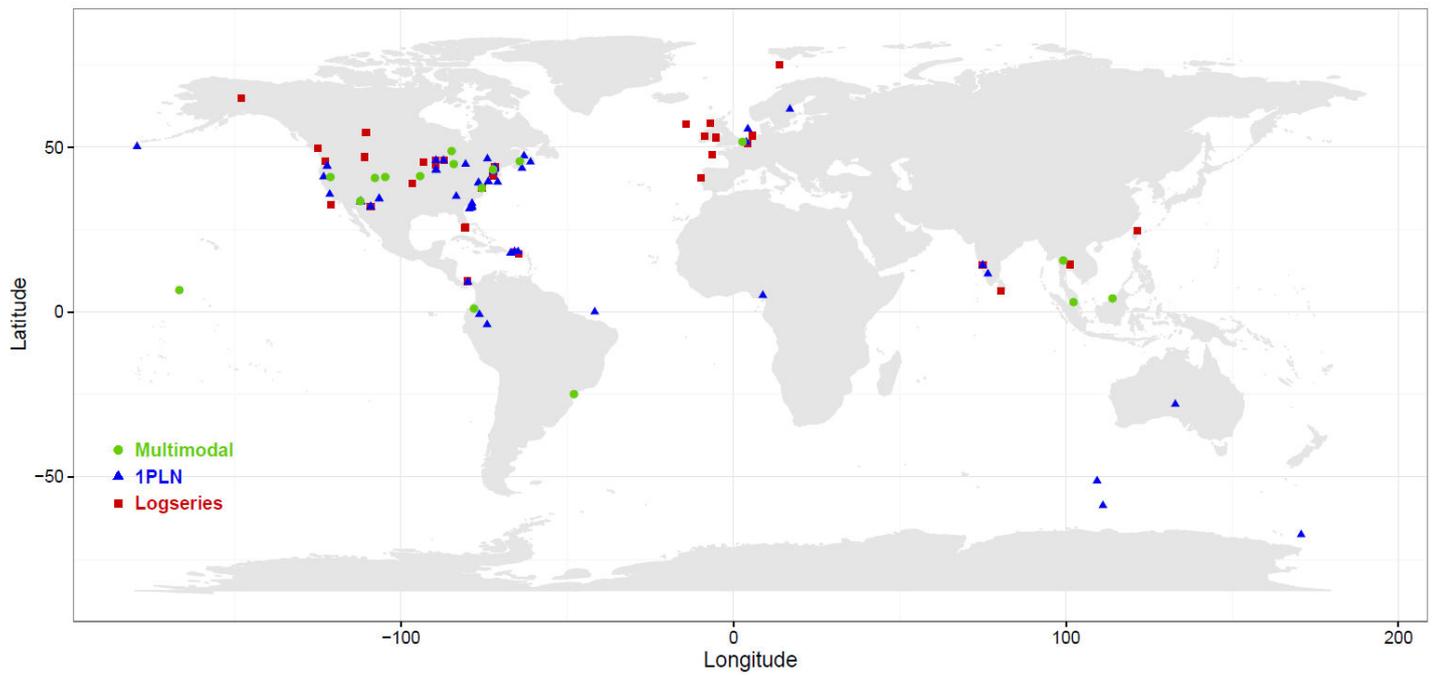


Figure 4.1 Map showing the 117 empirical SADs sampling locations and the model selected as best fit (each point corresponds to the mean latitude-longitude).

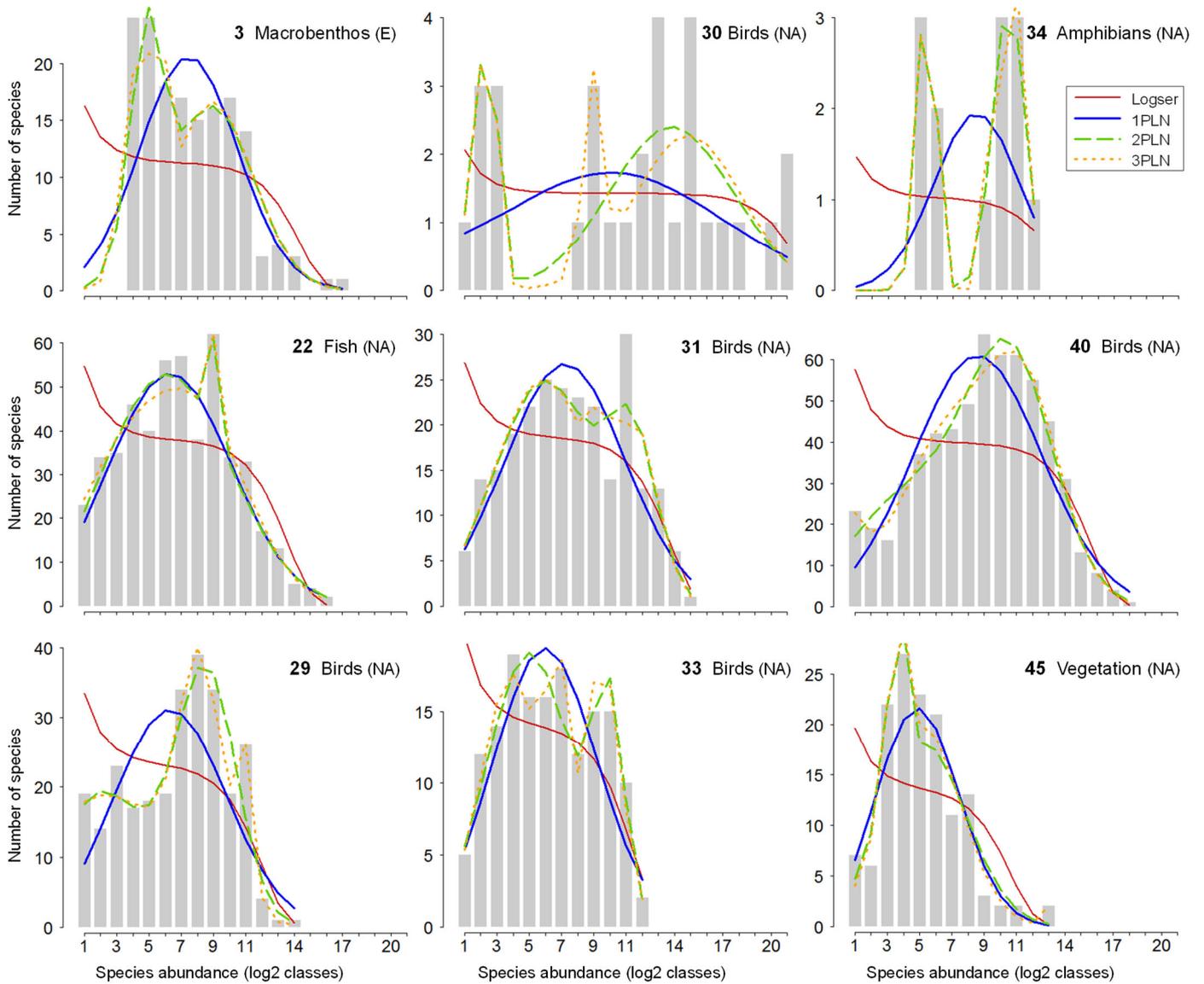
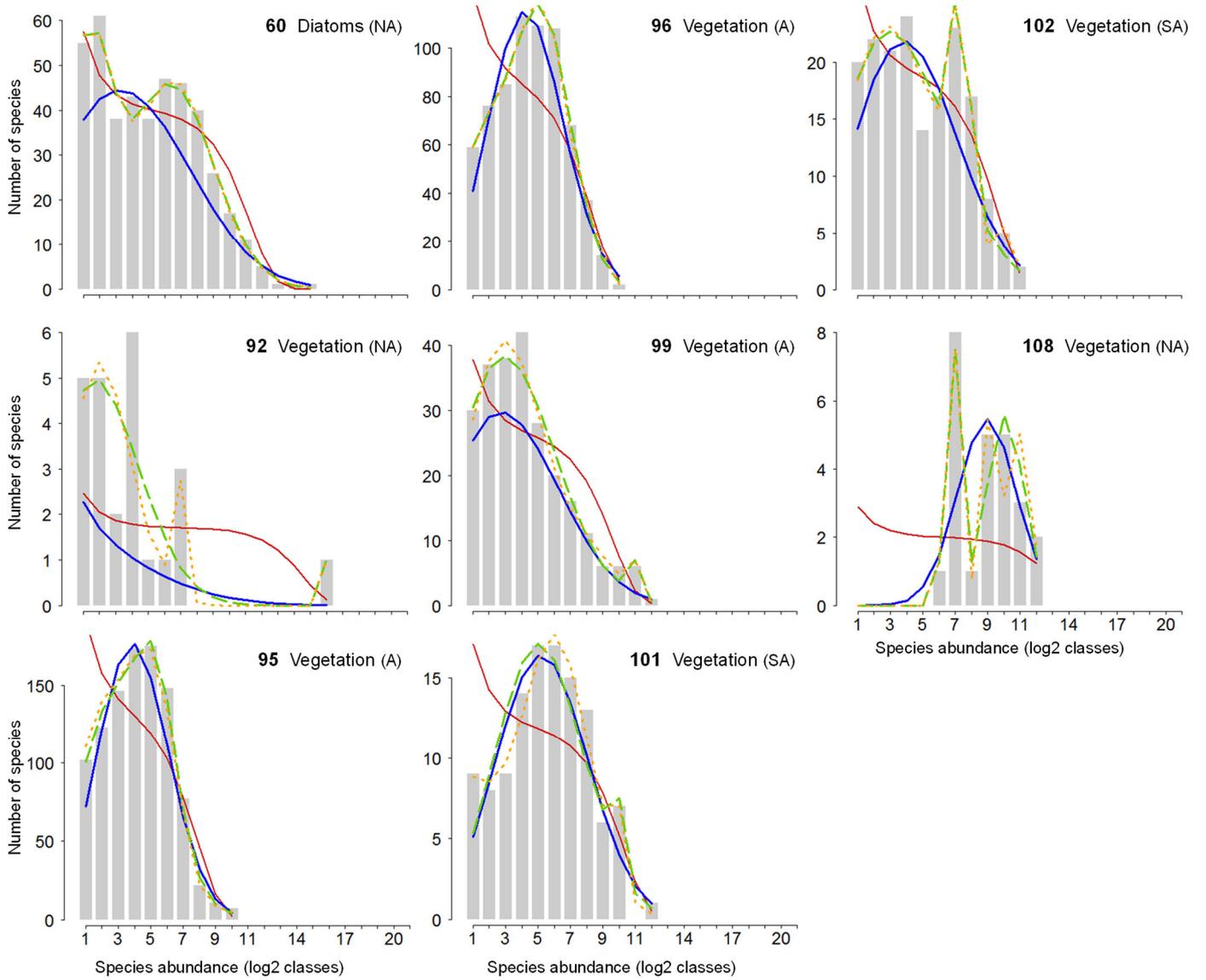


Figure 4.2 Species abundance distributions (SADs) of the empirical datasets selected as multimodal with high confidence, identified by the corresponding ID. For each SAD the taxon is identified, as well as the SAD location: A – Asia; E – Europe; NA – North America; and SA – South America. All the fitting routines were run on non-binned data. SADs were plotted with bins representing true doubling classes of abundance, following Gray *et al.* (2006). For all SADs the y-axis is the number of species and the x-axis is the species abundance in log₂ classes (the first bar represents species with abundance 1, the second one species with abundances 2-3, then 4-7, 8-15, etc). The best fitted curves are red line for the logseries, bold blue line for 1PLN, dashed green line for 2PLN and dotted orange line for 3PLN (continued next page).



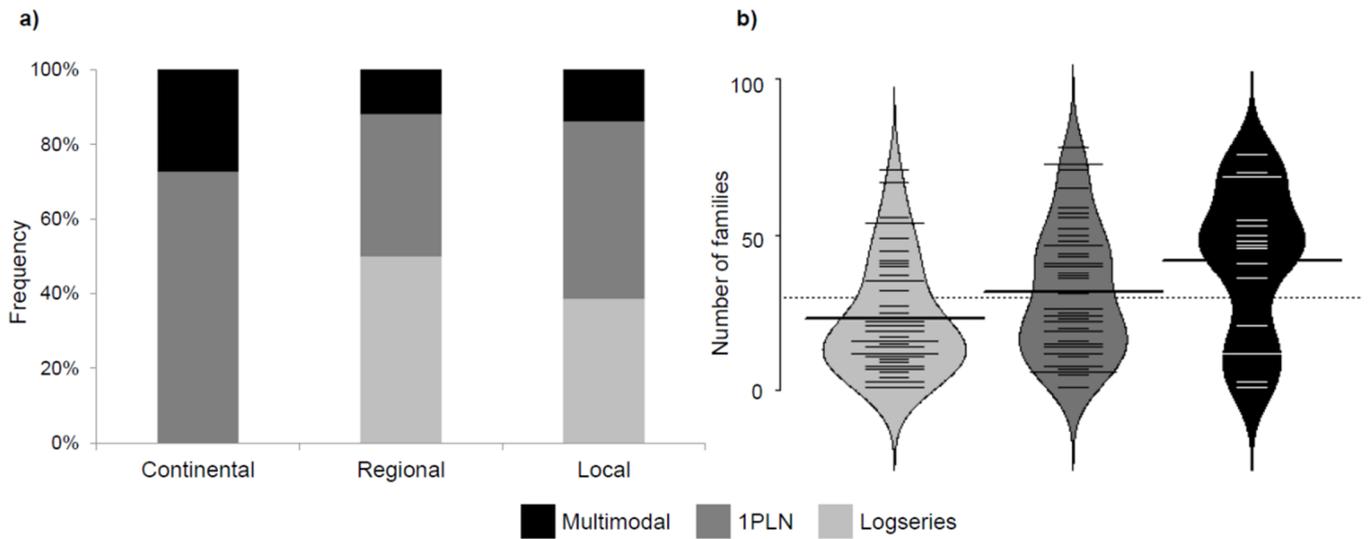


Figure 4.3 Model selection frequency versus spatial extent of species abundance distributions (SADs) (a) and taxonomic breadth as number of families (b). The absolute number of SADs per spatial extent is 11 continental, 42 regional and 64 local. The number of families was truncated at 100 for better visualization; the four SADs with the greater number of families were best fit by the one Poisson lognormal distribution (1PLN).

Table 4.1 Parametric bootstrap results comparing 1PLN vs 2PLN and vs 3PLN for the empirical SADs selected as multimodal by AIC_c ; for each SAD, the estimated 1PLN parameters are shown. The critical value is the equivalent of a p-value from the PBLRT distribution and the *empirical* LRT is calculated from the SAD fitting. Bold LRT values indicate support for a multimodal distribution (*empirical* LRT \geq bootstrap p-value).

Estimated 1PLN parameters			Likelihood Ratio test			
			1PLN vs 2PLN		1PLN vs 3PLN	
ID	$\hat{\mu}$	$\hat{\sigma}$	Parametric bootstrap p-value	<i>Empirical</i> LRT	Parametric bootstrap p-value	<i>Empirical</i> LRT
2	24.47	1.94	9.377	7.6038	14.688	13.811
3	124.50	2.10	11.160	21.297	18.693	35.714
4	19.21	5.31	14.707	13.430	20.241	16.606
9	95.30	5.23	12.095	9.246	19.752	14.435
10	150.78	5.46	12.745	6.808	17.386	7.404
12	3.54	2.92	12.588	8.644	18.103	11.315
13	6.76	2.83	10.674	6.860	15.502	14.417
14	5.75	3.92	12.204	10.441	16.467	13.752
16	5.46	4.01	10.761	7.396	18.038	12.537
18	0.26	4.14	11.497	7.719	17.555	10.329
21	0.70	4.30	11.873	9.657	16.518	12.368
22	58.47	2.61	13.907	17.693	19.992	21.597
23	23.27	3.94	11.580	6.398	18.328	16.761
24	1.18	2.55	10.825	8.463	17.580	9.422
29	57.55	2.39	12.481	43.531	23.296	55.787
30	747.59	4.78	12.940	17.442	17.610	22.306
31	107.28	2.56	14.647	15.502	20.957	17.948
33	45.10	2.20	15.236	17.542	21.788	20.903
34	249.84	1.83	11.724	15.723	18.878	18.941
35	4.75	3.00	12.523	8.538	16.868	10.088
38	35.70	2.47	12.942	11.105	19.575	13.125
39	41.27	2.56	14.331	12.185	23.251	19.210
40	275.14	2.74	15.717	35.574	24.019	42.367
41	5.42	2.95	13.875	10.999	18.959	10.999
43	1.27	4.48	11.558	6.698	16.688	11.533
45	20.27	1.79	11.052	11.048	16.423	20.835
49	4.00	3.69	11.392	7.905	16.739	10.408
50	9.60	2.58	12.370	8.004	19.615	18.391
51	1.12	3.48	12.745	6.982	18.678	8.747
55	18.57	2.26	14.210	8.282	24.062	11.914
60	8.96	2.82	10.519	22.863	16.974	23.217
90	4.64	2.78	13.386	2.968	19.279	16.158
91	12.07	2.54	12.866	6.831	17.417	7.803
92	0.19	3.86	12.789	16.989	20.083	22.405
93	9.29	2.13	10.707	7.517	16.712	5.097
94	7.95	2.71	11.788	8.197	18.315	13.076
95	9.91	1.57	12.813	25.698	17.673	31.901
96	12.95	1.63	15.230	19.944	26.657	22.725
99	5.65	2.39	12.735	15.141	17.760	17.325
100	25.51	2.82	12.768	12.234	19.488	16.313
101	26.19	1.97	12.369	10.358	15.550	15.938
102	11.03	2.26	12.463	14.316	19.487	20.841
104	16.14	2.03	11.724	10.474	19.039	12.096
108	348.56	1.25	12.196	14.874	17.454	18.779
110	135.25	2.10	12.602	11.018	21.157	17.723
113	49.12	1.74	13.111	8.733	20.980	17.107
116	22.50	2.11	12.442	9.184	17.841	13.840

Table 4.2 Results of binomial (a) and multinomial (b) generalized linear model (GLM) fitting, showing a positive effect of spatial scale or higher taxonomic breadth on the prevalence of multimodality.

a) Binomial GLM	Estimate	SE	z value	Pr(> z)
SpatialExtent.Continental	-0.2207	1.0135	-0.2180	0.8277
SpatialExtent.Regional	-1.5511	1.2830	-1.2090	0.2267
SpatialExtent.Local	-3.7396	1.3940	-2.6830	0.0073
NumberFamilies	-0.0127	0.0154	-0.8230	0.4105
SpatialExtent.Regional : NumberFamilies	0.0060	0.0247	0.2430	0.8084
SpatialExtent.Local : NumberFamilies	0.0747	0.0260	2.8730	0.0041

b) Multinomial GLM	Posterior mean	Lower 95% CI	Upper 95% CI	pMCMC
Reference: 1PLN-SpatialExtent.Continental				
Multimodality : SpatialExtent.Continental	-0.0013	-2.4390	2.3690	0.9930
Multimodality : SpatialExtent.Regional	0.0183	-3.1700	3.2370	0.9944
Multimodality : SpatialExtent.Local	-3.7030	-6.9170	-0.5603	0.0194
Multimodality : SpatialExtent.Continental : NumberFamilies	-0.0248	-0.0669	0.0098	0.1597
Multimodality : SpatialExtent.Regional : NumberFamilies	-0.0124	-0.0805	0.0541	0.7079
Multimodality : SpatialExtent.Local : NumberFamilies	0.0897	0.0301	0.1519	0.0011
Logser : SpatialExtent.Continental	-186.0000	-336.7000	-0.4936	0.0164
Logser : SpatialExtent.Regional	188.1000	1.9480	338.0000	0.0092
Logser : SpatialExtent.Local	186.2000	1.4960	337.8000	0.0158
Logser : SpatialExtent.Continental : NumberFamilies	-1.7840	-4.5770	0.7772	0.3401
Logser : SpatialExtent.Regional : NumberFamilies	1.7330	-0.8160	4.5390	0.3565
Logser : SpatialExtent.Local : NumberFamilies	1.7650	-0.8011	4.5560	0.3454

Pr, two-tailed p-value; pMCMC, Markov chain Monte Carlo *P*-values; PLN, Poisson lognormal distribution.

For the Bayesian GLM, the posterior mean estimates, the 95% credible intervals and the pMCMC values are shown. The parameter estimates were considered statistically significant when pMCMC < 0.05, and the 95% credible intervals (CIs) did not include 0. The term ‘Multimodality : SpatialExtent.Continental : NumberFamilies’ refers to the estimation of multimodality versus 1PLN at the continental scale with the interaction with number of families.

4.3 Discussion

This investigation showed that 17 out of 117 SADs analysed are multimodal with high confidence (~15%). Furthermore, there is a higher prevalence of multimodality for communities with broader spatial scale or higher taxonomic breadth, suggesting that multimodality increases with ecological heterogeneity. This warrants systematic consideration of multimodality in the quantification of SAD shape.

This analysis across different taxa, biomes and species richness indicates that multimodality is not an artefact of particular SADs. The only particularity of the SADs analysed here is that they were intensely sampled (more than 10,000 individuals), and there is no reason to suspect that this holds any influence as to whether the underlying ecological community is multimodal. Furthermore, because each empirical SAD analysed corresponded to only one year of sampling, multimodality reflects the structure of the community at a particular point in time. Additionally, multimodality was inferred only when it is supported by both AIC_c and PBLRT. Moreover, false negatives were more prevalent than false positives in the simulation study, thus rendering these conclusions highly conservative. A caveat of this study is that the SADs analysed here did not fully represent the spectrum of community variability in terms of spatial and taxonomic coverage. Furthermore, the sample of SADs analysed was not intended to be representative of taxa, habitat, climatic regions or even realm. Nevertheless, these results show a positive effect of both spatial scale and taxonomic breadth on the prevalence of multimodality, regardless of taxa and realm.

The prevalence of multimodality found in this study differs from that suggested by Barabás *et al.* (2013). The simulation study showed that depending on the parameter combination, sampled communities from a single PLN can indeed produce apparently multimodal SADs, as the authors suggested. However, I believe that the method developed here improves our ability to test for multimodality. Despite there being no direct correspondence between Barabás *et al.*'s parameterization and the one in this study, their Fig. 4 suggests that the mode of the average unimodal distributions is located around octave 6 of the SAD, with the distributions spanning 11 octaves. This could be compared to the 1PLN simulations for larger μ , σ and species richness values, which fall in the parameter space for which AIC_c had a higher chance of erroneously selecting multimodality.

Thus, it would be interesting to investigate whether performing the additional LRT to the SADs generated using Barabás *et al.*'s parameterization would still yield similar multimodality frequencies.

General explanations for multimodality

Scale is fundamental to understanding biodiversity patterns (Levin, 1992; McGill, 2010a). The results presented here indicate that multimodality is more likely to occur for regional to continental-scale SADs, albeit not exclusively. Some SADs selected as multimodal consist of local samples or plots, but all of these are taxonomically diverse (between 12 and 76 families): ID3 consists of macrobenthos samples from the Belgian Continental Shelf; IDs 95 and 96 of tropical forest plots in Malaysia, ID99 in Thailand, and IDs 101 and 102 of tropical plots in Brazil and Colombia, respectively; and IDs 45, 92 and 108 consist of vegetation plots in the USA (desert, shortgrass steppe and dune vegetation, respectively). This matches the regression analysis performed, for which local SADs with low family richness exhibited lower prevalence of multimodality than it did at high family richness or broad spatial scales.

The explanatory variables analysed in this study mirror the spatial and organizational scales suggested by Levin (1992) as underpinning the variability of ecological patterns, and they support previous explanations for multimodality. Multimodality has been proposed to arise as consequence of species differences in ecological or functional characteristics (e.g. Magurran & Henderson, 2003; Alonso *et al.*, 2008) and of environmental heterogeneity (Dornelas *et al.*, 2009). Both of these explanations are consistent with a greater prevalence of multimodality in communities with greater spatial extent or taxonomic diversity. The goal of this study was not to develop a predictive model for multimodality, but to quantify its prevalence and test its association with relevant ecological variables. Exploring the effects of environmental heterogeneity, functional diversity, and core-transient species in more detail will prove a fruitful avenue to further understand what aspects of ecological heterogeneity affect SAD shape and lead to multimodality.

An additional interesting research question is how temporal variability in the species abundances might affect SADs' shapes over time. In the present study, I was interested in removing the potential effect of temporal fluctuations of the relative abundances of species across years, to avoid the

possibility that multimodality could arise as an artefact of a single mode changing position over time. In principle, it is also possible that pooling could reduce multimodality, if changes in the position of modes over time make multiple modes more difficult to detect (for instance, if multimodality arises as a transient feature of communities, as an effect of particular stochastic environmental effects). Because the models used in this study implicitly account for sampling effects, and require actual counts (number of individuals sampled), an investigation into the effects of temporal averaging would require the development of an alternative statistical approach.

Rarity and commonness

SAD studies have often focused on the left-hand side of the distribution and on different theoretical models' ability to accommodate the rarest species mode (e.g. Hubbell 2001; McGill 2003b), and several studies have described the rarer mode as the one leading to a multimodal pattern (Magurran & Henderson, 2003; Borda-de-Água *et al.*, 2012; Matthews *et al.*, 2014). Although a mode was often fitted to the rarest species, some of the empirical SAD also exhibited modes for very abundant species (e.g. IDs 30, 92, 99 and 108 in Fig. 4.2). This highlights the observation that communities characterized by very high abundances of the most abundant species might not be accommodated within a single lognormal SAD, and a multimodal model provides a better description, similarly to communities with a very high prevalence of rare species. While the majority of species are rare and the universal 'hollow-curve' SAD is the definitive description of this, the few most common species disproportionately dominate communities in terms of abundance and ecological processes (Gaston, 2010, 2011), and might also have considerable influence on SAD shape (e.g. Connolly *et al.*, 2014).

The logseries was selected as best model relatively frequently, despite all of the data coming from intensely sampled communities. This suggests that, even for high sampling intensity, some communities are characterized by a very high proportion of rare species. The logseries was more often selected for communities encompassing smaller spatial scales, a finding consistent with the regression analysis results. Additionally, visual inspection suggests that there was a slight tendency for the logseries to be favoured when species richness was lower (see also results from Chapter 5), and in this analysis logseries was never the model with the best absolute fit (in terms of negative log-likelihood values only; c.f. Baldrige *et al.*, 2015). Interestingly, none of the SADs selected as logseries had the largest spatial extent, contrasting with the predictions of neutral theory with point-

mutation speciation (Hubbell, 2001), which predicts a logseries SAD for the metacommunity. On the other hand, the maximum entropy theory of ecology (METE; Harte *et al.*, 2008) predicts a logseries SAD, contrasting with the support for multimodality found in this study, and with the effect of spatial scale and taxonomic breadth on model frequency.

This investigation showed that multimodality occurs with high confidence in ~15% of the assemblages analysed. Additionally, I demonstrated that multimodality has higher prevalence for large scale or taxonomically heterogeneous communities. Broader spatial extent and higher taxonomic breadth (as measured by family diversity) underpin higher ecological heterogeneity, and hence these factors are suggested as potential explanations for multimodality in SADs.

Conclusions

Multimodal SADs occur at a non-negligible frequency. Larger spatial scale or higher taxonomic breadth can yield multimodal SADs. Greater spatial scale and taxonomic breadth of the communities imply higher ecological heterogeneity. In turn, this is expressed as different levels of species abundance, thus being reflected in the SAD shape and informing on community structure. This investigation showed that the dichotomy between logseries and lognormal as the sole adequate descriptors of SAD should be expanded to include multimodal models. This will enhance our ability to use SADs to detect the effects of ecological or functional mechanisms affecting the communities. Furthermore, differences in SAD shape across different scales provide important insights to the current endeavour of biodiversity scaling.

5. Species Abundance Distributions across scales

5.1 Introduction

SADs describe the relative abundance of the species in a community for a space and time, accounting for different aspects that univariate metrics measure separately, and readily integrating concepts such as rarity and dominance. Understanding how SAD shape varies with sampling scale is a long standing question in ecology (Fisher *et al.*, 1943; Preston, 1948; Zillio & He, 2010). Following the results from the previous chapter, showing that there is a higher prevalence of multimodal SADs for communities encompassing large spatial extents or higher taxonomic diversity, I performed an exploratory analysis of how SAD shape changed across a gradient in spatial scale. To my knowledge, this is the first empirical assessment of the shape of SADs across a scale gradient spanning several orders of magnitude and including different taxa. The overall goal of this chapter was to try to reconcile predictions for SAD from two different macroecological theories, with sampling theory predictions and empirical patterns. Specifically, I tested empirically the common better fit of lognormal distributions for larger samples on one hand, and the results in the previous chapter showing that multimodality occurs with higher prevalence for larger areas, on the other.

Two approaches to the scaling of SADs can be taken: downscaling and upscaling. In the former, some sampling approach is used to predict sampled SADs at smaller scales (Hubbell, 2001; Green & Plotkin, 2007), while in the latter some statistical method is employed to infer SADs for larger spatial scales than the focal one (Zillio & He, 2010; Borda-de-Água *et al.*, 2012). However, analysing SADs across different spatial scales has remained largely unexplored (but see Rosindell & Cornell, 2013). There is no *single adequate* scale for studying SADs (Wiens, 1989; Levin, 1992), hence systematically assessing SADs at different spatial scales allows us to make stronger inferences about the commonness and rarity of species across scales, and potentially disentangle which processes are determinant at different scales.

Sampling effects have long been recognized to severely affect SAD shape (e.g. sample size, random sampling from a metacommunity), as well as spatial scale (Fisher *et al.*, 1943; Preston, 1948; Pielou,

1977; Hubbell, 2001; Connolly *et al.*, 2005; Green & Plotkin, 2007). Generally, SAD studies have employed a sampling theory approach to the problem of analysing the relationship between the large regional community and a sampled local SAD. Fisher *et al.* (1943) proposed the logseries as a distribution for a random sample from a gamma distribution (SADs at smaller scales are random subsamples of SADs at larger scales). The “veil line” proposed by Preston (1948) was a first approximation to explain that the absence of rare species in small samples would lead to a truncation of the “true” underlying lognormal distribution. Numerous studies have shown that increasing sampling intensity does include more rare species in the empirical SAD. However, as subsequently shown by several authors, unveiling does not simply reveal the left-end of the distribution by rigidly moving the veil, but the shape of the distribution also changes (Pielou, 1977; Dewdney, 1998; McGill, 2003c). Specifically, McGill (2003c) showed that pooling repeated autocorrelated small samples can lead to the log-left-skew reported in many empirical SADs, i.e. the existence of more rare species than predicted by a lognormal distribution (Hubbell, 2001; Magurran & Henderson, 2003). This phenomenon can nonetheless be driven by a biological mechanism, where SAD shape reflects changes in the community structure, e.g. signature of core-transient species temporal dynamics (Magurran & Henderson, 2003).

Crucially SAD shape is affected by how species are distributed in space, and one of the fundamental patterns in ecology is that individuals are not randomly distributed across space (Condit *et al.*, 2000; McGill, 2010b). Green & Plotkin (2007) developed a statistical sampling theory for SAD incorporating conspecific spatial aggregation patterns. They showed that when sampling from regional SADs with randomly distributed populations (Poisson sampling), the sampled SADs would exhibit the same functional form as the regional SAD. In contrast, using a more realistic description of species spatial aggregation patterns (negative binomial sampling), sampled SADs diverged from the regional SAD. Specifically, this conspecific aggregation led to sampled SADs skewed towards both rare and more abundant species. Using theoretical models, Alonso *et al.* (2008) incorporated species asymmetries in terms of sampling or biological properties, and suggested that interspecific differences in aggregation rates can produce bimodal abundance distributions. Using an exceptionally large empirical sample, Dornelas & Connolly (2008) reported that species spatial aggregation partially explained the existence of multiple modes in a coral SAD (two, but not the three modes). Nonetheless, using a completely different approach and attempting to extrapolate SADs for larger areas, Borda-de-Água *et al.* (2012) predicted a bimodal larger scale-SAD, employing the method of moments without including any information on species aggregation patterns. Bimodality arises from an increase in the number of rare species with area (Borda-de-Água *et al.*,

2002, 2012), with one mode occurring for the singletons class and another mode for intermediate abundance classes. Previously, the multifractal approach proposed by Borda-de-Água *et al.* (2002) predicted that species abundance distributions obtained at different areas should collapse into a single curve after renormalization.

Two unified theories of biodiversity make predictions for SAD shape at different scales, the Neutral Theory (Hubbell, 2001) and the Maximum Entropy Theory of Ecology (METE) (Harte *et al.*, 2008). Both theoretical frameworks can be thought of as null models (functional equivalence between individuals, and no explicit mechanisms included, respectively). Systematic discrepancies between empirical data and theoretical predictions can help identify important mechanisms that should be accounted for in order to improve our ability to make stronger inferences about what is driving SAD shape. Both theories provide a suitable null expectation for what SAD shape should occur at different scales.

Neutral theory assumes all the individuals on the same trophic level have the same demographic rates of birth, death, dispersal and speciation, irrespective of species identity (Hubbell, 2001). Assuming equivalent demographic rates and fitness, stochastic drift and dispersal limitation are the processes explaining patterns of species abundance. The spatially implicit neutral model includes two distinct spatial scales: a local community that consists of a dispersal-limited sample from the metacommunity. In the original model (assuming the “point mutation” mode of speciation), the metacommunity follows a logseries distribution and the local community follows a zero-sum multinomial distribution (ZSM), which includes fewer rare species than the logseries and resembles a left-skewed lognormal distribution (Hubbell, 2001; Rosindell *et al.*, 2011). This latter distribution has been the focus of intense debate and numerous studies have assessed the ZSM and lognormal performances for empirical SADs (Hubbell, 2001; McGill, 2003b; Volkov *et al.*, 2003, 2007; Dornelas *et al.*, 2006). Several subsequent models have been developed that incorporate more realistic ecological settings, relaxing the neutrality assumption for some ecological characteristics, or including several local communities with different dispersal limitations linked to the metacommunity (Etienne, 2005, 2007, 2009; Janzen *et al.*, 2015). Recently, using a spatially explicit neutral model, Rosindell & Cornell (2013) have also derived a logseries SAD for the largest scale analysed, and while more realistic speciation modes have been proposed, data from 20 different local tree communities were actually better fitted by Hubbell’s original model (Etienne *et al.*, 2007). Finally, Rosindell *et al.* (2010) proposed a gradual protracted speciation mode as an improvement

on the classical neutral model that can produce both logseries and the “difference logseries (DLS)” distributions for the metacommunity level. The latter distribution is the difference between two logseries terms, it predicts fewer rare species, and reduces to a standard logseries when the number of generations equals 0 (equivalent to point mutation model). However, a logseries distribution is dependent on the size of the metacommunity, and the authors argue that the protracted model will provide a better fit only if there is a sufficiently large sample from the metacommunity. Therefore, assessing the performance of logseries for intensely sampled and large scale communities SADs, and for different taxa, provides a relevant test on current neutral models.

METE is a spatially explicit theory of biodiversity based on the principle of maximization of information entropy (MaxEnt). It requires knowledge only on four state variables to describe ecological communities: the area of an ecosystem (A_0), species richness (S_0), the total number of individuals (N_0), and total metabolic rate for the overall community (E_0) for a specified taxonomic group (original formulation ASNE model; the metabolic rate information has been disregarded when analysing SAD) (Harte *et al.*, 2008; Harte & Newman, 2014). MaxEnt rationale is that the least-biased inference of the shape of a probability distribution is as smooth and flat as possible given the constraints (Harte *et al.*, 2008). Using only these four state variables, i.e. what is known about the system *a priori* and without incorporating any specific ecological mechanisms, the most likely distribution for several macroecological patterns is found by maximizing information entropy. The logseries is the SAD distribution that emerges from the METE model across scales (Harte *et al.*, 2008; Harte & Newman, 2014).

Here, I performed a systematic assessment of SAD shape for different taxa across a scale gradient, employing the model fit comparison from the previous chapters, and assessed the effect of sampled area, taxonomic breadth, species richness and total abundance on SAD shape. I interpret the results in light of these two macroecological theories with explicit predictions for the expected SAD shape at different scales, hoping to reconcile the discrepancies between different theoretical predictions and empirical SADs.

5.2 Methods

5.2.1 Empirical Data

12 datasets were analysed¹, comprising different taxa, namely birds, fish, benthos and trees. I selected 11 datasets from the BioTIME database with spatial extent larger than 150 000 km² and for which the unique sampling locations were distributed across the study area so that the random splitting of the total extent would not result in portions without sampling locations (see section 5.2.2). For each dataset, the data corresponding to one year of sampling was used (the year with the most and more evenly distributed sampling locations). I also analysed the Forest Inventory and Analysis Database (FIA; <http://fia.fs.fed.us/>; USDA Forest Service, 2010; Woudenberg *et al.*, 2010), as I wanted to include a tree community data in this analysis to ensure the results are robust across taxonomic groups. I obtained the latter data using the EcoData Retriever (<http://data-retriever.org>; Morris & White, 2013; McGlenn & White, 2015), and selected data from 2013 only. For each dataset, information of the taxonomic family corresponding to each species was also retrieved. These empirical datasets cover a wide range of sampling grains (0.0001 to 400 km²) and total spatial extents (167 455 to 16 663 141 km²). The full list of datasets and their sources can be found in Table 5.1¹.

5.2.2 Implementing the scale gradient

All analyses were performed in the statistical software R (R Core Team, 2017). I established a scale gradient by using the fixed extent of the study area and systematically partitioning this area into smaller portions, thus varying “grain” size as follows. I drew a circle encompassing all the sampling locations from the community data and centred on the centroid of the sampling locations. A random point from the circle was selected to split the circle into halves, thirds, quarters, eighths and sixteenths, using the initial random point from the bisection as reference (Fig. 5.1). This way the spatial relationship between the sections and the sampling locations is maintained. Within each section, species abundances were pooled to build the Species Abundance Distributions, thus two SADs were produced for the bisection level, three for the third level, and so forth. At each level, each section

¹ For the analyses in chapters 5 and 6, 12 datasets were selected; however one of the datasets was excluded from this chapter due to model fitting issues – ID11 in Table 5.1 was not included in chapter 5 results.

was annotated with the area, species richness (S), total abundance (N), as well as the total number of families (to assess the effect of taxonomic diversity). The areas sampled were calculated using convex hull polygons encompassing the sampling locations within each section, using package rgeos (Bivand & Rundel, 2016) (Fig. 5.1).

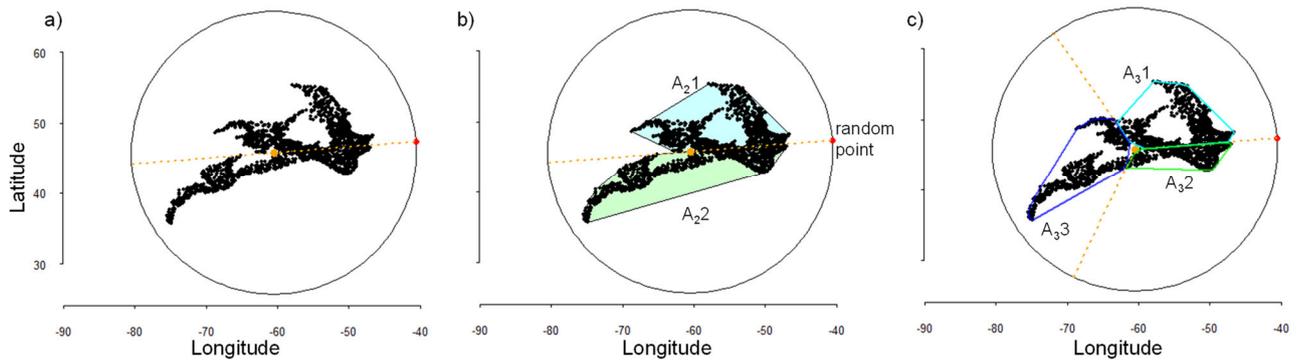


Figure 5.1 Schematic representation of the scale gradient, showing an example of how the encompassing circle was drawn and a random point was selected to establish bisections (a and b) and thirds (c).

5.2.3 Model fitting and analysis

Along the scale gradient and for each section, each SAD was fitted with the four alternative models as described in sections 3.2.1 and 3.2.2. I employed the maximum likelihood methods described in Chapter 3 to explicitly compare the fit of logseries distributions (Fisher *et al.*, 1943), and of mixtures of 1, 2 and 3 Poisson Lognormal distributions (1PLN, 2PLN and 3PLN, respectively) (Pielou, 1969; Bulmer, 1974). Model fitting was also performed for the SAD corresponding to the total extent. The second order Akaike's information criterion for small sample sizes (AIC_c , Burnham & Anderson, 2002) was used for model selection. In this investigation, I used AIC_c to compare the models, since the simulation study carried out illustrated that BIC was too conservative and can be insensitive to deviations in SADs shape (section 3.4). Furthermore, because I was not interested in detecting multimodality *per se*, but rather in detecting changes in SAD shape across scales, I used the best model as selected by AIC_c regardless of ΔAIC_c support².

I used the R package ggplot2 (Wickham, 2009) to plot smoothed density estimates relating the model selected (best model according to AIC_c) with the relevant variables across the scale gradient, namely area sampled, species richness, total abundance and number of families. I built these plots for each community individually and for all the SADs together, hence providing an overview of how these variables affect SAD shape across the different taxa analysed. In addition to using AIC_c as a model selection criterion, I also quantified the deviations between the empirical SADs and the predictions of each model, comparing the observed and expected number of species per octave.

Finally, I also described how a suite of α diversity metrics varied along the scale gradient for each community, namely species richness (S), total number of individuals (N), number of families, Fisher's α (Fisher *et al.*, 1943; Magurran, 2004), as the parameter estimated from the logseries fitting, and also Shannon's Diversity (H') (Pielou, 1975) and Evenness (J'). I plotted these metrics as a function of log₁₀ area. Specifically for the Shannon's Diversity, this allowed me to compare the results with a power-law relationship between the diversity index and area predicted by a multifractal approach (Borda-de-Água *et al.*, 2002).

² For dataset ID4 in this chapter, model selection for the total extent SAD was informed from the results in chapter 4, since the data analysed was the same as in the multimodality analysis (same year in both analyses).

Table 5.1 Community data used and data sources. For each community the taxon, species richness, spatial extent and grain are shown (data sources can be found in Appendix II).

ID	Dataset Title	Taxon	Usage notes	Spatial extent (Km ²)	Grain (Km ²)	Number of species	Number of samples	References
1	East Coast North America Strategic Assessment - ECNASAP	Fish	1994	7 229 693	0.33336	110	2 101	Brown et al., 2005
2	North American Breeding Bird Survey (BBS)	Birds	2015; USA data only (excluded Alaska)	13 104 786	25.42715	521	2 420	Pardieck et al., 2016
3	Reef Life Survey (RLS): Global reef fish dataset	Fish	Spatial subset around Australia	572 747	0.0005	1 847	6 666	Edgar & Stuart-Smith, 2014a,b
4	ICES North Sea International Bottom Trawl Survey for commercial fish species	Fish	2011	2 726 171	0.33336*	131	688	DATRAS, 2010c
5	Snow crab research trawl survey database (Southern Gulf of St. Lawrence. Gulf region. Canada) from 1988 to 2010 (OBIS Canada)	Benthos	2009	167 455	0.00642	32	354	Wade, 2011
6	Maritimes Breeding Bird Atlas (2006-2010) point count data	Birds	2009	480 235	0.031416*	163	3 243	NatureCounts a
7	Reef Life Survey (RLS): Invertebrates	Invertebrates	Spatial subset around Australia	572 747	0.0001	1 013	6 817	Edgar & Stuart-Smith, 2008, 2014b
8	Irish Ground Fish Survey for commercial fish species. ICES Database of trawl surveys	Fish	2004	967 879	0.177792	100	163	DATRAS, 2010d
9	Landbird Monitoring Program (UMT-LBMP)	Birds	2004	1 057 570	0.031416*	229	5 107	USFS
10	Ontario Breeding Bird Atlas (2001-2005) point count data	Birds	2003	3 545 420	0.031416	233	19 611	NatureCounts b
11 [‡]	North Pacific Groundfish Observer	Benthos	1993	6 794 596	400	220	1 007	North Pacific Groundfish Observer Program
12	Forest Inventory Analysis (FIA)	Trees	2013; excluded Alaska	16 663 141	0.004047	305	19 427	USDA Forest Service, 2010; Woudenberg et al., 2010

[‡]Dataset 11 was not included in chapter 5 analysis

*Grain was approximated to similar studies

5.3 Results

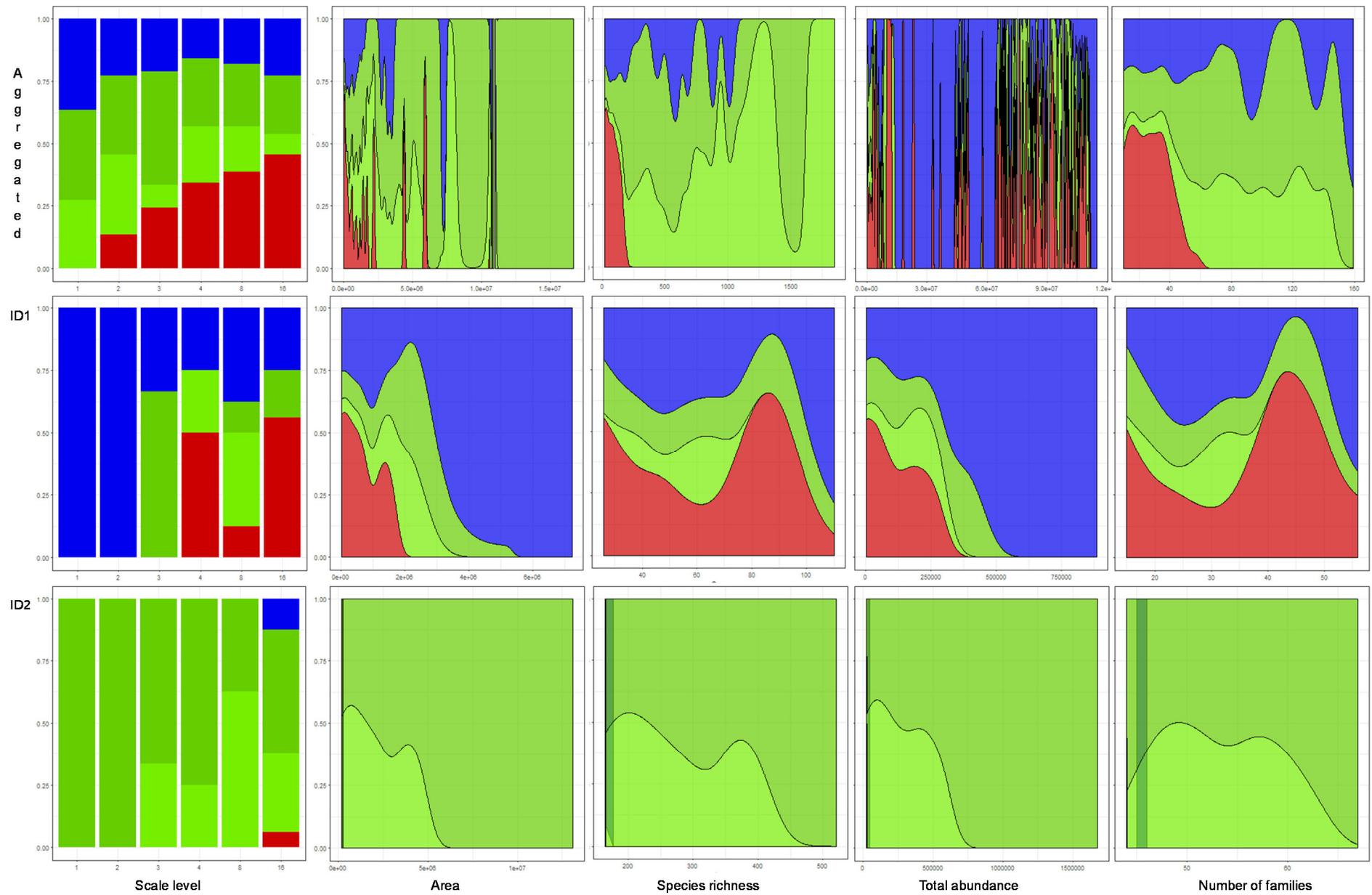
Overall, there was a higher prevalence of multimodal SADs for larger areas and for more taxonomically diverse communities (Fig. 5.2, aggregated communities (top row)), while nonetheless some smaller areas or less diverse communities were also multimodal. Logseries was never selected as best model for the total extent SAD, and was only selected for much smaller areas, and when species richness or number of families were proportionally much smaller (Figs. 5.2 and VI.1). As area sampled decreases both species richness (S) and total number of individuals (N) are also expected to decrease; however, while S showed a similar effect to that of area on model selection, there was no clear pattern for N (Fig. 5.2 top row). Note that not all the communities were multimodal at the total extent. For the SADs selected as multimodal with strong support at the total extent, multimodal models most often provided the best fit across the scale gradient. These are the BBS bird data, the RLS fish data, the Ontario Breeding Bird Atlas (OBBA), and the FIA tree inventory. The average ΔAIC_c for multimodality vs non-multimodality across the scale gradient was 11.01 for BBS, 9.77 for RLS fish, 6.39 for OBBA, and 5.53 for FIA (calculated as $(\min AIC_c_{2PLN/3PLN} - \min AIC_c_{1PLN/logser})$ for all the sections). On the other hand, some communities exhibited the expected pattern of progressing from multimodality to 1PLN or logseries as sampled area decreased (Fig. 5.2, IDs 6, 8 and 9). The communities better fit by 1PLN at the total extent showed some variability in the best fit models as area decreased, with 1PLN still being selected very frequently, but with both logseries and multimodal models being selected for smaller or intermediate levels (Fig. 5.2, IDs 1, 4, 5 and 7).

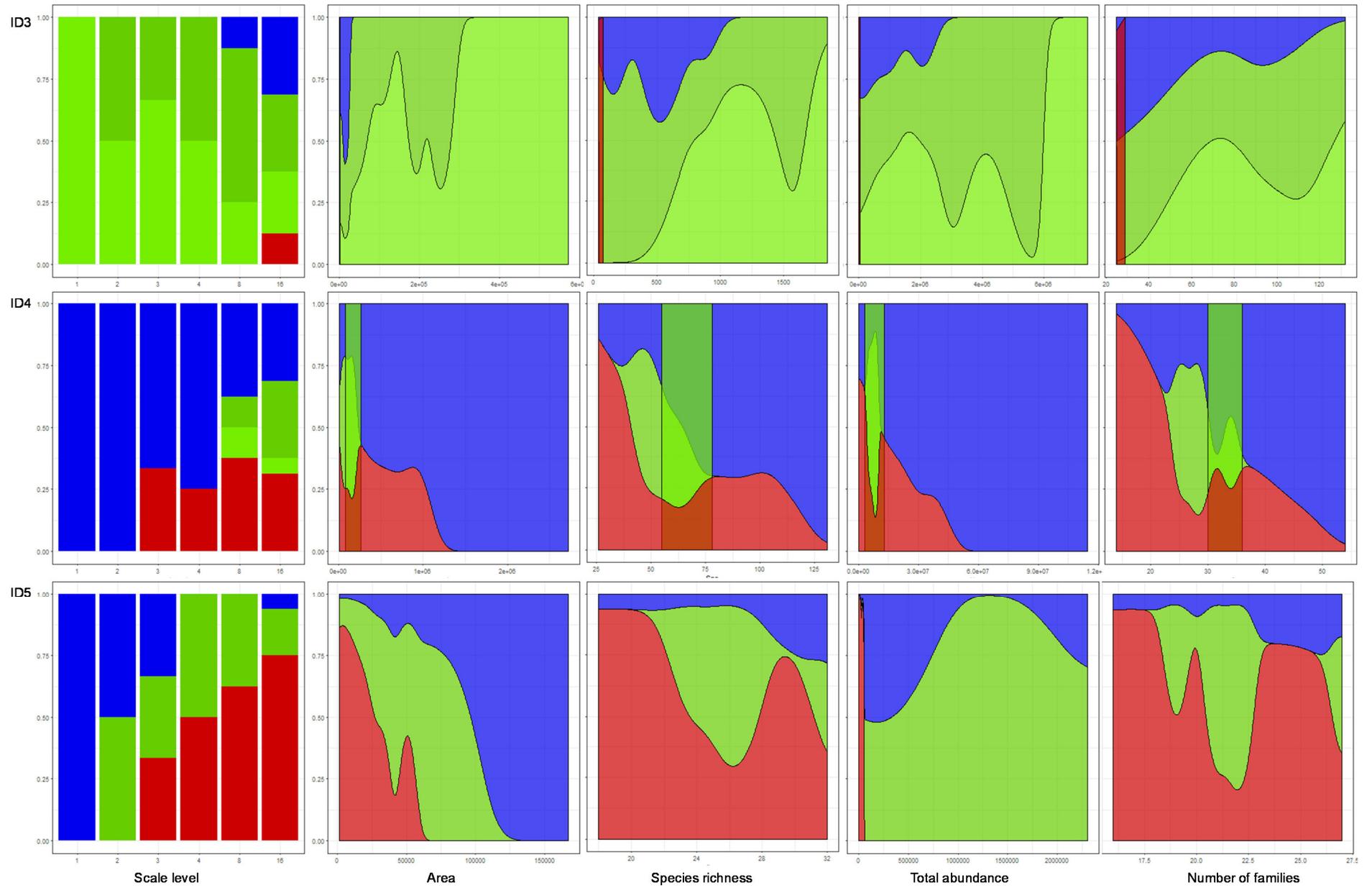
Visual inspection of the deviations between the empirical SADs and each model's predictions supports the abovementioned results. For the communities with consistent support for multimodality across the scale gradient, logseries consistently and severely overestimated the number of singletons (and rare species) across the scale gradient, while 1PLN often underestimated them, and both models either over or underestimated the number of species with intermediate to high abundances. On average, deviations are smaller for 2 or 3PLN at every scale (Figs. 5.3 and VI.2; IDs 2, 3, 10, 12). For the remaining multimodal SADs at the total extent, logseries overestimated the number of rare species again, and the PLN mixtures also exhibit large deviations between the observed number of species and the models' predictions across the distribution and across the scale gradient. For the SADs better fit by 1PLN at the total extent, for ID7, deviations are much smaller on average for 2PLN at every scale, while both logseries and 1PLN underestimate the number of rare species. For

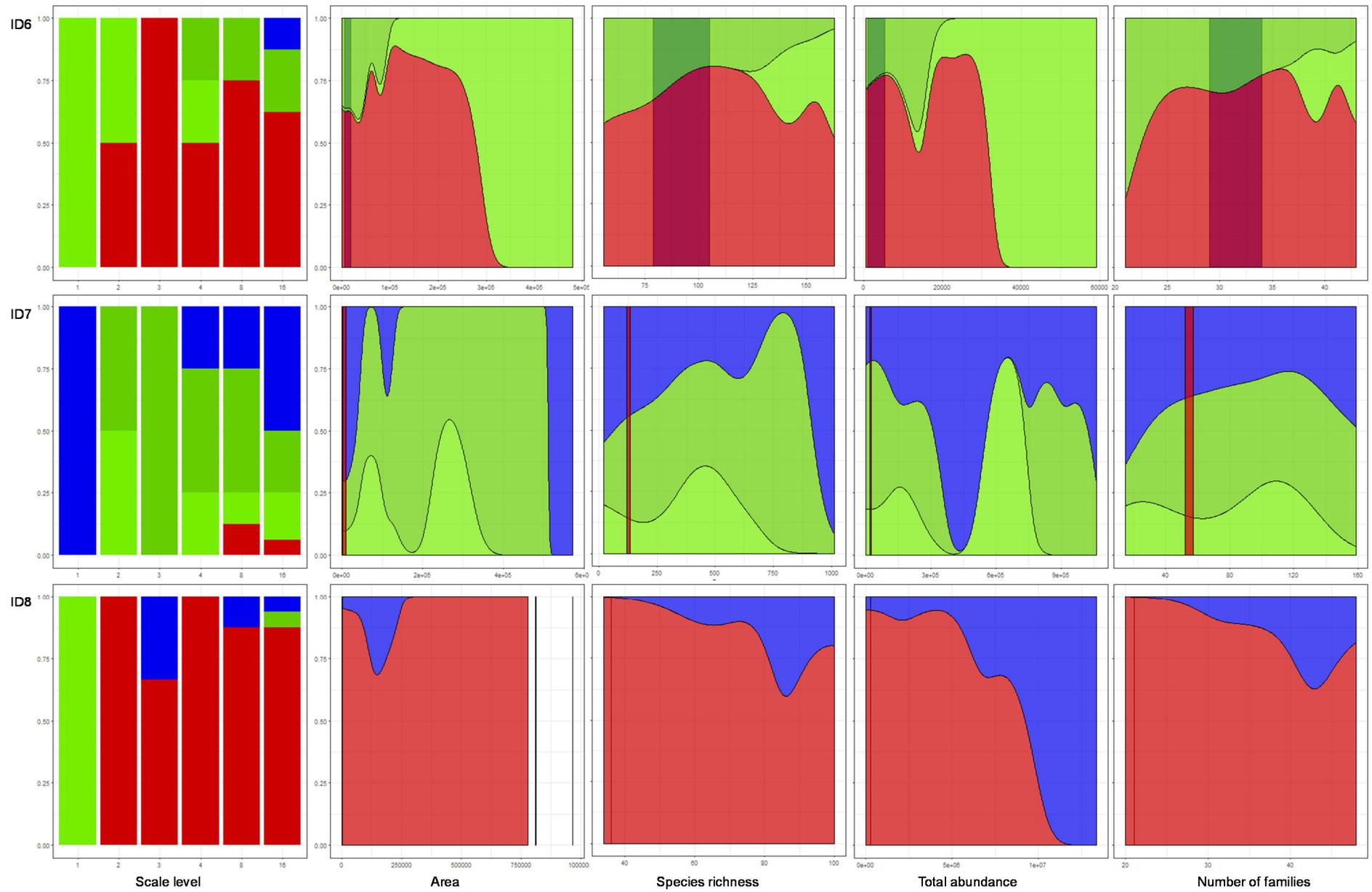
the remaining SADs, there is no clear pattern, but the logseries is systematically unable to accurately predict the rarest and the intermediately abundant species.

Regarding the scaling properties of the other α metrics, overall the expected linear relationship on log-log scale for species richness (S) was found, as well as a similar relationship for total abundance (N) and number of families (Fig. 5.4 a), b) and c)). There is some variability in the linear relationships for Fisher's α with log area, with increasing diversity for some communities, but more shallow relationships for other communities (Fig. 5.4 d)). The Shannon diversity index (H') increased with increasing area, and there was generally a significant linear relationship with log area (positive slopes), although the overall amount of variation explained varied depending on the community (for 6 communities adjusted R^2 values > 0.4 ; Table VI.1; Fig. 5.4 e)). Evenness (J') remained relatively stable across the scale gradient for the majority of communities (Fig. 5.4 f)). There was more variability in S, N and the other α metrics between the sections as area decreased. The exception to this general pattern was ID5, with Fisher's α , Shannon's H' and evenness decreasing as log area increased.

Figure 5.2 Effect of area sampled, species richness, total number of individuals and number of families on the best model selected for the SADs aggregated across all the communities analysed (top row) and for each individual community, identified by the corresponding ID1. 1PLN is represented in blue, 2PLN in darker green and 3PLN in lighter green, and logseries in red (next pages).







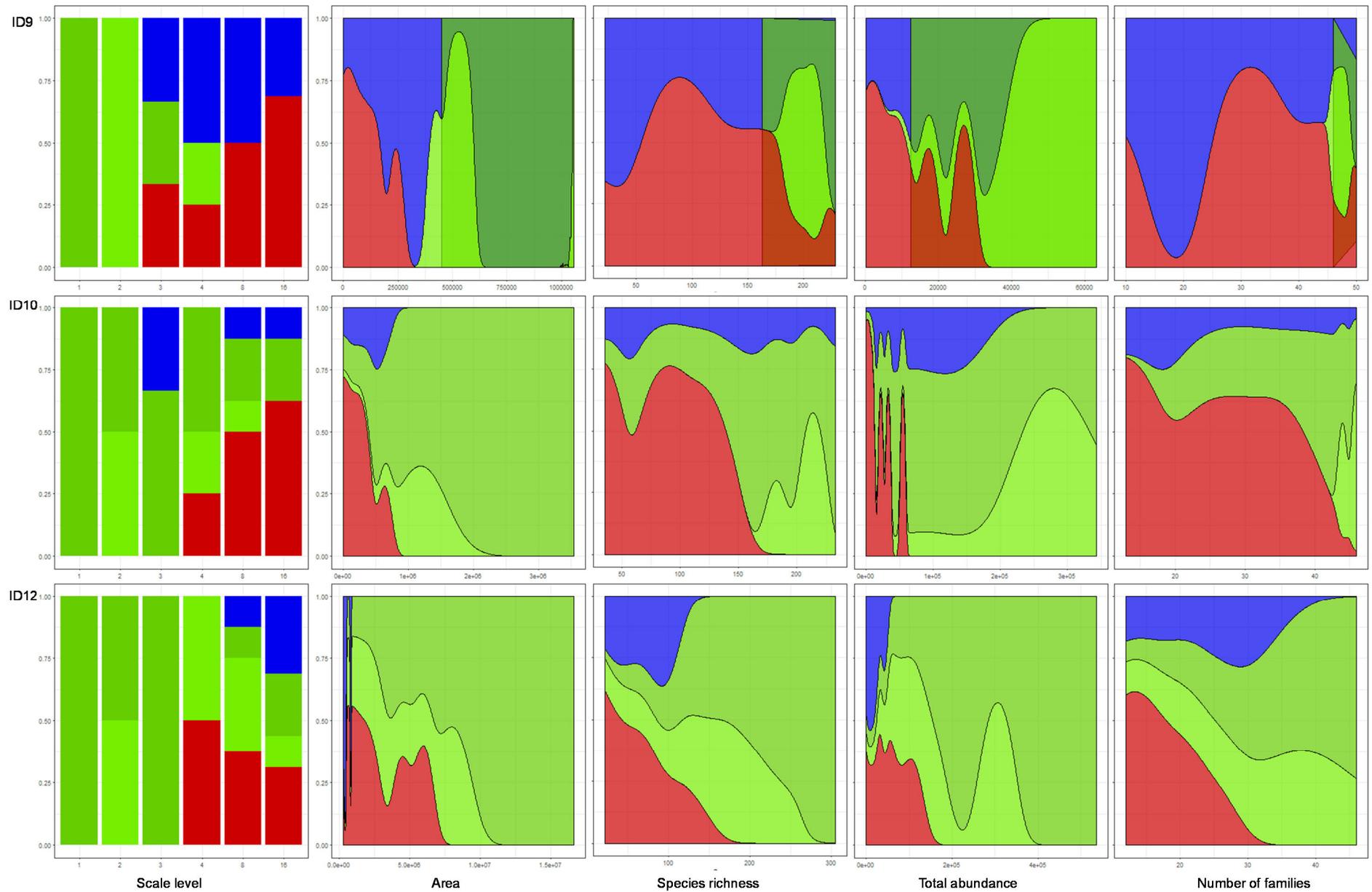
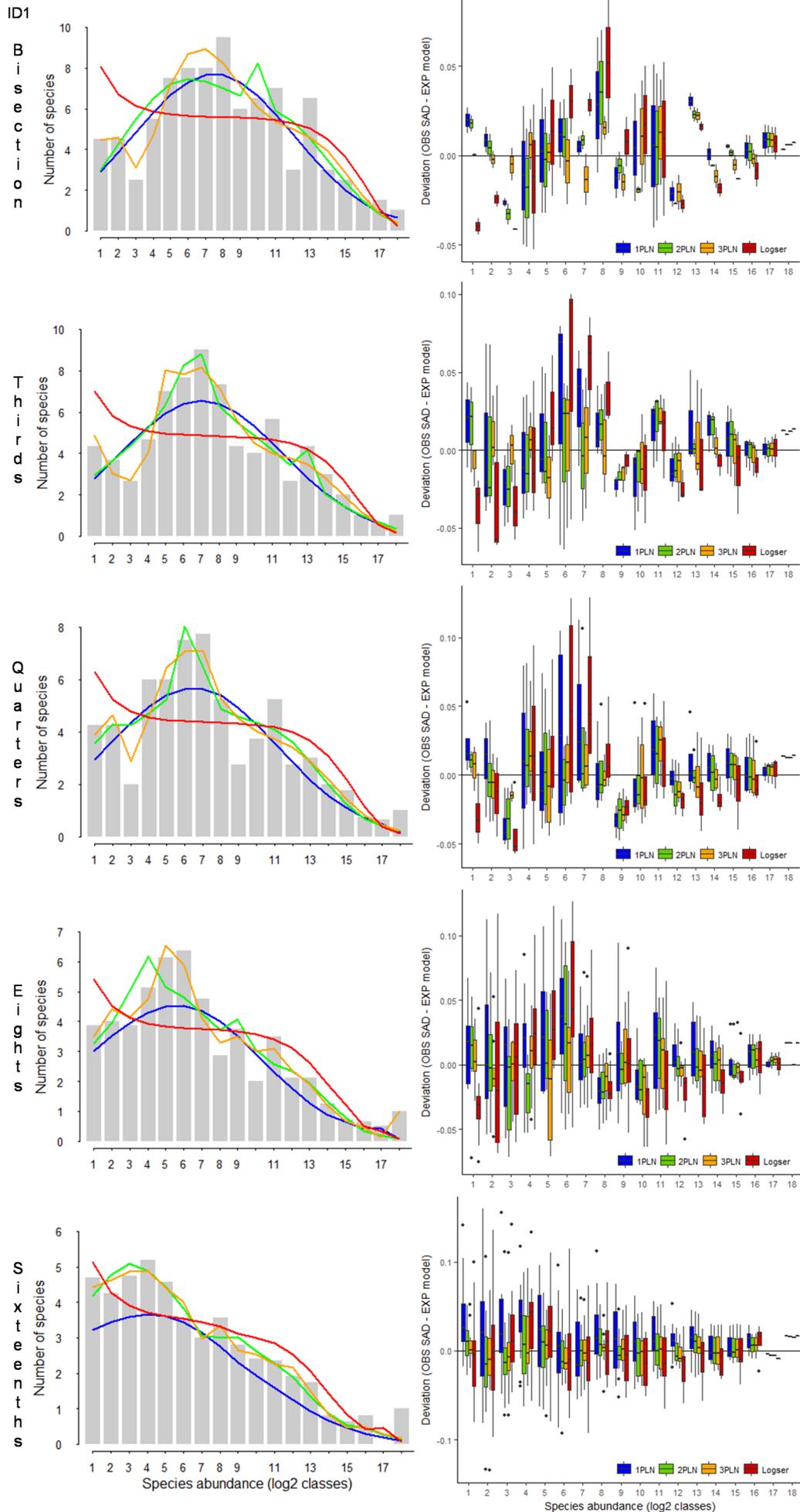


Figure 5.3 Left panel – Comparison of the mean empirical SADs (histograms) and the mean fitted models. Right panel – Deviation between the each empirical SAD and the best fit parameterization of each alternative model; deviations are calculated as the difference between the proportion of species observed and the predicted by each model for each octave of abundance. In both plots, 1PLN is represented in blue, 2PLN in green, 3PLN in orange, and logseries in red. Each community is identified by the corresponding ID (see panels for the remaining communities (IDs 5-12) in Fig. VI.2 (Appendix VI)).

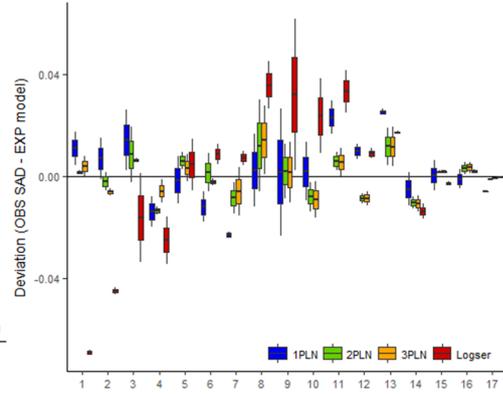
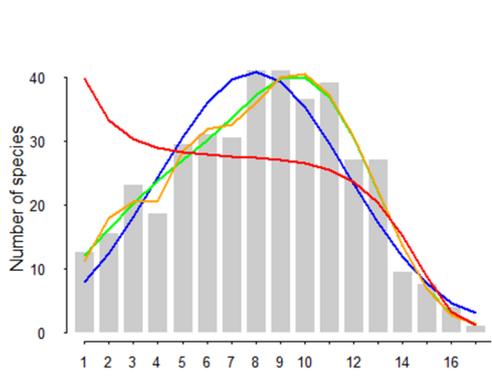
SADs across spatial scales



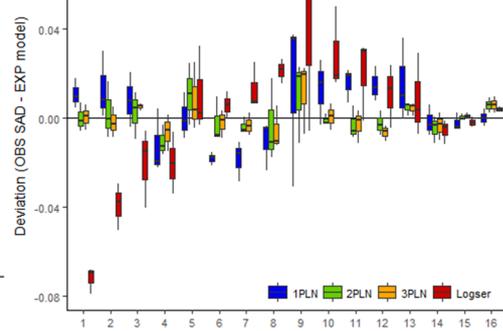
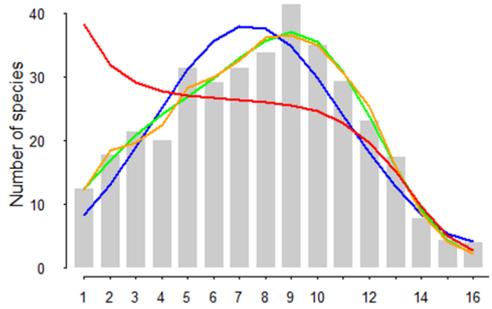
SADs across spatial scales

ID2

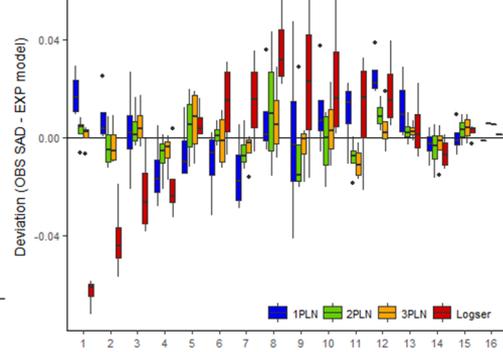
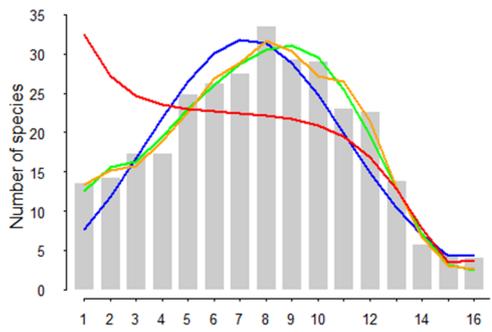
B
i
s
e
c
t
i
o
n



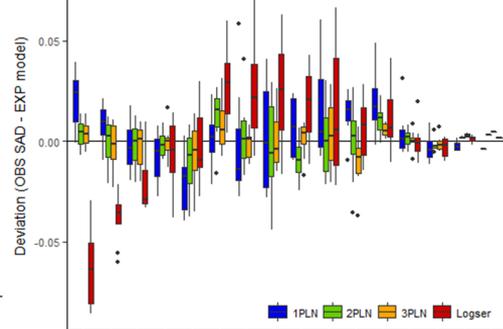
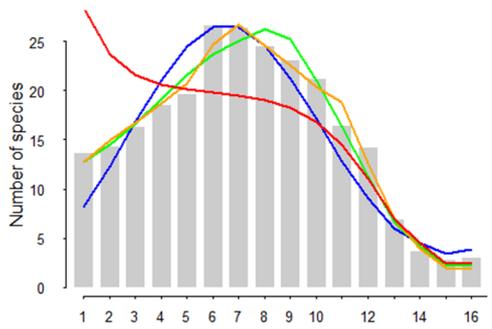
T
h
i
r
d
s



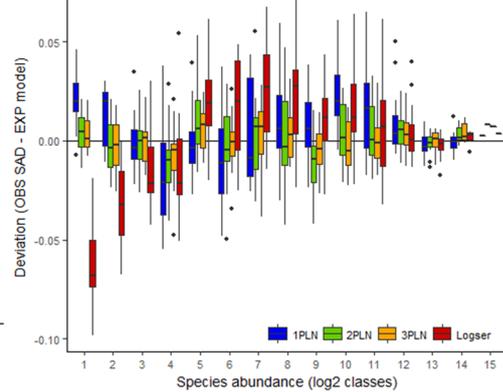
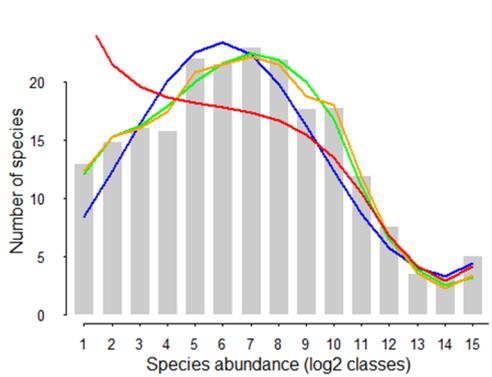
Q
u
a
r
t
e
r
s



F
i
g
h
t
s

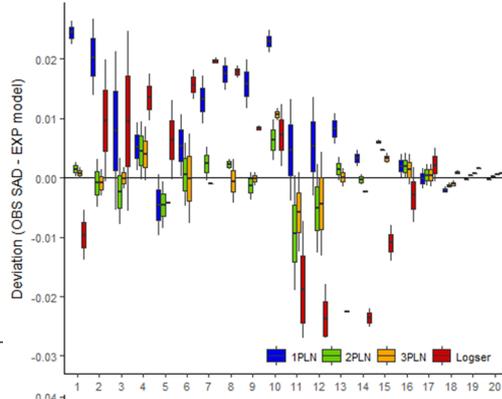
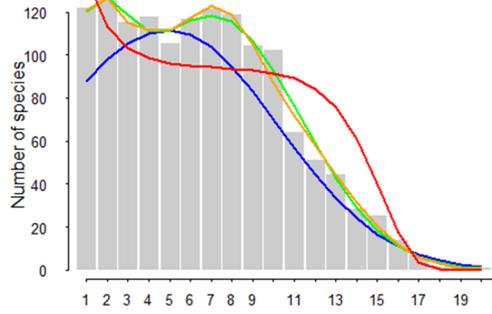


S
i
x
t
e
n
t
h
s

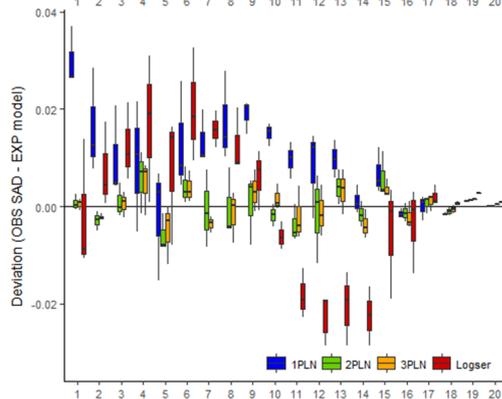
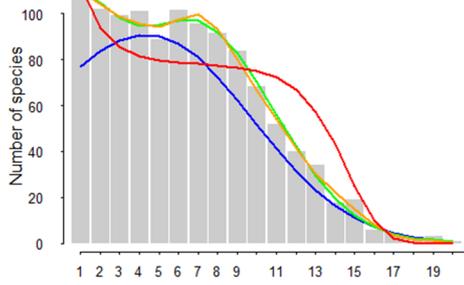


SADs across spatial scales

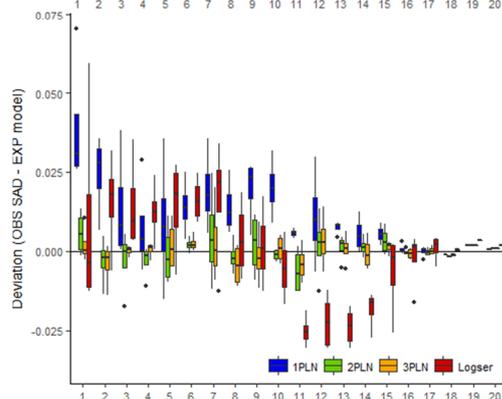
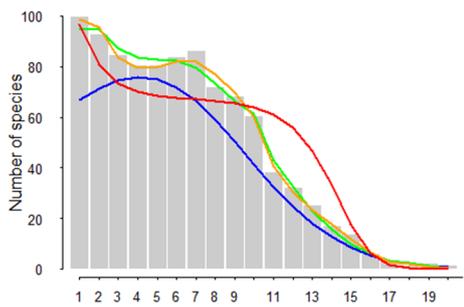
ID3
B
i
s
e
c
t
i
o
n



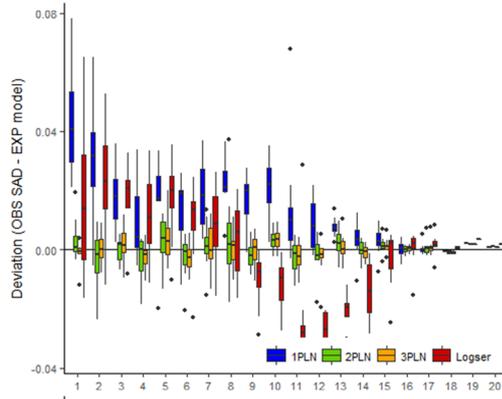
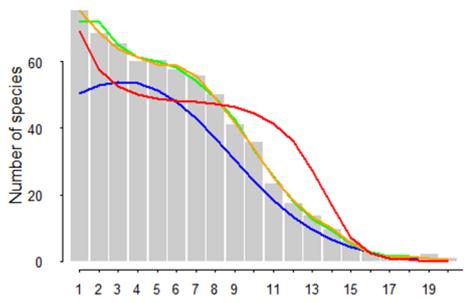
T
h
i
r
d
s



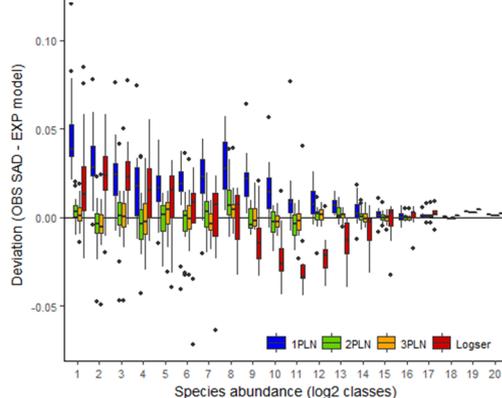
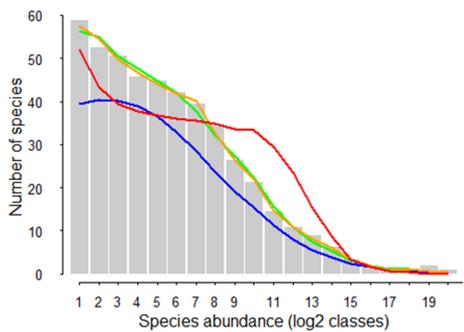
Q
u
a
r
t
e
r
s



F
i
g
h
t
s

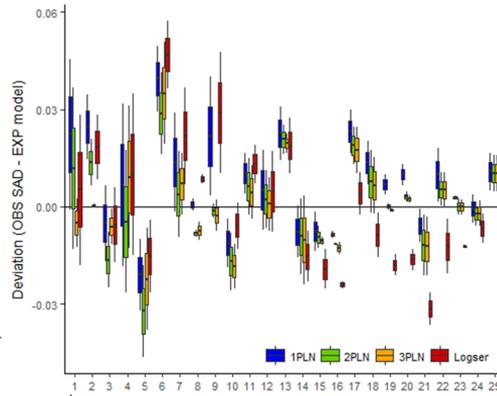
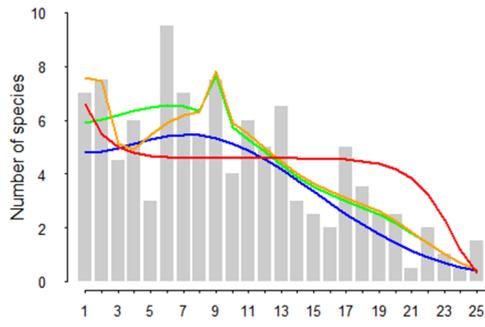


S
i
x
t
e
n
t
h
s

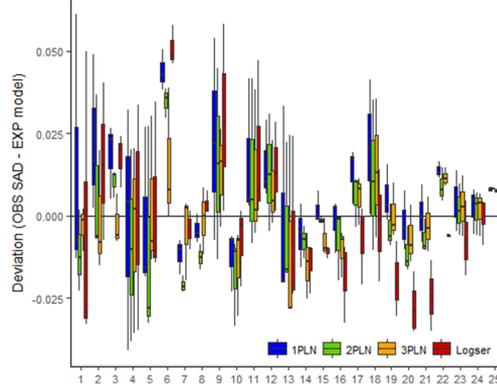
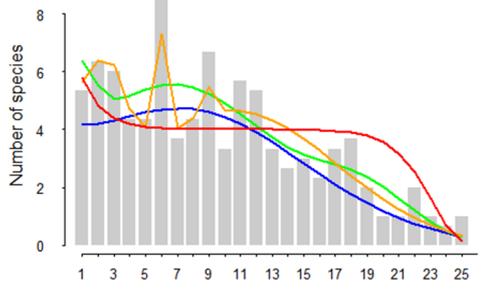


SADs across spatial scales

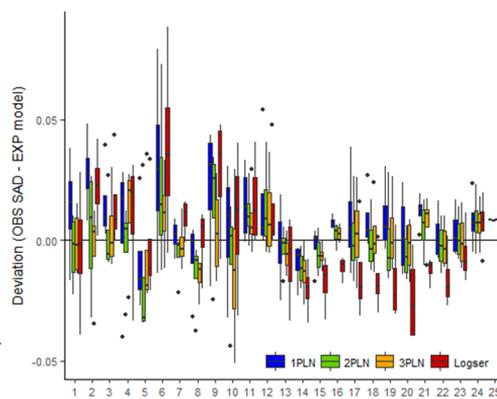
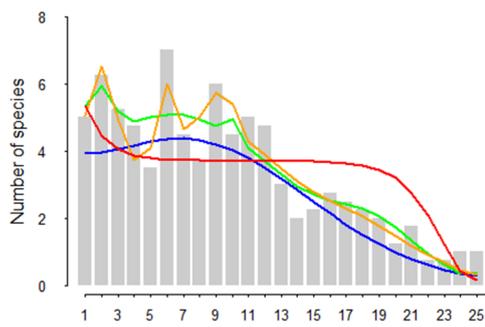
ID4
B
i
s
e
c
t
i
o
n



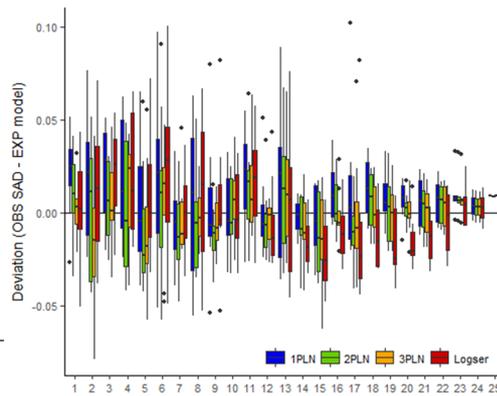
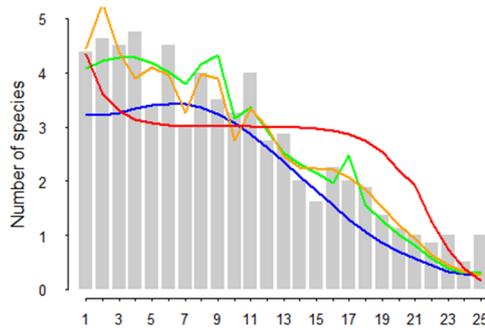
T
h
i
r
d
s



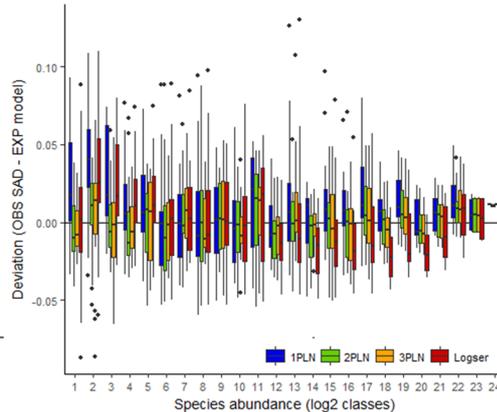
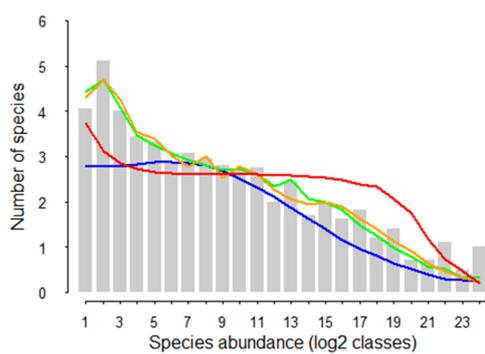
Q
u
a
r
t
e
r
s



E
i
g
h
t
s



S
i
x
t
e
n
t
s



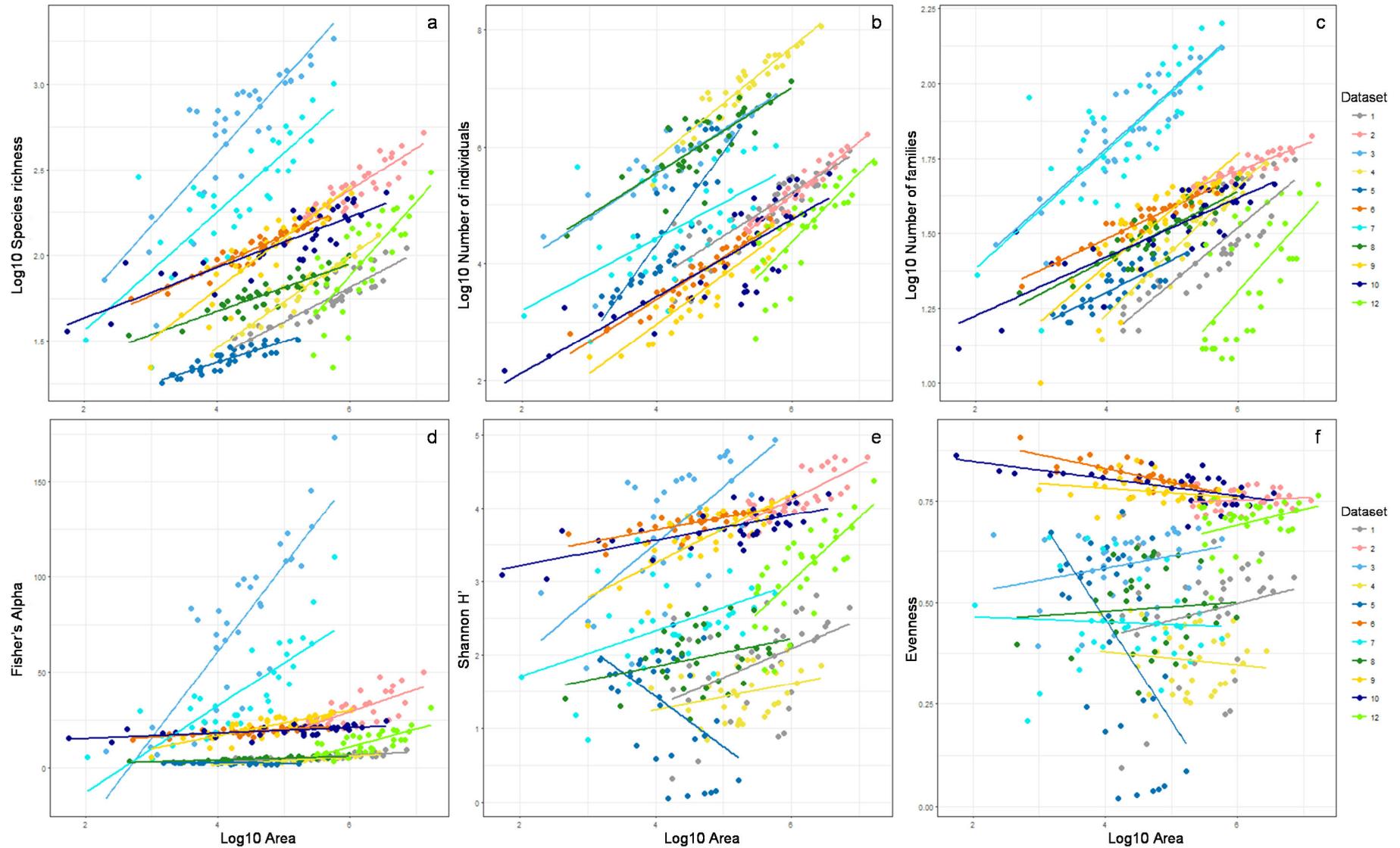


Figure 5.4 Scaling relationships with area for species richness (a), total abundance (b), and number of families (c) on a log-log scale, and for Fisher's α (d), Shannon's Diversity (H') (e), and Evenness (J') (f), on a semi-log scale.

5.4 Discussion

The systematic assessment of SAD shape showed consistent variation in SAD across the scale gradient. Furthermore, it supported the findings in the previous chapter: there is a higher prevalence of multimodal SADs for larger areas and for more taxonomically diverse communities, while logseries never provided an adequate fit for larger and more diverse communities. In addition, this analysis revealed a clear effect of area, species richness and taxonomic diversity in determining SAD shape, and a non-directional pattern for total abundance (for the aggregate SAD results). I compared the performance of different models to describe the SADs of different taxa across a scale gradient spanning several orders of magnitude, providing a comprehensive and robust analysis of how the relative abundance patterns depend on spatial scale and also on taxonomic breadth, without any *a priori* theoretical framework. The results here clearly depart from two macroecological theories predictions for the SAD.

Variability in SAD shape across scales

For the communities selected as multimodal with strong support at the total extent, multimodality is still strongly selected across the scale gradient, even for intense sampling effects of area, species richness and total abundance. This suggests that multimodality is a robust feature of the SADs and is indeed reflecting the structure of the underlying communities, rather than being a sampling (c.f. Barabás *et al.*, 2013) or scaling artefact. The sections created across the scale gradient for these spatially broad and taxonomically diverse datasets (The North American BBS bird survey, Reef Life Survey fish, Ontario Breeding Birds, and the FIA tree inventory), still represent very large spatial extents, and due to the way the scale gradient was established, the spatial relationship between the sections is maintained, thus suggesting that multimodality reflects the structure of the communities at these smaller scales, despite the marked decrease in number of species and total abundance as area sampled decreased. Since in this analysis I only used AIC_c to select the best fit model, it is possible that the prevalence of multimodal models is being overestimated (see results in section 3.3). Nonetheless, there was consistent and strong support for 2 or 3PLN for these communities. Hence, selection of multimodality across the scale gradient for these communities is robust. For these communities, SAD shape was more or less conserved across a wide range of areas sampled; dramatic

shifts in key aspects of community structure are required for the overall SAD shape to change (e.g. Supp & Ernest, 2014).

Clear shifts in SAD shape can provide information about relevant ecological and spatial aspects affecting community structure at different scales. For the non-multimodal SADs at the total extent, but for which multimodality was frequently selected for smaller sections, this might be due to haphazard spatial decomposition of the community when splitting the total extent, and/or because of sampling effects, namely if the SADs become dominated by both rare and very abundant species (Green & Plotkin, 2007). When the more abundant species are very abundant (hence dominating smaller samples, or smaller areas of the community in terms of number of individuals), 2 or 3PLN models are better able to accommodate both the rare and the most abundant species in the distribution, hence being selected despite the increase in number of parameters. Species aggregation patterns can lead to one or a few species becoming extremely (proportionally) abundant for smaller areas. Simultaneously there is also a higher number of rare species in smaller samples, and hence a multimodal model provides a better fit than the (original) 1PLN for the total extent. In contrast, logseries fails to accommodate both rare and very abundant species simultaneously. Differences in species aggregation rates have been suggested to be related to the existence of multiple modes in both theoretical and empirical SADs (Alonso *et al.*, 2008; Dornelas & Connolly, 2008). The results here suggest that multimodality occurring at smaller scales might be due to the spatial aggregation of individuals and species, where “hitting or missing” areas where a species is abundant can lead to the appearance of different modes. Hence, a multimodal model provides a better description for the SAD by accommodating both the rarest and the more abundant species, something that neither the logseries nor a single PLN are able to do. Conspecific aggregation is one of the fundamental features of ecological communities (McGill, 2010b). The results reported here suggest that species spatial aggregation is likely an important driver across scales (from local to truly continental scales) and taxa.

α metrics across scales

The expected relationship of species richness with area emerged for all the communities, and a similar pattern was found for total abundance and number of families as well. There seems to be support for a power-law relationship for both Fisher's α and Shannon's H' with area, while evenness remained largely stable across scales for the communities analysed. The exception was ID5, which might be related to the much smaller number of species in this dataset ($S=32$); thus for this community, species richness might be dominating the diversity patterns. The only theoretical framework I am aware of that explicitly predicts a relationship between the Shannon index and area is the multifractal approach proposed by Borda-de-Água *et al.* (2002). The method of moments used by the authors to derive the mathematical multifractal formalisms leads to power-law relationships between area and the Shannon, Simpson, and Berger-Parker diversity indices. Although the estimated Shannon diversity values did show a linear relationship as a function of log area, there was a lot of variability in the strength of the relationship depending on the community analysed. This suggests that the multifractal approach might provide a good fit for specific taxa, but is not able to reproduce the scaling relationship of Shannon diversity for all the different communities analysed. One possible aspect that might explain this discrepancy is that the relationships are derived for relatively small scales, whereas the spatial gradient explored here spans several orders of magnitude. The multifractal approach also predicts a scale invariant SAD after renormalization, which is not supported by the results here and in the previous chapter. The more stable relationship of evenness with sampled area can be related to community-level properties being relatively conserved. For instance, Supp & Ernest (2014) reported that species-level responses to disturbance were stronger than community-level properties, with species richness, evenness, and the form of the RAD being relatively resilient to disturbance (RAD, or Rank Abundance Distributions are an alternative way of plotting species abundance distributions). Although their study does not explicitly consider scale effects, this is in accordance with the results found here, if a parallel can be drawn between a large decrease in area sampled (and consequently number of species and number of individuals) and disturbance events. Furthermore, the fact that evenness remained relatively unchanged across the scale gradient can again be related to species aggregation properties.

Comparison with neutral and METE theories

This investigation not only assessed the effect of scale on empirical SADs, but also provided contrasting results with two important macroecological theories. Logseries distributions were never selected as best fit for the larger scales (with larger areas and higher number of individuals) and were unable to accurately predict the abundances of both rare species and more abundant species. This clearly deviates from neutral models' predictions of a logseries being the expected SAD for larger scales. Moreover, models with realistic speciation modes and that can produce more flexible metacommunity SADs and reduce the predicted number of singletons (e.g. Rosindell *et al.*, 2010) are still not able to accommodate multimodal SADs. On the other hand, the results here also show a strong departure from several studies reporting METE's success in characterizing the general shape of the SAD (Harte *et al.*, 2008; White *et al.*, 2012; Xiao *et al.*, 2015). White *et al.* (2012) reported that the logseries provided a better fit to several empirical SADs with a wide range of "anchor scales", including two datasets analysed here which have been selected as multimodal with strong support, specifically the North American BBS bird data and the FIA tree data. White *et al.* (2012) also reported that the logseries tended to overestimate richness for the lowest abundance classes, which is in agreement with my results, and the authors suggested that other METE's formulations or neutral models can be used as alternatives, since they predict fewer singletons. Although in this analysis I have not directly analysed either METE's or neutral models, the results here clearly illustrate that the logseries is not able to simultaneously deal with the rare and the abundant species tails. Furthermore, the fact that multimodal models have systematically outperformed the logseries, particularly for larger scales, and following the regression results performed in chapter 4 for over 100 empirical communities, it is unlikely that the logseries is an adequate descriptor of SADs across spatial scales.

These results do not invalidate the logseries as a "realistic functional form for SAD" as produced by METE (Pueyo *et al.*, 2007; Harte *et al.*, 2008) or neutral models (Hubbell, 2001; Rosindell *et al.*, 2010), nor that it is not a useful model in certain contexts. Logseries has indeed been selected as best model for several communities (in both chapters 4 and 5). What these results illustrate is that logseries is not adequate as the single SAD distribution, as METE suggests, and furthermore that it is more likely to describe SADs at smaller scales and for less taxonomically diverse communities, contrary to a logseries describing the metacommunity level in neutral models (Hubbell, 2001; Pueyo *et al.*, 2007; Rosindell *et al.*, 2010). Furthermore, in METE's framework, because derivations for other macroecological patterns depend on the SAD's formulation as a logseries (Harte *et al.*, 2008), this highlights the need to incorporate and test other SAD distributions to ensure those derivations are robust. On the other hand, the success of METE's predictions depends on selecting appropriate

state variables (Harte *et al.*, 2008; Harte & Newman, 2014), while small modifications can lead to different results for the SAD (Pueyo *et al.*, 2007). Furthermore, different assumptions for the configuration, the imposition of constraints, and the scale on which MaxEnt models are formulated can lead to different predictions for species spatial distributions (Haegeman & Etienne, 2010).

A recent model extending the neutral theory by incorporating size variation and growth dynamics (the size-structured neutral theory model (SSNT)) still assumes a logseries SAD (O'Dwyer *et al.*, 2009). A comparison of the ability of different model formulations for both METE and SSNT showed a better performance for the SSNT models (O'Dwyer *et al.*, 2009; Xiao *et al.*, 2016), with the authors arguing that METE's constraints are not fully capturing relevant biological processes that influence community structure. Nevertheless, neither neutral nor METE models account for multiple modes in SADs, not to their higher prevalence at larger scales and for more diverse communities. These results also call for a closer look at the notion of the "feasible set" (Haegeman & Loreau, 2008; Locey & White, 2013), specifically as the total abundance N across the scale gradient did not exhibit any directional effect on the SAD shape (for the aggregated results). Hence, the results in this chapter suggest that the combination of S and N is not sufficient to predict SAD shape across spatial scales (White *et al.*, 2012; Locey & White, 2013; Xiao *et al.*, 2015). Nonetheless, both area and species richness showed a strong influence on SAD shape, although the two variables are also strongly correlated. One of the advantages of using the METE approach is being able to interpret the deviations from the expected distributions solely constrained by richness and abundance as evidence that other ecological features must be important in structuring the communities analysed (Harte *et al.*, 2008; White *et al.*, 2012; Xiao *et al.*, 2016).

Moreover, the results here also depart from a purely idiosyncratic community structure; employing the MaxEnt approach under a Bayesian framework, and considering that species differ in all aspects (opposed to "equivalent species" in neutral theory), Pueyo *et al.* (2007) also derived a logseries SAD. However, the authors showed that small modifications can yield SADs that depart from the logseries, originating "bounded power law" or "bounded skewed lognormal-like" distributions (see also Pueyo (2006)). On the other hand, by analysing a smaller plot within one of the datasets in their 2008 paper, Harte *et al.* suggested that large-scale heterogeneity could be used as an additional constraint, noting that by examining a more homogeneous subplot the discrepancy in the estimated SAR slope decreased. Here (and in the previous chapter) I have clearly demonstrated that ecological heterogeneity, represented by larger scales, higher taxonomic diversity or species richness, is linked

to the existence of multiple modes in empirical SADs across taxa, and inversely linked to logseries SADs. Hence, modifying or adding other relevant constraints to METE might provide a fruitful avenue to test if the METE approach is able to incorporate the variability in SAD shape occurring in empirical data. The fundamental rationale behind the state-variable approach to ecology is to use *a priori* knowledge of the system and maximize entropy to find the most likely distribution. The results here clearly indicate that more information is required as input if the METE methodology is to accommodate the variability in SAD shape found in empirical communities, and crucially to reproduce different SADs at different scales, including multimodal SADs.

One potential source for the disparate findings here and the two theoretical predictions might arise from the scale gradient framework implemented, which spanned several orders of magnitude, and included very large areas, even for the smallest scale levels. Hence, it is possible that discrepancies found between my analyses and the theoretical predictions might be at least partially attributable to differences in the spatial scales investigated. For instance, the original METE formulation was designed for downscaling, and the original comparisons were made for relatively small plots (ranging from 64 m² to 50 ha) (Harte *et al.*, 2008). On the other hand, METE derivation was based on regular shaped plots that are formed by successive bisections of the “anchor scale”. More recent METE developments focused on upscaling SARs, but still assumed a logseries SAD (Harte *et al.*, 2009). It has been suggested that both Neutral theory and METE might be more adequate for smaller scales (smaller areas/fewer individuals) (McGill, 2010b). The results here support this suggestion, since there was a clear prevalence of the logseries distribution occurring at smaller spatial and taxonomic scales. Several of the studies mentioned here that derive scaling relationships do not use explicit “measures” of scale – a metacommunity, or a regional species pool, might constitute very different spatial extents depending on the particular community or taxa. The analysis here did not focus on any particular scale, nor did I intend to provide any operational definition for those scales. Nonetheless, these results highlight the importance of incorporating explicit scales, and have additionally expanded the usual scale ranges used to develop such theories.

Conclusions

Spatial scale emerged as a major driver of differences in SAD shape. The systematic analysis of several SADs at different spatial scales and for different taxa allows us to make stronger inferences about the commonness and rarity of species across scales. The results in this chapter clearly show that neither neutral nor METE formulations are able to accommodate the variability in SADs shape across spatial scales. The interplay of SAD shape at different scales can highlight important mechanisms acting on the communities, namely both inter- and intraspecific spatial patterns that lead to different SAD shape as spatial scale changes. A critical development for (current) macroecological theories is to predict or accommodate multimodal SADs, and crucially to incorporate the effect of spatial scale and ecological heterogeneity in determining SAD shape.

6. Multiscale spatial patterns of β diversity

Note: The work and results presented in Chapter 6 will be published in the form of a peer-reviewed article: **Multiscale spatial patterns of beta diversity and its components** (*in review in Ecology Letters*).

6.1 Introduction

β diversity quantifies the variation of species composition between assemblages or sites in a landscape (Whittaker, 1960). It is a fundamental component of biodiversity, with implications for community ecology, macroecology and conservation (Whittaker, 1960; Anderson *et al.*, 2011; Socolar *et al.*, 2016). However, as yet β diversity scaling patterns across space are poorly understood. In this chapter, I tested if β diversity shows systematic variation with scale.

Factors such as dispersal and niche limitations, along with environmental heterogeneity and species aggregation can affect β diversity patterns (Whittaker, 1960; Nekola & White, 1999; Gaston *et al.*, 2007; Morlon *et al.*, 2008; Barton *et al.*, 2013). As additional habitat types and different environmental features are included for larger geographical areas, β diversity patterns are expected to be scale dependent (Koleff *et al.*, 2003; Tuomisto, 2010b; Barton *et al.*, 2013). A lot of research has been dedicated to the scaling properties of species richness (Rosenzweig, 1995; Harte *et al.*, 2009; Storch *et al.*, 2012), but less attention has been devoted to the scaling of β diversity, with a lack of theoretical predictions about the form of β diversity scaling patterns (Koleff *et al.*, 2003; Gaston *et al.*, 2007; Barton *et al.*, 2013). Furthermore, little is known about how its two components, turnover and nestedness, behave across different scales. As turnover and nestedness are generated by fundamentally different processes, quantifying their relative contribution across spatial scales can provide insights into the mechanisms underlying β diversity (Baselga, 2010; Svenning *et al.*, 2011).

The measurement of β diversity is affected by the spatial scale of observation in terms of grain and extent (Wiens, 1989; Nekola & White, 1999; Mac Nally *et al.*, 2004; Qian, 2009; Keil *et al.*, 2012; Steinbauer *et al.*, 2012; Barton *et al.*, 2013; Nekola & McGill, 2014). For small grain sizes compared to the overall extent of the study, even close sampling units might be very dissimilar in their species composition, due to stochastic sampling effects and high variability in species occupancy patterns. As grain size increases, mean environmental variability decreases as a result of spatial averaging, and the probability of detecting more rare species increases (Wiens, 1989; Levin, 1992; Gaston *et al.*, 2007; Keil *et al.*, 2012; Barton *et al.*, 2013). Hence, a decrease in dissimilarity as grain size (area sampled) increases is expected. But what is the functional form of this relationship? And are the patterns system or taxon specific (Barton *et al.*, 2013)? This analysis provides the first attempt building β diversity scaling curves – akin to the triphasic Species-Area Relationship (Williams, 1943; Rosenzweig, 1995; Storch *et al.*, 2012), and by using empirical data from different communities, I provide the first assessment of their generality or idiosyncrasy.

One of the most widely used descriptions of β diversity is the distance decay of similarity (DDS) (Nekola & White, 1999; Morlon *et al.*, 2008). DDS arises from the decrease in environmental similarity with geographic distance and/or from differences in the dispersal abilities of the organisms (Nekola & White, 1999; Morlon *et al.*, 2008). Again as grain size increases, dissimilarity is expected to decrease, but previous studies have reported conflicting results. For instance, lower DDS rates for larger grains were reported by Nekola & White (1999) and Keil *et al.* (2012). In contrast, Morlon *et al.* (2008) found that grain size only affected the rate of decay at the smallest grain size analysed, while no consistent trend was found by both Steinbauer *et al.* (2012) and Zacaï *et al.* (2016). This suggests that the influence of grain size on DDS rates cannot be easily predicted, and might be context and/or taxa dependent. However, these studies have all employed different approaches, varying total extent and/or grain, using different taxa and analysing the patterns at different spatial scales (continents to plots). Hence, a systematic multiscale approach across multiple taxa to analyse the distance decay of similarity can help disentangle scale effects and ecological patterns.

A few studies have analysed turnover and nestedness patterns at large spatial extents for specific taxa, with their relative contributions being apparently contingent on the scale levels investigated (Svenning *et al.*, 2011; Wen *et al.*, 2016). On the other hand, and similarly to Baselga (2010), these studies analysed dissimilarity patterns at fixed spatial scales and along latitudinal/longitudinal gradients, while disentangling the effects of different environmental drivers. Thus, there is still no

general investigation of turnover and nestedness relative contributions independent of latitudinal, longitudinal or environmental gradients. Turnover can be expected to be lower between larger sampled areas and for more vagile organisms. It is plausible that the nestedness component is less relevant between smaller sampled areas, where turnover may be the dominant driver of β diversity. On the other hand, nestedness could also represent a smaller portion of β diversity in scenarios with high dispersal rates. Thus, it is not clear how sampled area could affect the two components relative importance (Si *et al.*, 2015), with possible interactions resulting from other mechanisms, e.g. metacommunity dynamics (Leibold *et al.*, 2004; Tonkin *et al.*, 2016; Gianuca *et al.*, 2017).

In this chapter, I systematically explored β diversity scaling patterns and provided the first empirical assessment of β diversity scaling curves for communities from different taxa. Additionally, I consistently investigated the behaviour of DDS across a scale gradient, by testing the hypothesis that DDS rates decrease as grain size increases. Finally, I tested whether the turnover component decreases with area sampled, along with total β diversity, while exploring the behaviour of the nestedness-resultant component.

6.2 Methods

In this study, I analyse β diversity patterns across a scale gradient established using the sampled areas from empirical datasets, rather than imposing static grids with arbitrary dimensions. Scale is broadly defined as area sampled, and for the purposes of this analysis the largest scale in the scale gradient was determined by the total extent sampled for each individual dataset, and the lowest level defined by its sampling grain (Tuomisto, 2010a). Finally, γ diversity was assumed constant for each community, corresponding to the total number of species sampled in the total extent, and because my goal was not to estimate latitudinal gradients of β diversity, I did not employ a null model approach (Tuomisto, 2010a; Socolar *et al.*, 2016; Ulrich *et al.*, 2016).

Using the scale gradient and the same datasets analysed in the investigation of SADs across different scales described in Chapter 5 (see sections 5.2.1 and 5.2.2; Table 5.1), an analysis of how β diversity and its components, turnover and nestedness, behave across spatial scales was carried out. A random point from the circle was selected to split the circle into halves, thirds, quarters, eighths and sixteenths, using the initial random point from the bisection as reference (Fig. 6.1). I performed sample rarefaction to obtain an equal number of samples in each subsection and then pooled species abundances across the retained samples within each section, before calculating β diversity metrics between each pair of sections (Baselga, 2010; Tuomisto, 2010b). Each study was randomly split ten times, to assess the generality of the results focusing on effects of scale independently of latitudinal or longitudinal effects. Finally, I also calculated β diversity between all the individual samples, representing the lowest level of the scale gradient.

The areas sampled (i.e. grain sizes) were calculated using convex hull polygons encompassing the sampling locations within each section, using package *rgeos* (Bivand & Rundel, 2016) (Fig. 6.1). At each level, sampled area is estimated as the minimum of the convex hull areas for each section – the estimated metrics are representative of the smallest area sampled. For the lowest level of the scale gradient, the grain size of individual samples was considered, assuming the grain size from a similar study if the exact information was not available (Table 5.1). Since the range of grains analysed covers several orders of magnitude, the results are not contingent on the exact grain size at the smallest scale. The geographic distances were calculated in km as the distance between the centroids of each

section, and between all the sampling locations for the lowest scaling level, using package `sp` (Pebesma & Bivand, 2005) (Fig. 6.1).

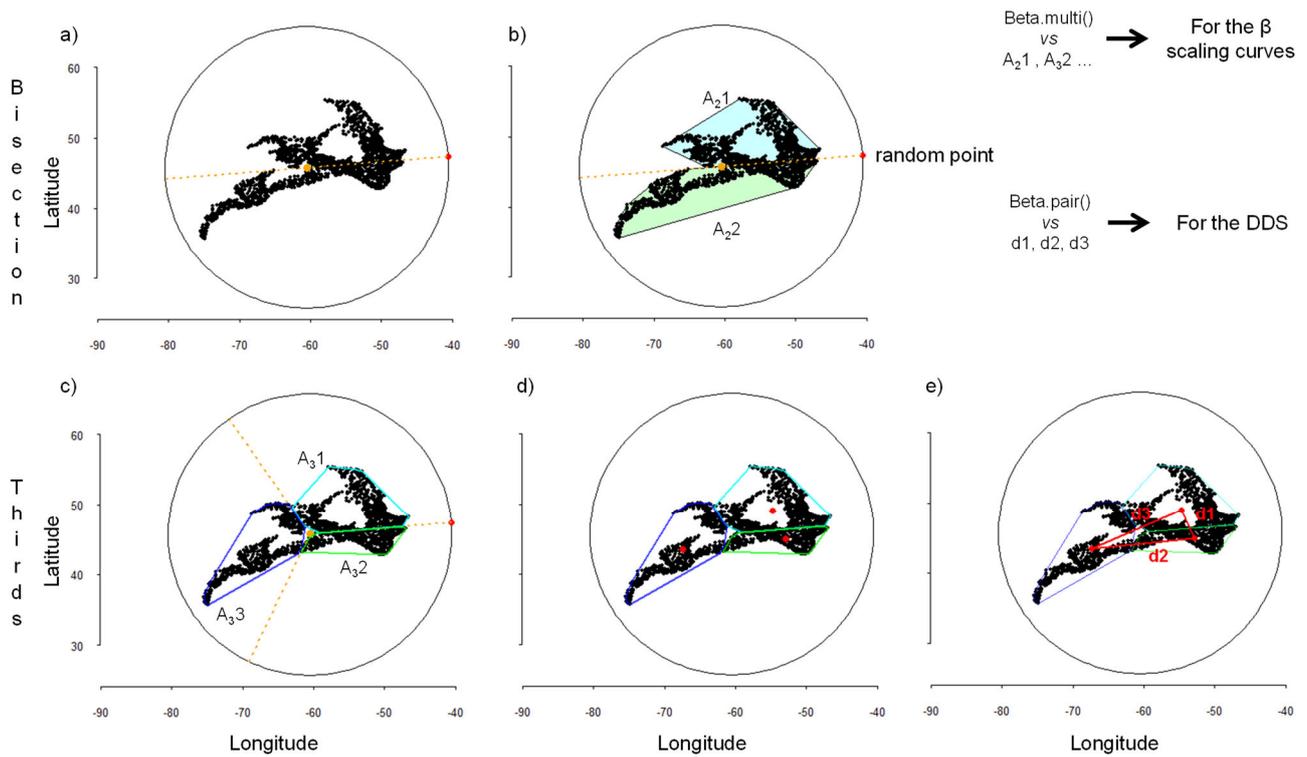


Figure 6.1 Schematic representation of the scale gradient, showing an example of how the encompassing circle was drawn and a random point was selected to establish bisections (a and b) and thirds (c – e). For the β diversity scaling curves, the smallest area at each level was used (b and c); for the DDS analyses the distances between the centroids of each section were calculated (d and e).

6.2.1 β diversity scaling curves

To quantify community dissimilarity I used the β diversity additive partition framework proposed by Baselga (2010), where the Sørensen index (Sørensen, 1948) represents total β diversity, accounting for all aspects of compositional variation, the Simpson index (Simpson, 1943) represents turnover (species replacement independent of species richness gradients), and their difference represents a measurement of the nestedness-resultant component:

$$\beta_{\text{Sørensen}} = \beta_{\text{Simpson}} + \beta_{\text{Nestedness-resultant}}$$

To build the β diversity scaling curves for each community I assessed how β diversity varied with sampled area (non-directional β diversity). I used the multiple-site dissimilarity function `beta.multi()` in the package `betapart` (Baselga & Orme, 2012) to calculate the total β diversity, turnover and nestedness-resultant components for each level of the scale gradient. I then plotted the estimated metrics *vs* sampled area in a log10 scale to build the β diversity scaling curves. Finally, I used generalized nonlinear least squares to fit a power law model to each β metric as a function of area, using function `gnls()` from the `nlme` package (Pinheiro *et al.*, 2016).

6.2.2 Distance Decay of Similarity

I used the function `beta.pair()` in `betapart` (Baselga & Orme, 2012) to obtain pairwise dissimilarities between all the sections at each scaling level and between all the sampling locations for the lowest level of the scale gradient (directional β diversity). This yielded three dissimilarity matrices for each level (one for each β metric) that can then be analysed with the corresponding geographic distances matrix. Since the FIA data represented over 188 million pairwise comparisons at the grain level, 10% of the pairs were randomly sampled for the DDS analysis. I follow the same notation as Baselga (2010), using upper-case letters to indicate the multiple-site measurements (β_{SOR} , β_{SIM} and β_{NES}), and lower-case letters for the pairwise comparisons (β_{sor} , β_{sim} and β_{nes}).

Because the distance matrices are calculated using non-independent sampling locations, Mantel tests were used to assess the significance of the Pearson correlations between the dissimilarity and

geographic distances matrices for the total β diversity and its components, and for each level of the scale gradient, using package *vegan* (Oksanen *et al.*, 2016) with 1000 permutations. The distance decay of similarity was quantified by fitting linear regression models to the total β diversity and its components *vs* geographic distance, at each scaling level. The slope of the regression represents the rate of distance decay, and the more constrained the dispersal of the organisms, the steeper the slopes of the relationships (Nekola & White, 1999). To assess if the intercepts and slopes of the linear models significantly differed between the levels in the scale gradient, I estimated the frequency distribution of the parameters by bootstrapping 1000 slopes and intercepts, using the package *boot* (Canty & Ripley, 2015). For the lowest level of sampling only 100 permutations were used due to the very high number of calculations to perform. For each β metric I evaluated if the slopes were steeper for lower scaling levels, testing if slopes for grain $> 1/16$ and $1/16 > 1/8$; 100 bootstrap values were randomly sampled from the $1/16$ level distributions for comparisons with grain. The same procedure was performed for the intercepts, again testing if coefficients were higher for lower scaling levels. Finally, using the same bootstrap distributions I assessed if coefficients for turnover (β_{sim}) were higher than for the nestedness component (β_{nes}) within each scaling level. These procedures were not performed for the bisection level because this level yields a single comparison, nor for the $1/3$ and $1/4$ levels due to the small number of comparisons available to accurately assess the strength of the linear relationships and generate the bootstrap distributions.

6.3 Results

β diversity scaling curves

All the datasets analysed showed strikingly similar β diversity scaling curves, with β_{SOR} decreasing with increasing sampled area according to a power law (area on a log scale) (Fig. 6.2). As these patterns were very consistent across the ten random splitting trials performed for each dataset, I report the results for a single split. To illustrate that the patterns are not contingent on using estimates from a single trial, I compared the median β_{SOR} values across all the trials (excluding the last one) with the values used in the analysis (Fig. VII.1). The lowest level of the scale gradient invariably exhibited very high β_{SOR} dissimilarity between the samples for all the communities (~ 1), which then decreased with increasing area, with some communities exhibiting more contracted curves along the y-axis; i.e. for the larger areas sampled β_{SOR} was close ~ 0.2 (e.g. the ICES North Sea International Bottom Trawl Survey, the Maritimes Breeding Bird Atlas, and the ICES Irish Ground Fish Survey – IDs 4, 6 and 8, respectively), whereas for others it was $\sim 0.4-0.6$ (e.g. the BBS bird data, the RLS Invertebrates, and the FIA tree inventory – IDs 2, 7 and 12, respectively). β_{SIM} exhibited a similar decreasing pattern and represented the largest portion of β diversity across all the communities and across the scale gradient. β_{NES} seemed to be relatively insensitive to changes in area sampled, and always had a much smaller contribution to total dissimilarity than turnover, although it seemed to increase slightly for some communities as area increased (Fig. 6.2; IDs 1, 2, 3, 10 and 11). Model fitting showed significant power law decrease with \log_{10} area for both β_{SOR} and β_{SIM} , and overall non-significant relationships for β_{NES} (Table 6.1). Moreover, the coefficients estimated fell in a relatively narrow parameter space – in many situations, the coefficients were not statistically different between the communities.

Distance decay of similarity

The Mantel tests indicated an overall positive correlation between both total β diversity (β_{SOR}) and turnover (β_{SIM}) matrices with geographic distance across all the datasets for the grain and the 1/16 levels. On the other hand, the nestedness-resultant component (β_{NES}) was never correlated with geographic distance, even at the grain level (always non-significant negative value). I used a single random trial for the DDS analysis since there was relatively small variability in estimating the linear

model coefficients for the lowest levels of the scale gradient among trials. There was more variability for the higher levels for some communities, but for these levels (1/3 and 1/4) I would not be able to estimate reliable bootstrap distributions, so I restricted the comparisons to the lowest levels of the scale gradient (see Fig. VII.2 for the estimated slopes for all the trials for each β metric).

As expected, total β diversity β_{sor} increased with geographic distance (Fig. 6.3). For all the communities, at the grain level there is a significant positive relationship for both β_{sor} and β_{sim} with geographic distance, and a significant negative relationship for β_{nes} . In other words, the total β diversity and the turnover component increased with the distance between the samples, while the nestedness-resultant component decreased (Fig. 6.3 and Table VII.1). Also as predicted, as the scale level increased, β diversity between the sections decreased. A similar pattern to that of grain occurred for β_{sor} and β_{sim} at higher scale levels (as sampled area increased), while β_{nes} did not show a consistent trend, but always exhibited much shallower slopes with geographic distance than β_{sor} or β_{sim} (Fig. 6.3), with significantly negative slopes occurring occasionally for the 1/16 level as well (Table VII.1). For all the communities, DDS slopes were not significantly different between the scale levels compared (from the bootstrapped coefficients comparing grain vs 1/16 and 1/16 vs 1/8), while the intercepts were always significantly higher comparing grain vs 1/16 levels, and occasionally also for 1/16 vs 1/8 for β_{sor} , with a very similar pattern for turnover (Table VII.2). For the nestedness component the slopes were again not significantly different and there were fewer situations where the scale levels had significantly different intercepts. Similarly to the scaling curves, turnover accounted for the largest portion of β diversity across all the communities and the scale levels. Pairwise comparisons between the bootstrapped slopes and intercepts across the different scale levels showed that both coefficients were consistently higher for turnover at the grain, 1/16 and 1/8 levels.

In summary, for both the scaling curves and the DDS analyses, turnover accounted for the largest portion of β diversity across scales, showing similar scaling properties to the total β diversity, while the nestedness component was less sensitive to large changes in spatial scale/area sampled. Furthermore, the rates of decay of similarity (β_{sor}) and of turnover (β_{sim}) were consistent across large increases in area sampled, while the absolute dissimilarity values for both metrics (the intercepts of the DDS relationships) were affected by increases in area. Finally, the nestedness component was negatively affected by geographic distance at the lowest levels only.

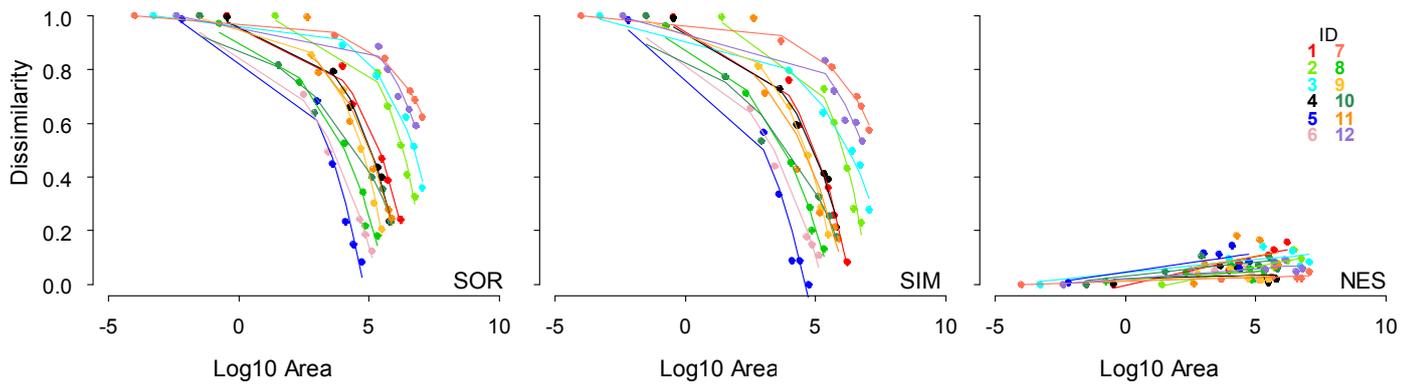
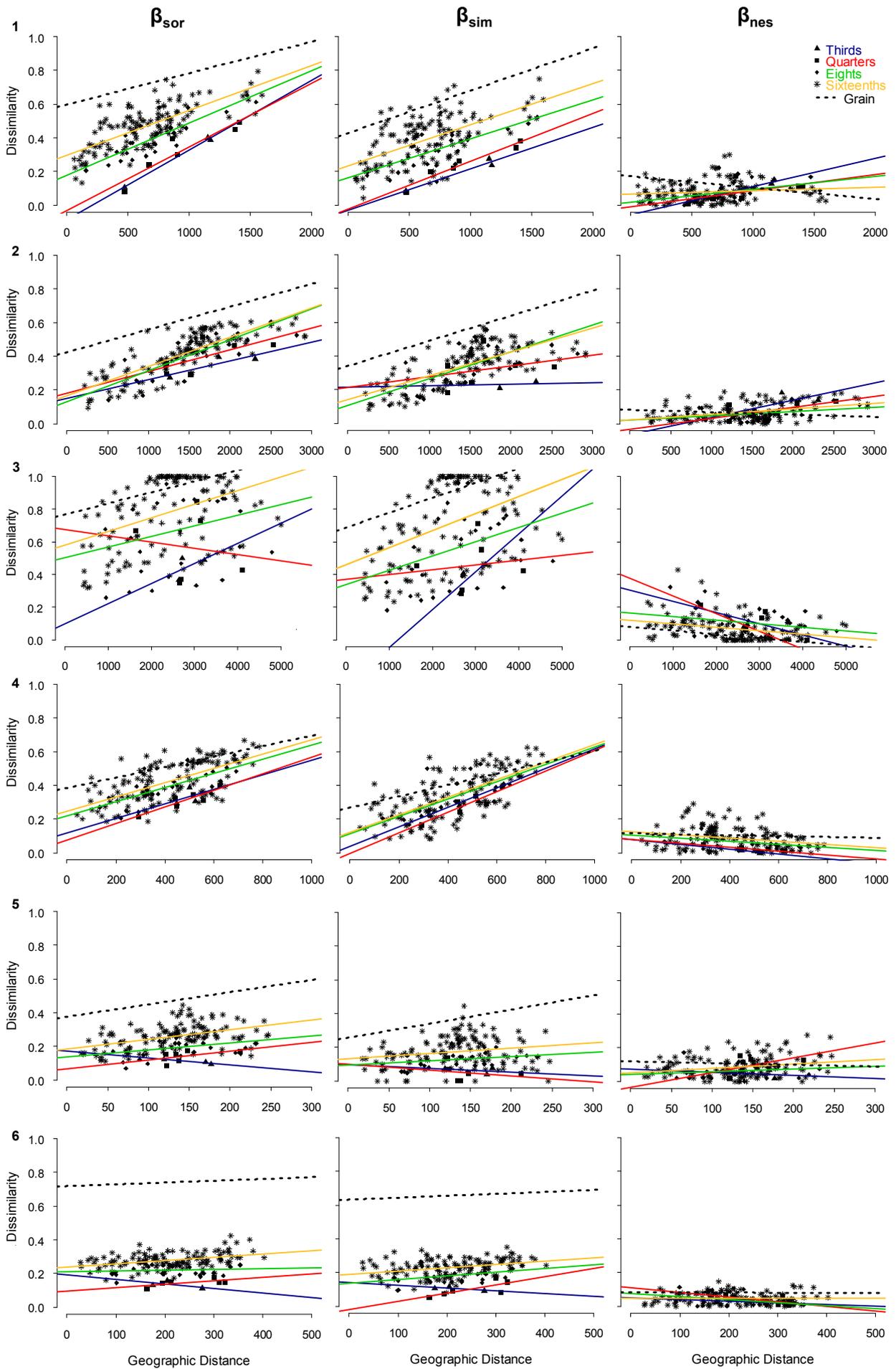


Figure 6.2 β diversity scaling curves, showing the decrease of total β_{SOR} and the turnover component β_{SIM} with increasing sampled area (on a log10 scale). The area plotted is the minimum convex hull value of the sections at each level. The lines represent the nonlinear model fitted to each dataset for each β metric (see Table 6.1).

Figure 6.3 Distance decay relationships for total β diversity (β_{SOR}), spatial turnover (β_{SIM}) and nestedness (β_{NES}) with geographic distance (km). Note that, for consistency with the previous analysis, the plots are showing the inverse of DDS – dissimilarity is expected to increase with distance. Each row is identified by the corresponding ID, with β_{SOR} on the left, β_{SIM} in the centre and β_{NES} on the right. The lines represent the linear model fitted to the pairwise comparisons, with a different colour for each level in the scale gradient (from a single random split; the linear coefficients are presented in Table VII.1). The different plotting symbols represent dissimilarity values for the different scale levels (grain was omitted due to the very large number of points) (next pages).



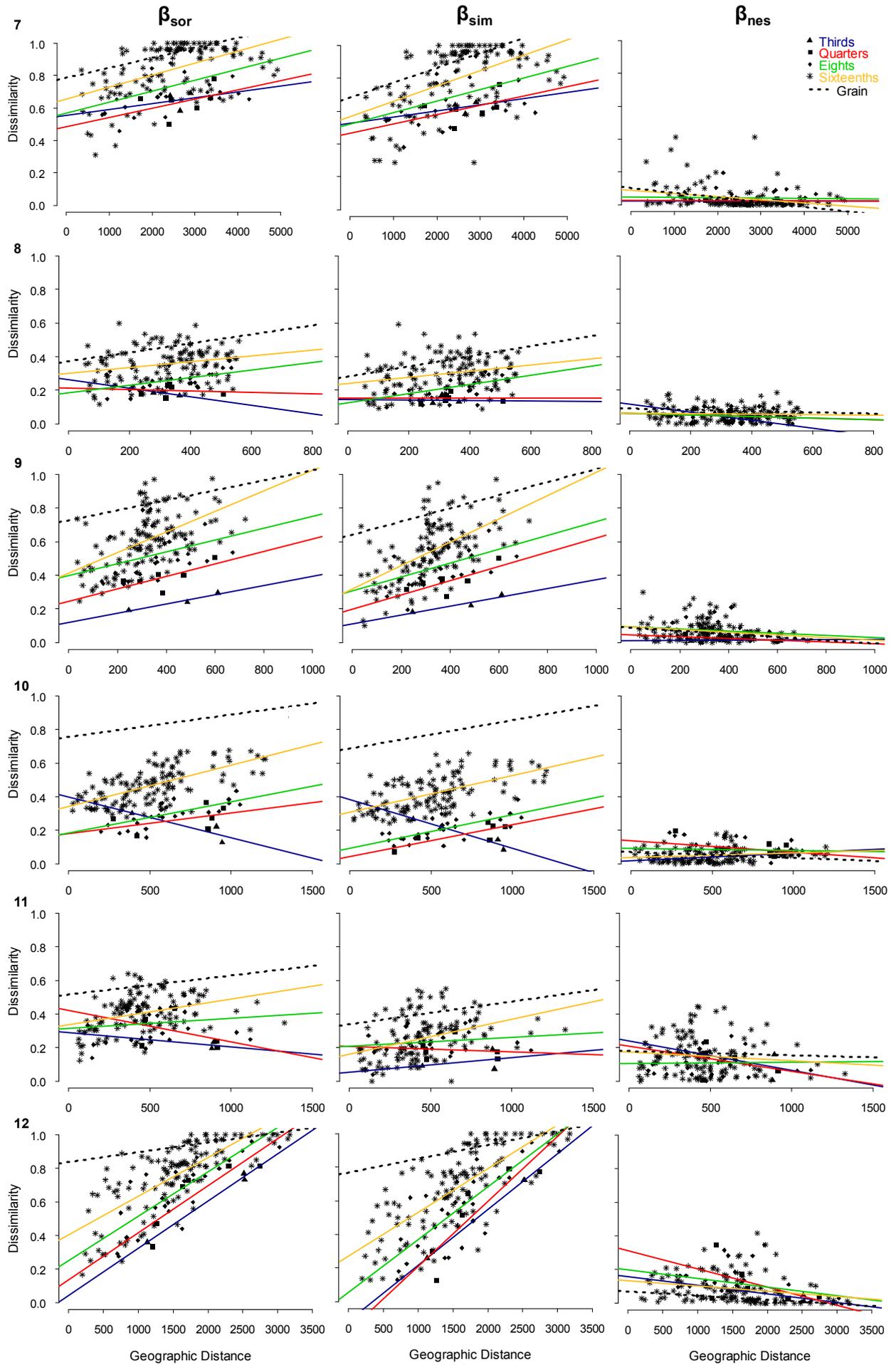


Table 6.1 Nonlinear model fitting results; significantly different estimated parameters are highlighted in bold. Model fitted to each metric was:

gnls ($\beta_{\text{metric}} \sim 1 - (a * \text{Area}^b)$, params = a + b ~ DatasetID).

Coefficients	Dataset ID	β_{SOR}				β_{SIM}				β_{NES}			
		Value	Std.Error	t-value	p-value	Value	Std.Error	t-value	p-value	Value	Std.Error	t-value	p-value
a	1	0.0296	0.0102	2.9009	0.0056	0.0392	0.0151	2.5933	0.0126	1.0009	0.0277	36.1521	0.0000
	2	-0.0241	0.0108	-2.2375	0.0299	-0.0306	0.0163	-1.8842	0.0656	0.0336	0.0501	0.6710	0.5055
	3	-0.0222	0.0113	-1.9633	0.0554	0.0014	0.0258	0.0539	0.9572	-0.0447	0.0341	-1.3098	0.1965
	4	-0.0021	0.0147	-0.1424	0.8873	0.0111	0.0265	0.4186	0.6774	-0.0227	0.0391	-0.5822	0.5632
	5	0.0489	0.0212	2.3083	0.0253	0.0945	0.0379	2.4937	0.0161	-0.0483	0.0343	-1.4081	0.1655
	6	0.0836	0.0242	3.4584	0.0011	0.1034	0.0365	2.8335	0.0067	-0.0211	0.0355	-0.5959	0.5540
	7	-0.0213	0.0131	-1.6341	0.1088	-0.0281	0.0200	-1.4091	0.1653	-0.0115	0.0337	-0.3403	0.7351
	8	0.0553	0.0227	2.4396	0.0184	0.0703	0.0345	2.0409	0.0468	-0.0221	0.0372	-0.5953	0.5545
	9	-0.0071	0.0131	-0.5407	0.5912	-0.0036	0.0218	-0.1663	0.8686	-0.0168	0.0358	-0.4673	0.6424
	10	0.0886	0.0247	3.5861	0.0008	0.1214	0.0364	3.3391	0.0016	-0.0329	0.0345	-0.9533	0.3452
	11	0.0057	0.0145	0.3943	0.6951	0.0321	0.0254	1.2621	0.2130	-0.0139	0.0578	-0.2403	0.8111
	12	-0.0266	0.0107	-2.4884	0.0164	-0.0278	0.0193	-1.4368	0.1573	-0.0216	0.0358	-0.6044	0.5484
b	1	0.2280	0.0260	8.7748	0.0000	0.2213	0.0291	7.5968	0.0000	-0.0097	0.0026	-3.7403	0.0005
	2	0.0825	0.0486	1.6976	0.0960	0.0708	0.0551	1.2842	0.2052	0.0008	0.0041	0.1884	0.8514
	3	0.0439	0.0492	0.8915	0.3771	-0.0476	0.0447	-1.0637	0.2928	0.0052	0.0031	1.7043	0.0948
	4	0.0194	0.0401	0.4848	0.6300	-0.0184	0.0454	-0.4056	0.6869	0.0086	0.0038	2.3049	0.0255
	5	0.0039	0.0353	0.1095	0.9133	-0.0308	0.0394	-0.7834	0.4373	0.0033	0.0036	0.9073	0.3688
	6	-0.0513	0.0315	-1.6269	0.1103	-0.0610	0.0363	-1.6817	0.0991	0.0073	0.0036	2.0217	0.0488
	7	0.0056	0.0686	0.0818	0.9351	-0.0001	0.0812	-0.0006	0.9995	0.0084	0.0030	2.8110	0.0071
	8	-0.0389	0.0336	-1.1567	0.2531	-0.0490	0.0386	-1.2695	0.2104	0.0071	0.0038	1.8762	0.0667
	9	0.0557	0.0403	1.3818	0.1734	0.0272	0.0473	0.5755	0.5676	0.0083	0.0035	2.3459	0.0232
	10	-0.0902	0.0302	-2.9832	0.0045	-0.0987	0.0337	-2.9323	0.0051	0.0051	0.0034	1.4706	0.1479
	11	0.0011	0.0346	0.0317	0.9749	-0.0355	0.0370	-0.9584	0.3427	0.0013	0.0055	0.2276	0.8209
	12	0.0872	0.0730	1.1950	0.2380	0.0175	0.0771	0.2272	0.8212	0.0063	0.0032	2.0068	0.0504

6.4 Discussion

The first comprehensive empirical analysis of β diversity scaling patterns reveals striking consistency: total beta diversity and turnover decreasing with the log of area sampled according to a power law, across the communities analysed. Furthermore, DDS rates were consistent over large increases of area sampled, while grain strongly affected dissimilarity values. In both analyses, turnover accounted for the larger portion of β diversity across all the communities, with the nestedness component being relatively insensitive to large changes in area sampled.

Some of the β diversity scaling patterns found are in accordance with the expectation that higher β diversity values are expected at smaller scales and for more dispersal limited organisms (Nekola & White, 1999; Qian, 2009; Barton *et al.*, 2013). In general, the smaller the area sampled the more dissimilar the compared sites will be (Nekola & White, 1999; Mac Nally *et al.*, 2004; Barton *et al.*, 2013). This is consistent with what I found in both analyses: a decrease in dissimilarity as area increased, and higher overall dissimilarity values for smaller grains for the DDS relationships (significantly higher intercepts for lower levels of the scale gradient). As grain size increases, more species are shared between the sampled areas, and environmental differences are attenuated. Moreover, larger areas will harbour more species and pooling samples to obtain coarser grains results in larger samples and consequently increased probability of sampling more rare species.

There is remarkable similarity in the scaling curves, as well as a consistent behaviour of the three β diversity metrics across the communities analysed. Although the datasets analysed do not include many taxonomic groups, they nonetheless comprise very different taxa, with different ecological and dispersal characteristics. Still, β_{SOR} , β_{SIM} and β_{NES} scaling curves were remarkably consistent, when some variability could be expected due to specific ecological and/or environmental underlying factors (Barton *et al.*, 2013). Similarly, DDS results were also consistent across the communities analysed. Moreover, the datasets analysed covered a wide range of total extent, grain sizes and also of species richness values, suggesting that these results are robust to large variation in these fundamental aspects of ecological studies. These findings are in accordance with one of the most general features of ecological communities – conspecific individuals are spatially aggregated (McGill, 2010b). Here I showed that, within fixed spatial extents and constant species pools, species

are spatially aggregated across a scale gradient spanning several orders of magnitude (from very small local samples to very large regional areas). Plotkin & Muller-Landau (2002) and Morlon *et al.* (2008) have previously shown that species spatial aggregation affects the expected similarity between samples within a regional landscape and DDS rates, respectively. Given the consistency of the β diversity scaling patterns reported here, these results suggest that neither the negative binomial nor the Poisson distributions might be able to adequately describe similarity patterns across scales (from local to regional scales). Specifically, the different functional form of the expected similarity with increasing area (Plotkin & Muller-Landau, 2002), and the inability of the Poisson cluster process completely reproducing the clustering patterns in empirical forest plots (Morlon *et al.*, 2008) suggest that these inconsistencies might be attributable to scale effects. For instance, Morlon *et al.* (2008) noted that the Poisson cluster process only assumes a single scale of aggregation. Finally, the fact that species replacement, rather than species richness, was the main driver of compositional change across the scale gradient offers important insights for studies of temporal change in β diversity, highlighting that despite species richness remaining seemingly stable, substantial changes in community structure might still be occurring across spatial scales, with obvious implications for both our understanding of spatiotemporal dynamics of biodiversity change and conservation (Dornelas *et al.*, 2014; Magurran *et al.*, 2015; McGill *et al.*, 2015; Socolar *et al.*, 2016).

β diversity scaling curves

There were nonetheless some differences between the scaling curves, namely a contraction along the y-axis for some communities. In other words, some communities still showed very high β diversity values even for very large sampled areas (i.e. the dissimilarity value for the bisection level, the highest area in the scaling plots). The overall β diversity for the bisection level was particularly high for the RLS Invertebrates survey around Australia, and for the FIA tree inventory (IDs 7 and 12, respectively). For more heterogeneous landscapes and for organisms with lower dispersal ability, sampled areas are expected to be more dissimilar (Qian, 2009; Si *et al.*, 2015). These two communities can be expected to be less vagile than either birds or marine fish (the majority of the other communities analysed here), and for the spatial framework implemented, i.e. within their respective fixed spatial extents and species pools. While the spatial configuration of the RLS data could also potentially affect β diversity patterns, and I have not tested for this effect (the sampling sites are distributed around Australia, whereas in the other datasets there is a contiguous cloud of more or less dispersed sampling locations across the spatial extent), the fish RLS data (ID3) scaling

patterns were similar to that of the North American BBS for instance (ID2), which lends support to the argument that ecological properties of the different taxa, rather than the spatial configuration of the sampling locations, more strongly affect β diversity patterns in this analysis. These latter two communities were the next ones with relatively high values for β diversity at the bisection level, which could be attributable to the very large spatial extent covered, and the very likely high diversity of habitats included.

Distance decay of similarity

The rate of decay of similarity did not change across very large ranges of area sampled for all the communities analysed. This contrasts with previous reports, namely Keil *et al.* (2012) and Steinbauer *et al.* (2012), but is in accordance with Morlon *et al.*'s (2008) results showing that grain affected overall similarity values, rather than the rate of decay, except at the smallest sampled area. The results from my analysis are consistent with the former, although I found no significant differences in the rates of decay of similarity even for the grain level comparisons. However, the range of area values investigated differs greatly between our studies: grain values ranged from 0.0004 to 6.25 ha in their study, while the scale gradient in my investigation spanned much larger sampled areas, with a very sharp increase from the smallest scale to the subsequent levels. Since Morlon *et al.*'s range is much narrower than my scale gradient, both studies can be reconciled in that DDS rates seem to be robust to large variations in grain, while those variations strongly affect overall dissimilarity values.

Regarding the disparate results to those of Keil *et al.* (2012) and Steinbauer *et al.* (2012), it is likely they can be attributed to the very different spatial framework in which I conducted this analysis. Firstly, both grain and extent were allowed to vary in those studies. Moreover, in Steinbauer *et al.*'s analysis the distance between plots was kept constant while increasing plot size, and DDS slopes were more strongly affected by varying extent than by grain size. In my analysis, I used a fixed spatial extent, so that community data was sampled from a uniform species pool, while the areas sampled and the distance between them were unconstrained and measured from the data for each scaling level. This isolates variation in β diversity from variation in gamma diversity. Secondly, I did not impose static grids overlaid over regional or continental extents, but analysed β diversity patterns across a scale gradient spanning several orders of magnitude. Finally, I used data collected with a very high degree of spatial resolution, from many small representative samples (although I did not use the abundance

information), rather than incidence data across large grid sizes, atlases or simulated data (Beck *et al.*, 2012). However, one caveat of this study was that I was not able to fully explore the scale gradient for the DDS analysis, as there was no statistically robust procedure of including information for the higher levels of the gradient. Nonetheless, these results showed that large increases in area sampled (grain) had no effect on the rates of decay, but did affect the overall dissimilarity values, i.e. the size of areas sampled, rather than the distance between them, more strongly affected dissimilarity patterns across the scale gradient.

β diversity components

Deconstructing total β diversity into its turnover and nestedness components revealed that the two components exhibit divergent spatial patterns, with turnover strongly driving the overall β diversity patterns, for all the communities and across the scale gradient. In contrast, the nestedness component contribution was systematically very low and generally scale insensitive. Since turnover followed the total β diversity pattern very closely, the nestedness component had to exhibit a contrasting pattern, due to the additive nature of the partitioning framework (Baselga, 2010). Interestingly, nestedness was negatively affected by geographic distance only at the lowest level of the scale gradient (and occasionally also at the next level; cf. Wen *et al.* 2016), being seemingly unaffected when calculating dissimilarity between sections at higher levels. Perhaps it is likely that nestedness would be invariant under this fixed extent setting, since new biogeographical areas or more species are not being added as area sampled increases, and hence turnover would be the dominant mechanism driving compositional differences between increasingly larger samples. Nonetheless, it is not straightforward to establish expectations for how nestedness should behave across scales, since the effects of several interacting factors must be considered, namely environmental heterogeneity and dispersal (Tonkin *et al.*, 2016; Gianuca *et al.*, 2017). For instance, Gianuca *et al.* (2017) demonstrated that the relative balance between turnover and nestedness completely depended both on dispersal rates and on landscapes being environmentally heterogeneous or homogeneous, using an experimental metacommunity. Complementary analyses including dispersal ability of organisms and habitat heterogeneity information might provide further insights to help discern how the two components are expected to behave.

Other studies have also reported turnover as the main contributor to β diversity for different taxa, albeit with some relevant nuances. Baselga (2010) showed that turnover was the dominant contributor of β diversity for Southern European longhorn beetles, while both components had similar contributions in Northern Europe. In contrast, Svenning *et al.* (2011) reported that mammal β diversity across Europe was also mainly driven by species turnover, but nestedness showed more equivocal responses at smaller regional scales; the authors suggested that scale could be involved in this discrepancy, due to different scale resolutions used (coarser grain size for the beetles study). Si *et al.* (2015) found a much stronger contribution of turnover for both breeding birds and lizards on islands in China, reporting that turnover decreased with larger differences in island area (and habitat richness), while the nestedness component showed the opposite trend. Finally, Wen *et al.* (2016) reported an increase of turnover with geographic distance, as well as a negative relationship of the nestedness component at the regional scale in their study of small mammals in China along latitudinal and longitudinal directions, while reporting contrasting results for both components at the grid (smaller scale) level. These findings illustrate the relevance of understanding the contributions of the two components systematically across different scales. All of the abovementioned studies however differ from the analyses presented here, in that I did not analyse β diversity patterns under a biogeographical or directional context. My analysis is purely descriptive of how the two components scale with area sampled, being the first investigation of turnover and nestedness components contributions independent of any latitudinal or longitudinal gradient.

Conclusions

The results reported here are remarkably consistent for the three β diversity metrics, across taxa and for both the scaling curves and the distance decay of similarity, providing valuable insights to understanding and synthesizing β diversity scaling patterns. Given the current need to both understand and quantify how biodiversity is changing in the Anthropocene, and the mounting evidence that ecological communities are undergoing biotic homogenization (Magurran *et al.*, 2015), it is of critical importance that we understand how spatial scale can influence such changes, and also how to best monitor communities so that β diversity can effectively inform conservation.

7. General Discussion

Quantifying and understanding how diversity patterns change with spatial scale is a central question for ecological research. By analysing how two fundamental biodiversity patterns are affected by scale, this thesis contributes to a better understanding of community structure across scales. The analyses performed examined empirical communities from a diverse array of taxonomic groups and ecosystems, I improved statistical methods and analysed complementary biodiversity metrics. Moreover, the results here highlight the need for current unified theories of biodiversity to accommodate and explain the variability in SAD shape, while also illustrating consistent patterns across taxa, providing further support for general processes explaining community structure, in terms of both community relative abundances and similarity patterns.

This investigation provided a systematic and comprehensive analysis of SADs and community similarity patterns across scales. First, I have demonstrated that relative abundance patterns in empirical communities deviate from the classical logseries and lognormal distributions, and have (conservatively) estimated that ~15% of the SADs analysed were multimodal. In addition, I have shown that the prevalence of multimodality is higher for larger scale or more taxonomically diverse communities. Currently, no unified theory of biodiversity predicts multiple modes in SAD. Second, I have explored how spatial scale affects the consistency of SAD shape across a scale gradient spanning several orders of magnitude. This analysis showed a clear effect of area, species richness and taxonomic diversity on SAD shape, while total abundance did not exhibit any directional effect. Furthermore, these results confirmed that multimodality is indeed reflecting the structure of the communities (even for intense declines in area sampled, species richness and total abundance), while reinforcing that logseries was never selected for broader scales, and was associated with communities with fewer species and taxonomic families. These findings differ significantly from two important macroecological theories' predictions for species abundances distributions across scales. Third, I have explored the scaling properties of two conceptual types of β diversity, providing the first empirical assessment of β diversity scaling curves for different communities, alongside with its two components, turnover and nestedness. These results revealed remarkable consistency in β diversity scaling patterns, with total β diversity and turnover decreasing with log area according to a power law. Moreover, species replacement, rather than species losses or gains, was the main driver of compositional differences across scales. Additionally, while the rate of decay of similarity was consistent across large variation in areas sampled, for the three β diversity metrics, grain size affected

the overall dissimilarity values. Spatial scale, ecological heterogeneity and species spatial aggregation patterns arose as critical components underlying these results.

7.1 Multimodality and SAD shape across scales

We still lack a thorough understanding of what processes influence the relative abundance of species. The results in chapters 4 and 5 have clearly demonstrated that we need to expand the suite of models used to describe SAD shape, and that multimodality is not a sampling artefact. In addition, this investigation clearly showed a scale effect on SAD shape, both in terms of spatial scale and of taxonomic diversity. No current unified biodiversity theory accommodates the multimodality patterns found here or the variability in SAD shape with scale. All the existing models produce unimodal SAD distributions (McGill *et al.*, 2007), with a single exception, the Emergent Neutrality (EN) model (Vergnon *et al.*, 2012). As both selection and ecological drift act simultaneously on ecological communities (Hubbell, 2001; Vellend, 2010), the current view is that both niche and neutral dynamics contribute to the maintenance of species diversity. The EN model considers that the self-organisation of species along a niche axis allows for very similar species to coexist “within niches”, while differentiation occurs “between niches”. For very similar species, the process of species exclusion is very slow, and hence similar species can coexist for long periods of time (Scheffer & van Nes, 2006; Vergnon *et al.*, 2012). The species occurring at the “core” of these niches are relatively abundant, while the remaining species are rare, and multimodal SADs are observed if the species within the different niches differ significantly in terms of their abundances (Vergnon *et al.*, 2012). This rationale is somewhat similar to the core-transient species dynamics described by Magurran & Henderson (2003), but instead of considering species ecological asymmetries, the EN model is symmetric, with neutrality emerging as a consequence of ecological and evolutionary processes (Holt, 2006; Scheffer & van Nes, 2006). However, the EN model has been criticized for not explicitly modelling “hidden” species differences (Barabás *et al.*, 2013), and an explicit test of the EN model showed no support for a link between a body-size axis and multimodality in SADs (Matthews *et al.*, 2014). On the other hand, producing SADs with multiple modes does not necessarily give support to the underlying mechanism (Vergnon *et al.*, 2012; Barabás *et al.*, 2013), and the authors themselves suggested environmental heterogeneity and species asymmetries as alternative explanations for multimodal SADs (Magurran & Henderson, 2003; Alonso *et al.*, 2008; Dornelas & Connolly, 2008; Dornelas *et al.*, 2009). Finally, there is no explicit consideration of either spatial scale or taxonomic diversity of the communities simulated in the EN model. Here, I have robustly shown that multimodality is linked to larger spatial scales and more diverse communities, and that it is not a sampling artefact (c.f. Barabás *et al.*, 2013).

Spatial scale emerged as a major driver of variability in SAD shape. The results here illustrate a strong departure from two important macroecological theories for SAD at different scales. On one hand, METE predicts a scale invariant logseries SAD, in clear contrast with the strong support for multimodality, and also for the lognormal distribution, in the analyses of over 100 diverse communities. On the other hand, different neutral models predict that SADs become more uneven for larger scales, proposing either a logseries or a “difference logseries” (DLS; with fewer rare species) for the metacommunity scale (Hubbell, 2001; Volkov *et al.*, 2007; Rosindell *et al.*, 2010). Several studies have shown that as scale increases, SADs become more even, i.e. transition from “logseries-like” to “lognormal-like” (Preston, 1948; Magurran, 2004; Connolly *et al.*, 2005; Zillio & He, 2010). Here, I have shown that the scaling of species abundance distributions can further transition into multimodality at larger scales. Dornelas & Connolly (2008) unveiled a multimodal coral SAD where the location of the modes shifted to the right as sample size increased, and Borda-de-Água *et al.* (2012) explicitly described a bimodal SAD when extrapolating the 50 ha BCI forest plot data for larger areas. Both these studies comprise relatively smaller areas compared to the ones analysed in chapters 4, and particularly chapter 5. Additionally, whereas a single lognormal still provided adequate fit for large scales and multimodality can occur for smaller areas, logseries distributions were strongly and consistently associated with smaller scales and with less diverse communities. Support for a logseries distribution at the metacommunity level has been questioned (Magurran, 2005b), and while recently developed neutral models improve the unlikely preponderance of singletons at larger areas (Rosindell *et al.*, 2010), the results here clearly suggest a link between logseries SAD and smaller areas and less diverse communities, both in terms of species richness and number of families. Interestingly, Rosindell & Cornell (2013) also predicted a bimodal SAD when extrapolating the BCI dataset for larger scales using a spatially explicit neutral model without the protracted speciation improvement, but argued that this was only a transient feature towards a logseries SAD for the larger scale. This clearly contrasts with logseries never being selected for any large scale empirical community analysed here.

Spatial scale is continuous, and there are no absolute definitions of scale that ecologists must use; the “adequate” scale will depend on the question, taxa and habitat investigated (Levin, 1992). Thus, cross-scale studies are crucial for improving our understanding of how diversity and the underlying mechanisms operate at different scales. On the other hand, explicit statements of what scale or scales are investigated are essential, as local, regional or metacommunity “scales” are loose definitions, and might represent very different areas for different taxa (Levin, 1992; Magurran, 2005b). Different organisms perceive their environments at inherently different scales, whereas both ecological

heterogeneity and habitat patchiness also depend on the scale of observation, thus suggesting the impossibility of an absolute scale for studying ecological patterns (Levin, 1992). Furthermore, some theoretical tools might be better suited to some particular scales (McGill, 2010b). In this thesis, I have analysed SADs from a broad range of spatial extents (from plots to continents), and I have explored SADs and community similarity patterns across a gradient in spatial scale spanning several orders of magnitude. In doing so, I have greatly expanded the range of scales at which theories deriving SAD have been developed and tested, namely the Neutral Theory and METE. It is possible that differences between my results and these theories' predictions might be partially due to the very different spatial scales analysed, as it is also possible that the exact setting of the spatial scale gradient may also have influenced the results found. Nevertheless, I would argue the results here are robust in that they analysed many communities and employed stringent statistical tools, with very consistent results across taxa and habitats.

Although a case-by-case comparison was never a goal in these investigations, a number of datasets analysed here have been frequently used as empirical support for both neutral and METE models, specifically several tropical forest community plots (e.g. BCI, Korup, Pasoh, Sinharaja, Yasuni, and Lambir), the North American BBS and the FIA tree inventory (e.g. Hubbell, 2001; Etienne *et al.*, 2007; Volkov *et al.*, 2007; Harte *et al.*, 2008; White *et al.*, 2012). Granted that differences in model selection might be due to variations in the specific data analysed (e.g. data from different years), and that model fitting does not indicate which model is “true”, a stringent assessment of PLN mixture models and logseries performance has shown that the SADs of several of these communities are multimodal. Because multimodality is rarely studied and the goal of the abovementioned studies was not to test for multimodality, such models were not considered. As one of the first steps for model comparison is to establish a set of appropriate models to compare (Burnham & Anderson, 2002), this again highlights the importance of including multimodal models in SAD studies, and clearly illustrates the need for macroecological theories to accommodate multimodality in the possible suite of SAD predicted. This integration of multimodal SADs can provide crucial insights into what mechanisms are affecting community structure at different spatial, temporal and organizational scales.

Temporal patterns in species abundance distributions have been largely overlooked (Magurran, 2007). Given the current rates of biodiversity loss and change it is of vital importance to understand how a temporal axis affects biodiversity. Although, no explicit temporal analysis was performed

here, two different years were analysed for the North American BBS data (2011 and 2015 in chapters 4 and 5, respectively), and for both years the SAD was multimodal with strong support. This strongly suggests that the species abundance distribution of this community is consistently multimodal. Further analyses to assess whether multiple modes are a consistent feature of SAD through time, or a transient pattern, will surely provide additional insights into the temporal dynamics of community structure (Magurran, 2007). Furthermore, SADs shape can also change along gradients other than spatial scale, e.g. communities at different stages of succession or under different rates of isolation can be better described by different models (Hubbell 2001, Magurran 2004, McGill *et al.* 2007). Thus, using SADs may also be useful in revealing temporal variation in community structure (Magurran 2007). Finally, the analyses here focused on number of individuals as a measure of abundance. While considering additional abundance measures laid outside the scope of my analyses, systematically analysing patterns for alternative abundance measures across space and time, such as biomass and resource use, may provide relevant insights into how communities are structured and also identify important links between patterns of numerical abundance, body size and energy (Harte *et al.*, 2008; Morlon *et al.*, 2009; Xiao *et al.*, 2015).

7.2 β diversity and spatial scale

β diversity studies are crucial to complement studies of α diversity, in order to better understand the processes that maintain species diversity across space and time. In this thesis, I provided the first empirical assessment of systematic variation of β diversity with spatial scale, showing a highly regular pattern for the scaling of three β diversity metrics with area for different taxa. Previous studies have also reported that species compositional differences are not scale invariant (Plotkin & Muller-Landau, 2002; McGlinn & Hurlbert, 2012). However, the scales explored in such studies have usually been much smaller and using more constrained spatial settings (e.g. Plotkin & Muller-Landau (2002) compared pairs of equally-sized small square plots, and McGlinn & Hurlbert (2012) analysed sets of four small quadrats with varying grain sizes). Furthermore, these studies commonly focus exclusively on terrestrial plants, and do not explicitly consider the two components of β diversity. On the other hand, while the explicit modelling of conspecific spatial aggregation has improved the predictions of community similarity analyses (Plotkin & Muller-Landau, 2002; Morlon *et al.*, 2008), it remains to be investigated if the negative binomial and the Poisson clustered distributions can adequately describe conspecific aggregation across larger ranges of spatial scales, and furthermore for different taxonomic groups, matching the consistency of the β diversity scaling relationships found here.

Across scales and taxonomic groups, turnover was the major driver of compositional patterns. It is possible that within the spatial setting implemented, turnover would be expected to be the main contributor, since a fixed species pool in space and time was analysed – no new areas or species were added, which arguably might not be a very realistic setting. Sampling larger areas and for longer periods results in more species being sampled. Hence, it would be interesting to investigate if a different partitioning approach, e.g. sampling circles with varying area within the total extent instead of using the overall encompassing circle, would yield similar results. It would also be worthwhile to analyse these scaling relationships within a partitioning framework that captured the underlying natural habitat patchiness. On the other hand, analysing how a temporal axis might affect the turnover and nestedness relative contributions will likely lead to a more thorough understanding of how community composition changes over time. Previous studies have also detected a more preponderant role for turnover than for species richness differences (Magurran *et al.*, 2015), and biotic homogenization is a key concept for understanding biodiversity change, as well as being a relevant

issue for ecosystem functioning and resilience (Tilman, 1999; Tilman *et al.*, 2014; Isbell *et al.*, 2015; Oliver *et al.*, 2015).

Regarding the distance decay of similarity with distance, conflicting results on the scale dependency of DDS rates have been reported (Morlon *et al.*, 2008; Steinbauer *et al.*, 2012). Here, I have shown that the rate of similarity remains constant across large increases in area sampled. On the other hand, the expected functional form of the distance decay of similarity remains unresolved (Nekola & McGill, 2014), with exponential, linear and power-law relationships being used. Nekola & McGill (2014) showed that the shape of this relationship is scale dependent: power law decay occurs from small grains or limited extents, where the species pool remains constant, while exponential decay is more prevalent for larger grains or extents, where species pools vary. For simplicity, I have only explored linear DDS patterns here; hence additional investigations using the gradient in spatial scale and comparing the performance of different functional forms would provide further insights into the scaling properties of β diversity and of its components. Finally, METE formulation also incorporates spatial patterns of species aggregation (Harte *et al.*, 2008). However, it has recently been shown that it is not able to accurately describe DDS patterns for several plant communities, consistently overestimating the rate of similarity decay and performing worse than a random placement model, which is known to be a poor model for this pattern, since it does not reproduce the decrease in similarity with distance (McGlenn *et al.*, 2014). It is worth noting again that these derivations are based on the assumption of a scale invariant logseries SAD. Incorporating additional information might prove useful for improving these models' performance in reproducing empirical patterns more accurately.

7.3 Contribution to Macroecology theory

General properties of ecological communities have emerged from the results in this thesis – spatial scale, environmental heterogeneity and spatial aggregation patterns of individuals and species. This investigation has demonstrated how these features impact SAD shape and patterns of community similarity from local to continental scales. On the other hand, there was a clear failure of important macroecological theories to accommodate the variability in SADs shape across spatial scales, and previous studies have also reported that such models cannot fully reproduce empirical community similarity patterns (e.g. Condit *et al.*, 2002; Dornelas *et al.*, 2006; McGlinn *et al.*, 2014). While the investigations in this thesis remained “theory-agnostic”, the extent to which current theories of biodiversity are able to accommodate and explain different biodiversity patterns, as well as incorporating potential variation in such patterns, is a critical criterion for their evaluation and application (McGill, 2003a, 2010b; Xiao *et al.*, 2015).

Unified theories of biodiversity strive not only to explain the pervasive diversity patterns, but also to identify links between the patterns and unite them under a single framework. Six unified theories, including the Neutral Theory (Bell, 2000; Hubbell, 2001), the continuum theory (McGill & Collins, 2003), spatial clustered Poisson (Plotkin & Muller-Landau, 2002; Morlon *et al.*, 2008) and MaxEnt applied to ecology (Pueyo *et al.*, 2007; Harte *et al.*, 2008) were recently reviewed and synthesized by McGill (2010b), who proposed three underlying principles shared by all theories, despite their clear differences in terms of biological assumptions, mathematical formulations and spatial scales involved. These three rules or assertions describe the stochastic geometry of biodiversity:

1. Conspecific individuals are spatially clumped;
2. Abundance between species at larger scales follows a hollow curve;
3. Individuals of different species can be treated as independent and placed regardless of other species (McGill, 2010b).

While there is some empirical evidence for intraspecific clumping of individuals, there might be less support for independent species placement, namely relating to the fact that species are strongly correlated to habitat properties and (can) interact with each other (Wiegand *et al.*, 2012; May *et al.*, 2016). Nonetheless, these assumptions have successively and parsimoniously reproduced

“adequately shaped” macroecological patterns, including the SAR, SAD, the decay of similarity with distance and abundance occupancy correlations (McGill & Collins, 2003; Harte *et al.*, 2008; Morlon *et al.*, 2008; McGill, 2010b). Critically, none of these theories of biodiversity are able to produce multimodal SAD, and some issues regarding the community similarity results across scales can also be discussed.

The findings here for both abundance and similarity patterns can be linked to the first assumption – conspecific individuals are spatially aggregated, one of the most common features of ecological communities. Spatial aggregation has been particular studied for terrestrial plants and at smaller spatial scales (Condit *et al.*, 2000, 2002; Plotkin & Muller-Landau, 2002; Plotkin *et al.*, 2002; Morlon *et al.*, 2008). These results suggest that conspecific aggregation is a relevant mechanism across spatial scales covering several orders of magnitude, and across taxonomic groups. The explicit modelling of conspecific spatial aggregation has considerably improved sampling theories and scaling properties of SAD (Green & Plotkin, 2007) and of species compositional analyses (Plotkin & Muller-Landau, 2002; Morlon *et al.*, 2008). On the other hand, incorporating species asymmetries in terms of aggregation rates has been linked to multiple modes in SAD at local scales, both theoretically and empirically (Alonso *et al.*, 2008; Dornelas & Connolly, 2008). The scaling analyses in chapters 5 and 6 showed that spatial aggregation of conspecifics impacts SAD shape and community similarity across scales. Thus, this is likely a relevant driver unifying the results for the two biodiversity patterns analysed here and explaining the variability in these patterns across scales.

On the other hand, the β diversity scaling analysis also suggested that independent species spatial distributions are unlikely to be able to explain the scaling properties of community similarity across scales. Such assumption has been successful in community similarity analyses using much smaller areas compared to the analysis in chapter 6 (Plotkin & Muller-Landau, 2002; Morlon *et al.*, 2008), and crucially for communities within a relatively homogeneous environment – see Morlon *et al.*'s comparison of forest community plots with different levels of heterogeneity. Finally, Morlon *et al.* (2008) also suggested that interspecific spatial aggregation could potentially affect distance decay relationships by indirectly influencing species abundances and intraspecific aggregation, while further noting that the Poisson cluster process only assumes one scale of aggregation. In addition, the analysis of SAD across the scale gradient suggested that “hitting or missing” areas where species are abundant will strongly affect the shape of abundance distributions, and this is likely connected with both intra- and interspecific spatial patterns of aggregation. Taking into consideration the effect

of both spatial scale and taxonomic diversity on SAD shape, and particularly on the prevalence of multimodality, these results crucially suggest that species spatial independence is unlikely to hold across different scales. Given the diversity of communities analysed and habitats covered across the spatial scale gradient explored, there is a strong suggestion that the assumption of independent interspecific placement is likely violated due to distinct habitat preferences and/or habitat heterogeneity. This links back to the notion that multimodality is associated to ecological heterogeneity, underpinned by larger spatial extents and higher taxonomic breadth.

Ecological heterogeneity, intentionally loosely defined to incorporate the spatial, environmental, taxonomic and functional aspects of ecological systems, emerged as a crucial feature explaining the results found here. The “amount” of heterogeneity incorporated will vary among ecological models, across any of these axes of variation, and will also be scale dependent. Neutral models, clustered Poisson and METE were suggested as more adequate for smaller scales (McGill, 2010b). Once more heterogeneity is integrated, such models are no longer able to accurately describe SAD and community similarity patterns. The findings here indicate that both inter- and intraspecific spatial patterns are relevant to explain SAD variability across scales and β diversity scaling properties. Overall, contrasting the results for the patterns of abundance and community similarity with the unifying rules (or the unified theories thus synthesized), indicates that more information is required to accurately describe biodiversity patterns across scales (Harte *et al.*, 2008; Morlon *et al.*, 2008; Xiao *et al.*, 2015; May *et al.*, 2016). Adding more realistic assumptions, or adjusting assumptions and processes to a scale dependent context can be a way to derive new predictions, and improve the ability of theoretical models to accommodate multimodal SADs on one hand, and incorporate variability in abundance and similarity patterns due to the effect of scale, on the other.

7.4 Conclusions

Understanding the underlying mechanisms of community abundance and composition is critical for biodiversity research. Additionally, understanding the links between relevant scales is imperative to solve the urgent challenge of conserving biodiversity, in the face of high rates of habitat fragmentation and loss, and of global climate change. Here, I have systematically analysed empirical communities focusing on establishing general scaling properties for SADs and community similarity patterns. The consistent deviation in empirical SAD from the predictions of established unified macroecological theories clearly indicates that further developments are necessary. Many authors have argued that, since different processes can originate similar abundance distributions, SADs hold little value in distinguishing different theories. Here, I show that SAD can provide valuable insights for community ecology and macroecology, and furthermore can provide insights into the mechanisms driving community structure across different scales. Additionally, I have provided empirical support for general scaling properties of β diversity metrics across taxa. Both these patterns can also be linked to species spatial patterns across different scales.

Consensus theories on how ecological systems are structured and maintained and how they respond to global change are still elusive. The results in this thesis clearly indicate that unified theories of biodiversity (or their underlying synthetic assumptions) are unable to accommodate the variability in SADs shape and cannot fully reproduce community similarity patterns across scales. On the other hand, elements at the very core of ecological research, namely spatial scale, ecological heterogeneity, as well as intra- and interspecific aggregation patterns emerged as essential for understanding the patterns of abundance and similarity, from macroecological to local scales. Ecological theories that (over) simplify ecological differences between species or ignore biological mechanisms are not able to reproduce either the variability or the scale dependence of these patterns. The rationale for simplification has provided many fruitful insights into community structure and macroecology. Crucially, the results here illustrate that additional information, e.g. incorporating ecological heterogeneity or scale dependent assumptions, will likely improve our ability to accurately describe species diversity patterns across scales. Theoretical frameworks need to be improved or developed to accommodate the empirical variability in diversity patterns across scales, if we are to understand biodiversity, the processes underpinning it, and moreover how biodiversity changes across space and time.

8. References

- Alonso, D., Ostling, A. & Etienne, R.S. (2008) The implicit assumption of symmetry and the species abundance distribution. *Ecology Letters*, **11**, 93–105.
- Anderson, M.J., Crist, T.O., Chase, J.M., Vellend, M., Inouye, B.D., Freestone, A.L., Sanders, N.J., Cornell, H. V., Comita, L.S., Davies, K.F., Harrison, S.P., Kraft, N.J.B., Stegen, J.C. & Swenson, N.G. (2011) Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecology Letters*, **14**, 19–28.
- Baldrige, E., Xiao, X. & White, E.P. (2015) An extensive comparison of species-abundance distribution models. *bioRxiv*.
- Barabás, G., D’Andrea, R., Rael, R., Meszéna, G. & Ostling, A. (2013) Emergent neutrality or hidden niches? *Oikos*, **122**, 1565–1572.
- Barton, P.S., Cunningham, S.A., Manning, A.D., Gibb, H., Lindenmayer, D.B. & Didham, R.K. (2013) The spatial scaling of beta diversity. *Global Ecology and Biogeography*, **22**, 639–647.
- Baselga, A. (2013) Multiple site dissimilarity quantifies compositional heterogeneity among several sites, while average pairwise dissimilarity may be misleading. *Ecography*, **36**, 124–128.
- Baselga, A. (2010) Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, **19**, 134–143.
- Baselga, A., Jiménez-Valverde, A. & Niccolini, G. (2007) A multiple-site similarity measure independent of richness. *Biology Letters*, **3**, 642–645.
- Baselga, A. & Orme, C.D.L. (2012) Betapart: An R package for the study of beta diversity. *Methods in Ecology and Evolution*, **3**, 808–812.
- Beck, J., Ballesteros-Mejia, L., Buchmann, C.M., Dengler, J., Fritz, S.A., Gruber, B., Hof, C., Jansen, F., Knapp, S., Krefth, H., Schneider, A.-K., Winter, M. & Dormann, C.F. (2012) What’s on the horizon for macroecology? *Ecography*, **35**, 673–683.
- Bell, G. (2001) Neutral macroecology. *Science*, **293**, 2413–2418.
- Bell, G. (2000) The distribution of abundance in neutral communities. *The American Naturalist*, **155**, 606–617.

- Bivand, R. & Rundel, C. (2016) rgeos: Interface to Geometry Engine - Open Source (GEOS). R package version 0.3-19.
- Borda-de-Água, L., Borges, P.A.V., Hubbell, S.P. & Pereira, H.M. (2012) Spatial scaling of species abundance distributions. *Ecography*, **35**, 549–556.
- Borda-de-Água, L., Hubbell, S.P. & McAllister, M. (2002) Species-area curves, diversity indices, and species abundance distributions: a multifractal analysis. *The American Naturalist*, **159**, 138–155.
- Brown, J.H. (1995) *Macroecology*, The University of Chicago Press, Chicago.
- Brown, J.H. & Maurer, B.A. (1989) Macroecology: The Division of Food and Space Among Species on Continents. *Science*, **243**, 1145–1150.
- Bulmer, M. (1974) On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics*, **30**, 101–110.
- Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multimodel inference: a practical information-theoretic approach*, 2nd edn. Springer, New York.
- Burnham, K.P. & Anderson, D.R. (2004) Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, **33**, 261–304.
- Butchart, S.H.M., Walpole, M., Collen, B., van Strien, A., Scharlemann, J.P.W., Almond, R.E.A., Baillie, J.E.M., Bomhard, B., Brown, C., Bruno, J., Carpenter, K.E., Carr, G.M., Chanson, J., Chenery, A.M., Csirke, J., Davidson, N.C., Dentener, F., Foster, M., Galli, A., Galloway, J.N., Genovesi, P., Gregory, R.D., Hockings, M., Kapos, V., Lamarque, J.-F., Leverington, F., Loh, J., McGeoch, M.A., McRae, L., Minasyan, A., Hernández Morcillo, M., Oldfield, T.E.E., Pauly, D., Quader, S., Revenga, C., Sauer, J.R., Skolnik, B., Spear, D., Stanwell-Smith, D., Stuart, S.N., Symes, A., Tierney, M., Tyrrell, T.D., Vié, J.-C. & Watson, R. (2010) Global biodiversity: indicators of recent declines. *Science*, **328**, 1164–1168.
- Canty, A. & Ripley, B. (2015) boot: Bootstrap R (S-Plus) Functions. R package version 1.3-15.
- Caswell, H. (1976) Community Structure: A Neutral Model Analysis. *Ecological Monographs*, **46**, 327–354.
- Condit, R., Pitman, N., Leigh, E.G., Chave, J., Terborgh, J., Foster, R.B., Núñez, P., Aguilar, S., Valencia, R., Villa, G., Muller-Landau, H.C., Losos, E. & Hubbell, S.P. (2002) Beta-diversity

- in tropical forest trees. *Science*, **295**, 666–669.
- Condit, R.S., Ashton, P.S., Baker, P.J., Bunyavejchewin, S., Gunatilleke, S., Gunatilleke, N., Hubbell, S.P., Foster, R.B., Itoh, A., LaFrankie, J. V, Lee, H.-S., Losos, E.C., Manokaran, N., Sukumar, R. & Yamakura, T. (2000) Spatial patterns in the distribution of tropical tree species. *Science*, **288**, 1414–1418.
- Connolly, S.R. & Dornelas, M. (2011) *Fitting and empirical evaluation of species abundance models. Biological Diversity: Frontiers in Measurement and Assessment* (ed. by A.E. Magurran) and B.J. McGill), pp. 123–140. Oxford University Press, Oxford.
- Connolly, S.R., Dornelas, M., Bellwood, D.R. & Hughes, T.P. (2009) Testing species abundance models: a new bootstrap approach applied to Indo-Pacific coral reefs. *Ecology*, **90**, 3138–3149.
- Connolly, S.R., Hughes, T.P., Bellwood, D.R. & Karlson, R.H. (2005) Community structure of corals and reef fishes at multiple scales. *Science*, **309**, 1363–1365.
- Connolly, S.R., MacNeil, M.A., Caley, M.J., Knowlton, N., Cripps, E., Hisano, M., Thibaut, L.M., Bhattacharya, B.D., Benedetti-Cecchi, L., Brainard, R.E., Brandt, A., Bulleri, F., Ellingsen, K.E., Kaiser, S., Kroncke, I., Linse, K., Maggi, E., O’Hara, T.D., Plaisance, L., Poore, G.C.B., Sarkar, S.K., Satpathy, K.K., Schuckel, U., Williams, A. & Wilson, R.S. (2014) Commonness and rarity in the marine biosphere. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 8524–8529.
- Connolly, S.R. & Thibaut, L.M. (2012) A comparative analysis of alternative approaches to fitting species-abundance models. *Journal of Plant Ecology*, **5**, 32–45.
- Dewdney, A.K. (1998) A general theory of the sampling process with applications to the “veil line.” *Theoretical Population Biology*, **54**, 294–302.
- Diserud, O.H. & Ødegaard, F. (2007) A multiple-site similarity measure. *Biology Letters*, **3**, 20–22.
- Dornelas, M. & Connolly, S.R. (2008) Multiple modes in a coral species abundance distribution. *Ecology Letters*, **11**, 1008–1016.
- Dornelas, M., Connolly, S.R. & Hughes, T.P. (2006) Coral reef diversity refutes the neutral theory of biodiversity. *Nature*, **440**, 80–82.
- Dornelas, M., Gotelli, N.J., McGill, B., Shimadzu, H., Moyes, F., Sievers, C. & Magurran, A.E. (2014) Assemblage Time Series Reveal Biodiversity Change but Not Systematic Loss. *Science*,

- 344, 296–299.
- Dornelas, M., Moonen, A.C., Magurran, A.E. & Bärberi, P. (2009) Species abundance distributions reveal environmental heterogeneity in modified landscapes. *Journal of Applied Ecology*, **46**, 666–672.
- Engen, S. & Lande, R. (1996) Population dynamic models generating the lognormal species abundance distribution. *Mathematical biosciences*, **132**, 169–183.
- Enquist, B., Sanderson, J. & Weiser, M. (2002) Modeling macroscopic patterns in ecology. *Science*, **295**, 1835–1837.
- Etienne, R.S. (2007) A neutral sampling formula for multiple samples and an “exact” test of neutrality. *Ecology Letters*, **10**, 608–618.
- Etienne, R.S. (2005) A new sampling formula for neutral biodiversity. *Ecology Letters*, **8**, 253–260.
- Etienne, R.S. (2009) Maximum likelihood estimation of neutral model parameters for multiple samples with different degrees of dispersal limitation. *Journal of Theoretical Biology*, **257**, 510–514.
- Etienne, R.S., Apol, M.E.F., Olf, H. & Weissing, F.J. (2007) Modes of speciation and the neutral theory of biodiversity. *Oikos*, **116**, 241–258.
- Fauth, J., Bernardo, J., Camara, M., Resetarits, W., Van Buskirk, J. & McCollum, S. (1996) Simplifying the jargon of community ecology: a conceptual approach. *The American Naturalist*, **147**, 282–286.
- Fisher, R., Corbet, A. & Williams, C. (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, **12**, 42–58.
- Gaston, K.J. (2011) Common Ecology. *BioScience*, **61**, 354–362.
- Gaston, K.J. (2010) Valuing common species. *Science*, **327**, 154–155.
- Gaston, K.J. & Blackburn, T.M. (2000) *Pattern and Process in Macroecology*, (ed. by K.J. Gaston and T.M. Blackburn) Blackwell Science Ltd, Malden, MA, USA.
- Gaston, K.J., Evans, K.L. & Lennon, J.J. (2007) *The scaling of spatial turnover: pruning the thicket*. *Scaling Biodiversity* (ed. by D. Storch, P. Marquet, and J. Brown), pp. 181–222. Cambridge

- University Press, Cambridge.
- Gelman, A. & Hill, J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 1st ed. Cambridge University Press, Cambridge.
- Gianuca, A.T., Declerck, S.A.J.J., Lemmens, P. & De Meester, L. (2017) Effects of dispersal and environmental heterogeneity on the replacement and nestedness components of β -diversity. *Ecology*, **98**, 525–533.
- Gray, J.S., Bjørgesæter, A. & Ugland, K.I. (2006) On plotting species abundance distributions. *The Journal of Animal Ecology*, **75**, 752–756.
- Gray, J.S., Bjørgesæter, A. & Ugland, K.I. (2005) The impact of rare species on natural assemblages. *Journal of Animal Ecology*, **74**, 1131–1139.
- Green, J.L. & Plotkin, J.B. (2007) A statistical theory for sampling species abundances. *Ecology Letters*, **10**, 1037–1045.
- Grøtan, V. & Engen, S. (2008) poilog: Poisson lognormal and bivariate Poisson lognormal distribution. R package version 0.4.
- Hadfield, J.D. (2010) MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software*, **33**.
- Haegeman, B. & Etienne, R.S. (2010) Entropy maximization and the spatial distribution of species. *The American Naturalist*, **175**, E74–E90.
- Haegeman, B. & Loreau, M. (2008) Limitations of entropy maximization in ecology. *Oikos*, **117**, 1700–1710.
- Halpern, B.S., Frazier, M., Potapenko, J., Casey, K.S., Koenig, K., Longo, C., Lowndes, J.S., Rockwood, R.C., Selig, E.R., Selkoe, K.A. & Walbridge, S. (2015) Spatial and temporal changes in cumulative human impacts on the world's ocean. *Nature Communications*, **6**, 7615.
- Hankin, R.K.S. (2007) Introducing untb, an R package for simulating ecological drift under the unified neutral theory of biodiversity. *Journal of Statistical Software*, **22**.
- Harrison, S., Ross, S.J. & Lawton, J.H. (1992) Beta Diversity on Geographic Gradients in Britain. *Journal of Animal Ecology*, **61**, 151–158.
- Harte, J., Kinzig, A. & Green, J. (1999) Self-similarity in the distribution and abundance of species.

- Science*, **284**, 334–336.
- Harte, J. & Kinzig, A.P. (1997) On the Implications of Species-Area Relationships for Endemism, Spatial Turnover, and Food Web Patterns. *Oikos*, **80**, 417–427.
- Harte, J. & Newman, E.A. (2014) Maximum information entropy: A foundation for ecological theory. *Trends in Ecology and Evolution*, **29**, 384–389.
- Harte, J., Smith, A.B. & Storch, D. (2009) Biodiversity scales from plots to biomes with a universal species-area curve. *Ecology Letters*, **12**, 789–797.
- Harte, J., Zillio, T., Conlisk, E. & Smith, A. (2008) Maximum entropy and the state-variable approach to macroecology. *Ecology*, **89**, 2700–2711.
- Henson, J.M., Reise, S.P. & Kim, K.H. (2007) Detecting Mixtures From Structural Model Differences Using Latent Variable Mixture Modeling: A Comparison of Relative Model Fit Statistics. *Structural Equation Modeling: A Multidisciplinary Journal*, **14**, 202–226.
- Holt, R.D. (2006) Emergent neutrality. *Trends in Ecology and Evolution*, **21**, 531–533.
- Hubbell, S.P. (2001) *The Unified Neutral Theory of Biodiversity and Biogeography*, Princeton University Press, Princeton.
- Hutchinson, G. (1959) Homage to Santa Rosalia or why are there so many kinds of animals? *The American Naturalist*, **93**, 145–159.
- Isbell, F., Craven, D., Connolly, J., Loreau, M., Schmid, B., Beierkuhnlein, C., Bezemer, T.M., Bonin, C., Bruelheide, H., de Luca, E., Ebeling, A., Griffin, J.N., Guo, Q., Hautier, Y., Hector, A., Jentsch, A., Kreyling, J., Lanta, V., Manning, P., Meyer, S.T., Mori, A.S., Naeem, S., Niklaus, P. a., Polley, H.W., Reich, P.B., Roscher, C., Seabloom, E.W., Smith, M.D., Thakur, M.P., Tilman, D., Tracy, B.F., van der Putten, W.H., van Ruijven, J., Weigelt, A., Weisser, W.W., Wilsey, B. & Eisenhauer, N. (2015) Biodiversity increases the resistance of ecosystem productivity to climate extremes. *Nature*, **526**, 574–577.
- Janzen, T., Haegeman, B. & Etienne, R.S. (2015) A sampling formula for ecological communities with multiple dispersal syndromes. *Journal of Theoretical Biology*, **374**, 94–106.
- Keil, P., Schweiger, O., Kühn, I., Kunin, W.E., Kuussaari, M., Settele, J., Henle, K., Brotons, L., Pe'er, G., Lengyel, S., Moustakas, A., Steinicke, H. & Storch, D. (2012) Patterns of beta diversity in Europe: the role of climate, land cover and distance across scales. *Journal of*

- Biogeography*, **39**, 1473–1486.
- Knape, J. & de Valpine, P. (2012) Are patterns of density dependence in the Global Population Dynamics Database driven by uncertainty about population abundance? *Ecology Letters*, **15**, 17–23.
- Koleff, P., Gaston, K.J. & Lennon, J.J. (2003) Measuring beta diversity for presence–absence data. *Journal of Animal Ecology*, **72**, 367–382.
- Leibold, M.A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J.M., Hoopes, M.F., Holt, R.D., Shurin, J.B., Law, R., Tilman, D., Loreau, M. & Gonzalez, A. (2004) The metacommunity concept: a framework for multi-scale community ecology. *Ecology Letters*, **7**, 601–613.
- Lennon, J.J., Koleff, P., Greenwood, J.J.D. & Gaston, K.J. (2001) The geographical structure of British bird distributions: Diversity, spatial turnover and scale. *Journal of Animal Ecology*, **70**, 966–979.
- Levin, S. (1992) The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. *Ecology*, **73**, 1943–1967.
- Levin, S.A. (2000) Multiple Scales and the Maintenance of Biodiversity. *Ecosystems*, **3**, 498–506.
- Locey, K.J. & White, E.P. (2013) How species richness and total abundance constrain the distribution of abundance. *Ecology Letters*, **16**, 1177–1185.
- MacArthur, R. (1960) On the relative abundance of species. *American Naturalist*, **94**, 25–36.
- MacArthur, R.H. (1972) *Geographical ecology*, Princeton University Press, Princeton, NJ.
- MacArthur, R.H. (1957) On the relative abundance of bird species. *Proceedings of the National Academy of Sciences of the United States of America*, **43**, 293–295.
- Magurran, A.E. (2005a) Biological diversity. *Current Biology*, **15**, R116–R118.
- Magurran, A.E. (2004) *Measuring biological diversity*, Blackwell Science, Oxford.
- Magurran, A.E. (2005b) Species abundance distributions: pattern or process? *Functional Ecology*, **19**, 177–181.
- Magurran, A.E. (2007) Species abundance distributions over time. *Ecology Letters*, **10**, 347–354.
- Magurran, A.E. & Dornelas, M. (2010) Biological diversity in a changing world. *Philosophical*

- transactions of the Royal Society of London. Series B, Biological sciences*, **365**, 3593–3597.
- Magurran, A.E., Dornelas, M., Moyes, F., Gotelli, N.J. & McGill, B. (2015) Rapid biotic homogenization of marine fish assemblages. *Nature Communications*, **6**, 8405.
- Magurran, A.E. & Henderson, P.A. (2003) Explaining the excess of rare species in natural species abundance distributions. *Nature*, **422**, 714–716.
- Marquet, P., Fernández, M., Navarrete, S. & Valdovinos, C. (2004) *Diversity emerging: toward a deconstruction of biodiversity patterns. Frontiers of Biogeography: New directions in the Geography of Nature* (ed. by M. Lomolino) and L. Heaney), pp. 191–209. Cambridge University Press, Cambridge.
- Matthews, T.J., Borges, P.A. V & Whittaker, R.J. (2014) Multimodal species abundance distributions: A deconstruction approach reveals the processes behind the pattern. *Oikos*, **123**, 533–544.
- May, F., Wiegand, T., Lehmann, S. & Huth, A. (2016) Do abundance distributions and species aggregation correctly predict macroecological biodiversity patterns in tropical forests? *Global Ecology and Biogeography*, **25**, 575–585.
- May, R.M. (1975) *Patterns of species abundance and diversity. Ecology and Evolution of Communities* (ed. by M.L. Cody) and J.M. Diamond), pp. 81–120. Belknap Press of Harvard University Press, Cambridge.
- McGill, B. (2003a) Strong and weak tests of macroecological theory. *Oikos*, **102**, 679–685.
- McGill, B. & Collins, C. (2003) A unified theory for macroecology based on spatial patterns of abundance. *Evolutionary Ecology Research*, **5**, 469–492.
- McGill, B., Maurer, B. & Weiser, M. (2006) Empirical evaluation of neutral theory. *Ecology*, **87**, 1411–1423.
- McGill, B.J. (2003b) A test of the unified neutral theory of biodiversity. *Nature*, **422**, 881–885.
- McGill, B.J. (2003c) Does Mother Nature really prefer rare species or are log-left-skewed SADs a sampling artefact? *Ecology Letters*, **6**, 766–773.
- McGill, B.J. (2010a) Matters of scale. *Science*, **328**, 575–576.
- McGill, B.J. (2011) *Species abundance distributions. Biological Diversity: Frontiers in*

- Measurement and Assessment* (ed. by A.E. Magurran) and B.J. McGill), pp. 105–122. Oxford University Press, Oxford.
- McGill, B.J. (2010b) Towards a unification of unified theories of biodiversity. *Ecology Letters*, **13**, 627–642.
- McGill, B.J., Dornelas, M., Gotelli, N.J. & Magurran, A.E. (2015) Fifteen forms of biodiversity trend in the anthropocene. *Trends in Ecology and Evolution*, **30**, 104–113.
- McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K., Dornelas, M., Enquist, B.J., Green, J.L., He, F., Hurlbert, A.H., Magurran, A.E., Marquet, P.A., Maurer, B.A., Ostling, A., Soykan, C.U., Ugland, K.I. & White, E.P. (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, **10**, 995–1015.
- McGlenn, D. & White, E. (2015) ecoretriever: R Interface to the EcoData Retriever. R package version 0.2.1.
- McGlenn, D.J. & Hurlbert, A.H. (2012) Scale dependence in species turnover reflects variance in species occupancy. *Ecology*, **93**, 294–302.
- McGlenn, D.J., Xiao, X., Kitzes, J. & White, E.P. (2014) Exploring the spatially explicit predictions of the Maximum Entropy Theory of Ecology. *Global Ecology and Biogeography*, **24**, 675–684.
- McLachlan, G. & Peel, D. (2000) *Finite Mixture Models*, John Wiley & Sons, New York.
- McLachlan, G.J. (1987) On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture. *Applied Statistics*, **36**, 318–324.
- Morlon, H., Chuyong, G., Condit, R., Hubbell, S., Kenfack, D., Thomas, D., Valencia, R. & Green, J.L. (2008) A general framework for the distance-decay of similarity in ecological communities. *Ecology Letters*, **11**, 904–917.
- Morlon, H., White, E.P., Etienne, R.S., Green, J.L., Ostling, A., Alonso, D., Enquist, B.J., He, F., Hurlbert, A., Magurran, A.E., Maurer, B.A., McGill, B.J., Olf, H., Storch, D. & Zillio, T. (2009) Taking species abundance distributions beyond individuals. *Ecology Letters*, **12**, 488–501.
- Morris, B.D. & White, E.P. (2013) The EcoData Retriever: Improving Access to Existing Ecological Data. *PLoS ONE*, **8**, e65848.

- Motomura, I. (1932) On the statistical treatment of communities. *Zoological Magazine*, **44**, 379–383.
- Mac Nally, R., Fleishman, E., Bulluck, L.P. & Betrus, C.J. (2004) Comparative influence of spatial scale on beta diversity within regional assemblages of birds and butterflies. *Journal of Biogeography*, **31**, 917–929.
- Nekola, J.C. & McGill, B.J. (2014) Scale dependency in the functional form of the distance decay relationship. *Ecography*, **37**, 309–320.
- Nekola, J.C. & White, P.S. (1999) The distance decay of similarity in biogeography and ecology. *Journal of Biogeography*, **26**, 867–878.
- O'Dwyer, J.P., Lake, J.K., Ostling, A., Savage, V.M. & Green, J.L. (2009) An integrative framework for stochastic, size-structured community assembly. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 6170–6175.
- Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E. & Wagner, H. (2016) *vegan: Community Ecology Package*. R package version 2.4-0.
- Oliver, T.H., Heard, M.S., Isaac, N.J.B., Roy, D.B., Procter, D., Eigenbrod, F., Freckleton, R., Hector, A., Orme, C.D.L., Petchey, O.L., Proença, V., Raffaelli, D., Suttle, K.B., Mace, G.M., Martín-López, B., Woodcock, B.A. & Bullock, J.M. (2015) Biodiversity and Resilience of Ecosystem Functions. *Trends in Ecology & Evolution*, **30**, 673–684.
- Pebesma, E.J. & Bivand, R.S. (2005) Classes and methods for spatial data in R. *R News* 5 (2), <http://cran.r-project.org/doc/Rnews/>.
- Pereira, H.M., Leadley, P.W., Proença, V., Alkemade, R., Scharlemann, J.P.W., Fernandez-Manjarrés, J.F., Araújo, M.B., Balvanera, P., Biggs, R., Cheung, W.W.L., Chini, L., Cooper, H.D., Gilman, E.L., Guénette, S., Hurtt, G.C., Huntington, H.P., Mace, G.M., Oberdorff, T., Revenga, C., Rodrigues, P., Scholes, R.J., Sumaila, U.R. & Walpole, M. (2010) Scenarios for global biodiversity in the 21st century. *Science*, **330**, 1496–1501.
- Pereira, H.M., Navarro, L.M. & Martins, I.S. (2012) Global biodiversity change: the bad, the good, and the unknown. *Annual Review of Environment and Resources*, **37**, 25–50.
- Pielou, E.C. (1969) *An Introduction to Mathematical Ecology*, Wiley-Interscience, New York.
- Pielou, E.C. (1975) *Ecological diversity*, Wiley-Interscience, New York.

- Pielou, E.C. (1977) *Mathematical Ecology*, John Wiley & Sons, New York.
- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team (2016) nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-128.
- Plotkin, J.B., Chave, J. & Ashton, P.S. (2002) Cluster analysis of spatial patterns in Malaysian tree species. *The American naturalist*, **160**, 629–644.
- Plotkin, J.B. & Muller-Landau, H.C. (2002) Sampling the Species Composition of a Landscape. *Ecology*, **83**, 3344–3356.
- Preston, F. (1948) The commonness, and rarity, of species. *Ecology*, **29**, 254–283.
- Pueyo, S. (2006) Diversity: between neutrality and structure. *Oikos*, **112**, 392–405.
- Pueyo, S., He, F. & Zillio, T. (2007) The maximum entropy formalism and the idiosyncratic theory of biodiversity. *Ecology Letters*, **10**, 1017–1028.
- Qian, H. (2009) Global comparisons of beta diversity among mammals, birds, reptiles, and amphibians across spatial scales and taxonomic ranks. *Journal of Systematics and Evolution*, **47**, 509–514.
- R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.
- Ricklefs, R. (1987) Community diversity: relative roles of local and regional processes. *Science*, **235**, 167–171.
- Ricklefs, R.E. (2008) Disintegration of the ecological community. *The American Naturalist*, **172**, 741–750.
- Rosenzweig, M.L. (1995) *Species Diversity in Space and Time*, Cambridge University Press.
- Rosindell, J. & Cornell, S.J. (2013) Universal scaling of species-abundance distributions across multiple scales. *Oikos*, **122**, 1101–1111.
- Rosindell, J., Cornell, S.J., Hubbell, S.P. & Etienne, R.S. (2010) Protracted speciation revitalizes the neutral theory of biodiversity. *Ecology Letters*, **13**, 716–727.
- Rosindell, J., Hubbell, S.P. & Etienne, R.S. (2011) The unified neutral theory of biodiversity and biogeography at age ten. *Trends in Ecology and Evolution*, **26**, 340–348.

- Sax, D.F. & Gaines, S.D. (2003) Species diversity: From global decreases to local increases. *Trends in Ecology and Evolution*, **18**, 561–566.
- Scheffer, M. & van Nes, E.H. (2006) Self-organized similarity, the evolutionary emergence of groups of similar species. *Proceedings of the National Academy of Sciences*, **103**, 6230–6235.
- Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.
- Si, X., Baselga, A. & Ding, P. (2015) Revealing Beta-Diversity Patterns of Breeding Bird and Lizard Communities on Inundated Land-Bridge Islands by Separating the Turnover and Nestedness Components. *PLOS ONE*, **10**, e0127692.
- Simpson, G.G. (1943) Mammals and the nature of continents. *American Journal of Science*, **241**, 1–31.
- Socolar, J.B., Gilroy, J.J., Kunin, W.E. & Edwards, D.P. (2016) How Should Beta-Diversity Inform Biodiversity Conservation? *Trends in Ecology & Evolution*, **31**, 67–80.
- Sørensen, T.J. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab*, **5**, 1–34.
- Steinbauer, M.J., Dolos, K., Reineking, B. & Beierkuhnlein, C. (2012) Current measures for distance decay in similarity of species composition are influenced by study extent and grain size. *Global Ecology and Biogeography*, **21**, 1203–1212.
- Storch, D., Keil, P. & Jetz, W. (2012) Universal species-area and endemics-area relationships at continental scales. *Nature*, **488**, 78–81.
- Sugihara, G. (1980) Minimal community structure: an explanation of species abundance patterns. *American Naturalist*, **116**, 770–787.
- Supp, S. & Ernest, S. (2014) Species-level and community-level responses to disturbance: a cross-community analysis. *Ecology*, **95**, 1717–1723.
- Svenning, J.-C., Fløjgaard, C. & Baselga, A. (2011) Climate, history and neutrality as drivers of mammal beta diversity in Europe: insights from multiscale deconstruction. *Journal of Animal Ecology*, **80**, 393–402.
- Tilman, D. (1999) The ecological consequences of changes in biodiversity: a search for general principles. *Ecology*, **80**, 1455–1474.

- Tilman, D., Isbell, F. & Cowles, J.M. (2014) Biodiversity and Ecosystem Functioning. *Annual Review of Ecology, Evolution, and Systematics*, **45**, 471–493.
- Tokeshi, M. (1990) Niche apportionment or random assortment: species abundance patterns revisited. *The Journal of Animal Ecology*, **59**, 1129–1146.
- Tokeshi, M. (1996) Power Fraction: A New Explanation of Relative Abundance Patterns in Species-Rich Assemblages. *Oikos*, **75**, 543–550.
- Tokeshi, M. (1993) Species Abundance Patterns and Community Structure. *Advances in Ecological Research*, **24**, 111–186.
- Tokeshi, M. (1999) *Species coexistence: ecological and evolutionary perspective*, Blackwell Science Ltd, Oxford.
- Tonkin, J.D., Stoll, S., Jähnig, S.C. & Haase, P. (2016) Contrasting metacommunity structure and beta diversity in an aquatic-floodplain system. *Oikos*, **125**, 686–697.
- 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. *Ecography*, **33**, 2–22.
- Tuomisto, H. (2010b) A diversity of beta diversities: straightening up a concept gone awry. Part 2. Quantifying beta diversity and related phenomena. *Ecography*, **33**, 23–45.
- Ugland, K. & Gray, J. (1982) Lognormal distributions and the concept of community equilibrium. *Oikos*, **39**, 171–178.
- Ulrich, W., Baselga, A., Kusumoto, B., Shiono, T., Tuomisto, H. & Kubota, Y. (2016) The tangled link between β - and γ -diversity: a Narcissus effect weakens statistical inferences in null model analyses of diversity patterns. *Global Ecology and Biogeography*, **26**, 1–5.
- Vellend, M. (2010) Conceptual synthesis in community ecology. *The Quarterly Review of Biology*, **85**, 183–206.
- Venter, O., Sanderson, E.W., Magrath, A., Allan, J.R., Beher, J., Jones, K.R., Possingham, H.P., Laurance, W.F., Wood, P., Fekete, B.M., Levy, M.A. & Watson, J.E.M. (2016) Sixteen years of change in the global terrestrial human footprint and implications for biodiversity conservation. *Nature Communications*, **7**, 12558.
- Vergnon, R., van Nes, E.H. & Scheffer, M. (2012) Emergent neutrality leads to multimodal species abundance distributions. *Nature Communications*, **3**, 663.

- Volkov, I., Banavar, J.R., Hubbell, S.P. & Maritan, A. (2003) Neutral theory and relative species abundance in ecology. *Nature*, **424**, 1035–1037.
- Volkov, I., Banavar, J.R., Hubbell, S.P. & Maritan, A. (2007) Patterns of relative species abundance in rainforests and coral reefs. *Nature*, **450**, 45–49.
- Wen, Z., Yang, Q., Quan, Q., Xia, L., Ge, D. & Lv, X. (2016) Multiscale partitioning of small mammal β -diversity provides novel insights into the Quaternary faunal history of Qinghai-Tibetan Plateau and Hengduan Mountains. *Journal of Biogeography*, **43**, 1412–1424.
- White, E. (2007) *Spatiotemporal scaling of species richness: patterns, processes, and implications*. *Scaling Biodiversity* (ed. by D. Storch, P.A. Marquet, and J.H. Brown), pp. 325–346. Cambridge University Press.
- White, E.P., Thibault, K.M. & Xiao, X. (2012) Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model. *Ecology*, **93**, 1772–1778.
- Whittaker, R.H. (1972) Evolution and Measurement of Species Diversity. *Taxon*, **21**, 213–251.
- Whittaker, R.H. (1960) Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs*, **30**, 279–338.
- Wickham, H. (2009) *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York, New York.
- Wiegand, T., Huth, A., Getzin, S., Wang, X., Hao, Z., Gunatilleke, C.V.S. & Gunatilleke, I.A.U.N. (2012) Testing the independent species' arrangement assertion made by theories of stochastic geometry of biodiversity. *Proceedings of the Royal Society B*, **279**, 3312–3320.
- Wiens, J.A. (1989) Spatial Scaling in Ecology. *Functional Ecology*, **3**, 385–397.
- Williams, C.B. (1943) Area and number of species. *Nature*, **152**, 264–267.
- Xiao, X., McGlenn, D.J. & White, E.P. (2015) A strong test of the Maximum Entropy Theory of Ecology. *The American Naturalist*, **185**, E70–E80.
- Xiao, X., O'Dwyer, J.P. & White, E.P. (2016) Comparing process-based and constraint-based approaches for modeling macroecological patterns. *Ecology*, **97**, 1228–1238.
- Zacari, A., Brayard, A., Dommergues, J.-L., Meister, C., Escarguel, G., Laffont, R., Vrielynck, B. & Fara, E. (2016) Gauging scale effects and biogeographical signals in similarity distance decay

analyses: an Early Jurassic ammonite case study. *Palaeontology*, **59**, 671–687.

Zillio, T. & He, F. (2010) Inferring species abundance distribution across spatial scales. *Oikos*, **119**, 71–80.

Appendix I

Table I.1 Empirical datasets information, showing the number of families and classification in terms of spatial extent, as well as the selected model for each SAD. Datasets ID 1-25 were retrieved from OBIS, ID 26-110 from Ecological Data Wiki and ID 111-117 from GBIF.

ID	References	Dataset Name	Taxon	Organism	Realm	Climatic region	Habitat / Biome	Data usage	Species Richness	Number of individuals	Spatial extent	Number of families	Model selected
1	Woehler, 1999	Seabirds of the Southern and South Indian Ocean (Australian Antarctic Data Centre)	Birds	Seabirds	Marine	Polar / Temperate	Coastal habitats	Used last year sampled with enough records - 2005; incidental records of marine mammals discarded	46	14450	Continental	5	1PLN
2	Bakker & Herman, 1990; Bakker et al., 1994	Phytoplankton in the Oosterschelde before, during and after the storm-surge barrier (1982-1990) (EUROBIS)	Marine plants	Phytoplankton	Marine	Temperate	Estuarine	Used last year sampled - 1990	85	13058	Local	40	1PLN
3	Degraer et al., 2006	Macrobelt: Long term trends in the macrobenthos of the Belgian Continental Shelf (EurOBIS)	Benthos	Macrobenthos	Marine	Temperate	Benthic	Used last year sampled - 2001	157	248815	Local	70	Multimodal
4	DATRAS, 2010f	ROCKALL: Scottish Rockall Survey for commercial fish species (EurOBIS)	Fish	Fish	Marine	Temperate	Pelagic / Bottom waters	Used last year sampled - 2009 (only Fish data)	39	3821469	Regional	22	Logser
5	DATRAS, 2010e	Northern Irish Ground Fish Trawl Survey (EurOBIS)	Fish	Groundfish	Marine	Temperate	Pelagic / Bottom waters	Used last year sampled - 2008 (only Fish data)	75	504318	Regional	35	Logser
6	DATRAS, 2010d	Irish Ground Fish Survey for commercial fish species (EurOBIS)	Fish	Groundfish	Marine	Temperate	Pelagic / Bottom waters	Used last year sampled - 2008 (only Fish data)	110	8932753	Regional	54	Logser
7	DATRAS, 2010b	ICES French Southern Atlantic Bottom Trawl Survey for commercial fish species (EurOBIS)	Fish	Fish	Marine	Temperate	Pelagic / Bottom waters	Used last year sampled - 2007 (only Fish data)	124	8143303	Regional	54	Logser

8	DATRAS, 2010a	ICES Beam Trawl Survey for commercial fish species (EurOBIS)	Fish	Fish	Marine	Temperate	Pelagic / Bottom waters	Used last year sampled - 2010 (only Fish data)	39	962807	Regional	22	Logser
9	DATRAS, 2010g	Scottish West Coast Survey for commercial fish species (EurOBIS)	Fish	Fish	Marine	Temperate	Pelagic / Bottom waters	Used last year sampled - 2010 (only Fish data)	74	12978222	Regional	35	Logser
10	DATRAS, 2010c	ICES North Sea International Bottom Trawl Survey for commercial fish species (EurOBIS)	Fish	Fish	Marine	Temperate	Pelagic / Bottom waters	Used last year sampled - 2011 (only Fish data)	132	114794968	Regional	52	1PLN
11	NOAA/NOS /NCCOS/CCMA, 2007b	St. Croix, USVI Fish Assessment and Monitoring Data (2002 - Present) (NOAA-CCMA)	Fish	Fish	Marine	Tropical	Reef	Used last year sampled - 2010	134	27278	Local	42	Logser
12	NOAA/NOS /NCCOS/CCMA, 2007a	La Parguera, Puerto Rico Fish Assessment and Monitoring Data (2002 - Present) (NOAA-CCMA)	Fish	Fish	Marine	Tropical	Reef	Used last year sampled - 2010	155	30758	Local	44	1PLN
13	NOAA/NOS /NCCOS/CCMA, 2007c	St. John, USVI Fish Assessment and Monitoring Data (2002 - Present) (NOAA-CCMA)	Fish	Fish	Marine	Tropical	Reef	Used last year sampled - 2010	164	40033	Local	43	1PLN
14	CSIRO	CSIRO Marine Data Warehouse (OBIS Australia)	Fish	Fish	Marine	Tropical / Temperate	Mixed	Used last year sampled with enough records - 1988	206	337485	Continental	78	1PLN
15	Wade, 2011	Snow crab research trawl survey database (Southern Gulf of St. Lawrence, Gulf region, Canada) from 1988 to 2010 (OBIS Canada)	Benthos	Benthos	Marine	Temperate	Benthic	Used last year sampled - 2009	32	2308239	Regional	26	1PLN
16	Clark & Branton, 2007	DFO Maritimes Research Vessel Trawl Surveys Fish Observations (OBIS Canada)	Fish	Demersal Fish	Marine	Temperate	Demersal	Used last year sampled - 2011	129	230445	Regional	57	1PLN
17	Brown et al., 2005	ECNASAP - East Coast North America Strategic Assessment (OBIS Canada)	Fish	Groundfish	Marine	Temperate	Pelagic / Bottom waters	Used last year sampled with enough records - 1994 (only Fish data)	110	883726	Continental	56	1PLN
18	Reichert, 2009	MARMAP Chevron Trap Survey 1990-2009 (OBIS-USA)	Fish	Fish	Marine	Temperate	Reef	Used last year sampled - 2000	46	15717	Regional	19	1PLN

19	Reichert, 2010a	MARMAP Fly Net 1990-2009 (OBIS-USA)	Fish	Fish	Marine	Temperate	Coastal habitats	Used last year sampled with enough records - 1985	48	10145	Regional	24	1PLN
20	Reichert, 2010b	MARMAP Yankee Trawl 1990-2009 (OBIS-USA)	Fish	Fish	Marine	Temperate	Benthic	Used last year sampled - 1980	147	17839	Regional	59	1PLN
21	Northeast Fisheries Science Center, 2005	Northeast Fisheries Science Center Bottom Trawl Survey Data (OBIS-USA)	Benthos	Benthos	Marine	Temperate	Benthic	Used last year sampled - 2008	401	1467769	Continental	146	1PLN
22	Coral Reef Ecosystem Division, 2011	CRED Rapid Ecological Assessments of Fish Belt Transect Surveys and Fish Stationary Point Count Surveys in the Pacific Ocean 2000-2010 (OBIS-USA)	Fish	Fish	Marine	Tropical / Temperate	Reef	Used last year sampled - 2010	499	496205	Continental	53	Multimodal
23	Marine Resources Research Institute, 2011	Southeast Area Monitoring and Assessment Program (SEAMAP) South Atlantic (OBIS-USA)	All	Fish and Marine invertebrates	Marine	Temperate	Pelagic (shallow waters)	Used last year sampled - 2010	164	684363	Regional	73	1PLN
24	NIWA	South Western Pacific Regional OBIS Data Specify Subset	All	Fish and Marine invertebrates	Marine	Polar / Temperate	Pelagic	Used last year sampled with enough records - 2004	824	28569	Continental	302	1PLN
25	Silveira & Lopes, 2011	Previous fisheries REVIZEE Program (Tropical and Subtropical Western South Atlantic OBIS)	All	Fish and Marine invertebrates	Marine	Tropical	Pelagic / Bottom waters	Used last year sampled with enough records - 1979	165	35409	Continental	107	1PLN
26	Jones & Miller, 2005	Spatial and temporal distribution and abundance of moths in the Andrews Experimental Forest	Terrestrial invertebrates	Macromoths	Terrestrial	Temperate	Coniferous forest	Used last year sampled - 2004	367	13500	Local	16	1PLN
27	Johnson & Farrand, 2014	Aquatic insect sampling in Lookout Creek at the H.J. Andrews Experimental Forest	Terrestrial invertebrates	Arthropoda	Freshwater	Temperate	Streams in coniferous forest	Use all dataset	87	32926	Local	23	1PLN

28	Harmon & Franklin, 2012	Tree growth and mortality measurements in long-term permanent vegetation plots in the Pacific Northwest (LTER Reference Stands)	Terrestrial plants	Trees	Terrestrial	Temperate	Coniferous forest	Used last year sampled - 2009	25	11228	Regional	10	Logser
29	Ballard et al., 2008	PRBO Conservation Science - Point Counts	Birds	Birds	Terrestrial	Temperate	Mixed	Used last year sampled - 2003	268	100038	Regional	48	Multimodal
30	HMANA	Hawk Migration Association of North America (HMANA)	Birds	Raptor birds	Terrestrial	Tropical / Temperate	Mixed	Used last year sampled - 2008	30	4310625	Continental	1	Multimodal
31	NatureCounts b	Ontario Breeding Bird Atlas (2001-2005): point count data	Birds	Birds (breeding)	Terrestrial	Temperate	Mixed	Used last year sampled - 2005	247	262128	Regional	47	Multimodal
32	USFS	Landbird Monitoring Program (UMT-LBMP)	Birds	Landbirds	Terrestrial	Temperate	Mixed	Used last year sampled - 2006	165	25557	Regional	40	Logser
33	NatureCounts a	Maritimes Breeding Bird Atlas (2006-2010): point count data	Birds	Birds (breeding)	Terrestrial	Temperate	Mixed	Used last year sampled - 2010	154	40476	Regional	41	Multimodal
34	Bird Studies Canada, 2012b	Marsh Monitoring Program - Amphibian Surveys	Amphibians	Anuran (frogs and toads)	Terrestrial	Temperate	Wetlands	Used last year sampled - 2011	13	10046	Regional	3	Multimodal
35	Bird Studies Canada, 2012b	Marsh Monitoring Program - Bird Surveys	Birds	Waterbirds	Terrestrial	Temperate	Wetlands	Used last year sampled - 2011	154	29328	Regional	41	1PLN
36	Nilon, 2010	Biodiversity - Fauna - Bird Survey (Table 1 of 4 - Birds)	Birds	Birds (breeding)	Terrestrial	Temperate	Urban / Rural areas	Use all dataset	102	41077	Local	36	1PLN
37	Brush, 2007	Permanent Plot Vegetation Sampling, 2003 Shrub and Vine Data	Terrestrial plants	Shrub and Vine	Terrestrial	Temperate	Urban and non-urban forests	Use all dataset	24	26304	Local	15	1PLN
38	Viereck et al., 2005	Vegetation Plots of the Bonanza Creek LTER Control Plots: Species Count (1975 - 2004)	Terrestrial plants	Boreal vegetation	Terrestrial	Temperate	Taiga	Used last year sampled with enough records - 2003	40	12796	Local	7	Logser
39	Pardieck et al., 2015	Canadian pre-1997 50-stop data for the North American Breeding Bird Survey (BBS)	Birds	Birds (breeding)	Terrestrial	Temperate	Mixed	Used last year sampled - 1996	250	108951	Regional	49	Logser

40	Pardieck et al., 2015	The North American Breeding Bird Survey (BBS)	Birds	Birds (breeding)	Terrestrial	Tropical / Temperate / Polar	Mixed	Used last year sampled - 2011	604	2098071	Continental	69	Multimodal
41	CCE LTER b	Bird and mammal counts for CalCOFI cruises off the west coast of the United States (ID112)	Mammals & Birds	Mammals & Birds	Marine	Temperate	Coastal upwelling habitats	Used last year sampled with enough records - 2006	98	17521	Regional	16	Logser
42	CCE LTER a	Bird and mammal counts for National Marine Fisheries cruises as part of the Rockfish Recruitment Survey (ID117)	Mammals & Birds	Mammals & Birds	Marine	Temperate	Coastal upwelling habitats	Used last year sampled - 2006	77	16781	Regional	12	1PLN
43	CCE LTER c	Bird and mammal counts for continuous plankton recorder cruises on the great circle route from Vancouver, BC to Tokyo (ID118)	Mammals & Birds	Mammals & Birds	Marine	Temperate	Open oceanic habitats	Used last year sampled - 2006	98	247679	Continental	15	1PLN
44	Reich et al.; Cavender-Bares & Reich, 2012	Effect of Burning Patterns on Vegetation in the Fish Lake Burn Compartments (Experiment 133 - Shrub Survey)	Terrestrial plants	Shrub	Terrestrial	Temperate	Savanna/Tallgrass prairie	Used last year sampled - 2005	39	22011	Local	19	Logser
45	Grimm et al., 2005	Survey 200 long term study of multiple sites in central Arizona-Phoenix (27 individual count 1)	Terrestrial plants	Desert vegetation	Terrestrial	Temperate	Desert	Used last year sampled - 2005	139	22940	Local	36	Multimodal
46	Walker et al., 2004	Point Count Bird Censusing Data Subset for Paper 'Effects of land use and vegetation cover on bird communities' Walker et. al (127_birds_2003_1)	Birds	Birds	Terrestrial	Temperate	Urban / Desert / Riparian / Agricultural	Use all dataset	128	22582	Local	40	1PLN
47	Warren et al., 2005	Ecological and Social Interactions in Urban Parks: Bird surveys in local parks in the CAP-LTER study area (52_pp52_birds_1)	Birds	Birds	Terrestrial	Temperate	Urban	Used last year sampled - 2002	78	44444	Local	31	1PLN

48	Shochat et al., 2004	Point count bird censusing: long-term monitoring of bird distribution and diversity in central Arizona-Phoenix: period 2000 to 2011 (34_birds_1)	Birds	Birds	Terrestrial	Temperate	Urban / Desert / Riparian / Agricultural	Used last year sampled - 2011	133	15212	Local	41	1PLN
49	Ohmart et al., 2003	Transect bird survey with data synthesis from multiple transects in the central Arizona-Phoenix area: period 1998 to 2000 (12_birds_2000_1)	Birds	Birds	Terrestrial	Temperate	Urban / Desert	Used last year sampled - 2000	133	154334	Local	38	1PLN
50	Hale et al., 2002	Coastal ecological data from the Virginian Biogeographic Province, 1990-1993 - benthic_species_abun_data	Benthos	Benthos	Marine	Temperate	Estuarine and coastal waters	Used last year sampled - 1993	514	101118	Regional	168	1PLN
51	Hale et al., 2002	Coastal ecological data from the Virginian Biogeographic Province, 1990-1993 - fish_species_abun_data	Fish	Demersal Fish	Marine	Temperate	Estuarine and coastal waters	Used last year sampled - 1993	82	13007	Regional	47	1PLN
52	McLarney et al., 2010	Upper Little Tennessee River Biomonitoring Program Database - LTWA Biomonitoring Database	Fish	Fish	Freshwater	Temperate	Streams in deciduous forest	Used last year sampled - 2011	36	12148	Local	8	1PLN
53	Gaiser, 2010	Macrophyte count data collected from Northeast Shark Slough, Everglades National Park (FCE) from September 2006 to Present	Macrophytes	Macrophytes	Marine	Temperate	Estuarine	Used last year sampled - 2008	42	19352	Local	21	Logser
54	Trexler, 2007	Consumer Stocks: Fish, Vegetation, and other Non-physical Data from Everglades National Park, South Florida	Fish	Fish	Marine	Temperate	Estuarine	Used last year sampled with enough records - 2003 (only Fish data)	33	10481	Local	12	Logser
55	Ramesh et al., 2010	Forest stand structure and composition in 96 sites along environmental gradients in the central Western Ghats of India (macroplots)	Terrestrial plants	Woody plants	Terrestrial	Tropical	Mixed	Use data from macroplot only	399	61965	Regional	73	1PLN

56	Ramesh et al., 2010	Forest stand structure and composition in 96 sites along environmental gradients in the central Western Ghats of India (microplots)	Terrestrial plants	Woody plants	Terrestrial	Tropical	Mixed	Use data from microplot only	334	14848	Regional	71	Logser
57	Hosley, 2003	HF039. Vegetation Inventory of Harvard Forest 1937	Terrestrial plants	Trees	Terrestrial	Temperate	Deciduous forest	Use all dataset	32	18173	Local	12	1PLN
58	Gould & Foster, 2000	HF037. Vegetation Inventory of Harvard Forest 1986-1993	Terrestrial plants	Trees	Terrestrial	Temperate	Deciduous forest	Use all dataset	55	30282	Local	12	Logser
59	Sipe et al., 2009	HF141. Tree Seedlings in CRUI Land Use Project at Harvard Forest 1996	Terrestrial plants	Trees (seedlings)	Terrestrial	Temperate	Deciduous forest & Agricultural	Use all dataset	18	14801	Local	8	Logser
60	Foster et al., 2006a	HF078. Influence of Little Ice Age on New England Vegetation from 2000 BP to Present	Chromista	Diatoms	Freshwater	Temperate	Ponds in deciduous forests	Use all dataset	430	102079	Regional	21	Multimodal
61	Foster & Motzkin, 2003	HF015. Land Use and Forest Dynamics at Harvard Forest 1937-1995	Terrestrial plants	Trees	Terrestrial	Temperate	Deciduous forest	Use all dataset	38	26154	Local	14	Logser
62	Foster et al., 2006b	HF044. Land Use on the Southern New England and New York Coasts 1600-2001	Terrestrial plants	Trees	Terrestrial	Temperate	Deciduous forest	Used last year sampled with enough records - 1999	33	13511	Regional	14	Logser
63	Ellison & Gotelli, 2009	HF147. Ant Distribution and Abundance in New England since 1990: bog ants 1999	Terrestrial invertebrates	Ants	Terrestrial	Temperate	Deciduous forest and bogs	Use all dataset	40	19013	Regional	1	Logser
64	Ellison & Gotelli, 2009	HF147. Ant Distribution and Abundance in New England since 1990: bog vegetation 1999	Terrestrial plants	Bog vegetation	Terrestrial	Temperate	Deciduous forest and bogs	Use all dataset	91	64105	Regional	47	1PLN
65	Kittredge et al., 2009	HF127. Timber Harvesting Field Study in Western Massachusetts 2004-2005: stand trees	Terrestrial plants	Trees	Terrestrial	Temperate	Deciduous forest	Use all dataset	49	9946	Regional	16	Logser
66	Kittredge et al., 2009	HF127. Timber Harvesting Field Study in Western Massachusetts 2004-2005: stand saplings	Terrestrial plants	Trees (saplings)	Terrestrial	Temperate	Deciduous forest	Use all dataset	51	14857	Regional	15	Logser

67	Kittredge et al., 2009	HF127. Timber Harvesting Field Study in Western Massachusetts 2004-2005: stand seedlings	Terrestrial plants	Trees (seedlings)	Terrestrial	Temperate	Deciduous forest	Use all dataset	49	18909	Regional	16	Logser
68	Battles et al., 2003b	Forest Inventory of a Northern Hardwood Forest: Watershed 6 2002, Hubbard Brook Experimental Forest	Terrestrial plants	Trees	Terrestrial	Temperate	Deciduous forest	Use all dataset	15	12373	Local	6	Logser
69	Battles et al., 2003a	Forest Inventory of a Northern Hardwood Forest: Watershed 5 (before the whole-tree harvest) 1982	Terrestrial plants	Trees	Terrestrial	Temperate	Deciduous forest	Use all dataset	14	18275	Local	6	1PLN
70	Driscoll et al., 2003	Forest Inventory of a Northern Hardwood Forest: Watershed 1 (before the calcium addition) 1996	Terrestrial plants	Trees	Terrestrial	Temperate	Deciduous forest	Use all dataset	14	10454	Local	7	Logser
71	Battles & Fahey	Tree inventory data for the Hubbard Brook Valley Plots	Terrestrial plants	Trees	Terrestrial	Temperate	Deciduous forest	Use all dataset	21	14964	Local	8	Logser
72	Gido	Fish population on selected watersheds at Konza Prairie - CFP012 - Konza fish population	Fish	Fish	Freshwater	Temperate	Pool / riffle in tallgrass prairie	Used last year sampled with enough records - 1997	11	19442	Local	4	Logser
73	Woods, 2009	Multi-decade, spatially explicit population studies of canopy dynamics in Michigan old-growth forests: all_plots_1974-1980	Terrestrial plants	Woody stems	Terrestrial	Temperate	Northern hardwood forest (old-growth forests)	Use all dataset	21	28986	Local	11	1PLN
74	Woods, 2009	Multi-decade, spatially explicit population studies of canopy dynamics in Michigan old-growth forests: upland_plots_89-07	Terrestrial plants	Woody stems	Terrestrial	Temperate	Northern hardwood forest (dominated by deciduous trees)	Use all dataset	23	13624	Local	9	Logser
75	Woods, 2009	Multi-decade, spatially explicit population studies of canopy dynamics in Michigan old-growth forests: swamp_all_modern	Terrestrial plants	Woody stems	Terrestrial	Temperate	Northern hardwood forest (swamp forests)	Use all dataset	25	10764	Local	12	Logser

76	Zimmerman & Brokaw	Census of species, diameter and location at the Luquillo Forest Dynamics Plot (LFDP), Puerto Rico - LFDP census 3 (Part1 & part2)	Terrestrial plants	Woody stems	Terrestrial	Tropical	Tropical forest	Use all dataset	149	93703	Local	45	Logser
77	Zimmerman	LFDP phenology plot seedlings – 16 ha plot - LFDP Phenology Seedlings Data for 2011	Terrestrial plants	Woody stems (seedlings)	Terrestrial	Tropical	Tropical forest	Used last year sampled - 2011	71	16238	Local	22	1PLN
78	Carpenter et al., 2006a	Biocomplexity at North Temperate Lakes LTER; Coordinated Field Studies: Fish / Crayfish Abundance 2001 - 2004	Fish / Crayfish	Fish and Crayfish	Freshwater	Temperate	Temperate lake	Used last year sampled with enough records - 2002	34	10165	Regional	11	Logser
79	Lathrop, 2000	Madison Wisconsin Lakes Zooplankton 1976 - 1994 (old net)	Freshwater invertebrates	Zooplankton	Freshwater	Temperate	Temperate lake	Used last year for which 4 lakes were sampled - 1985	55	136545466	Local	19	1PLN
80	Lathrop, 2000	Madison Wisconsin Lakes Zooplankton 1976 - 1994 (new net)	Freshwater invertebrates	Zooplankton	Freshwater	Temperate	Temperate lake	Used last year sampled - 1994	15	11896183	Local	6	1PLN
81	NLT LTER, 2012	North Temperate Lakes LTER: Fish Abundance 1981 - current	Fish	Fish	Freshwater	Temperate	Temperate lake	Used last year sampled - 2012	45	14084	Regional	17	Logser
82	NLT LTER, 2011	North Temperate Lakes LTER: Zooplankton - Madison Lakes Area 1997 - current	Freshwater invertebrates	Zooplankton	Freshwater	Temperate	Temperate lake	Used last year sampled - 2010	16	82123230	Local	8	1PLN
83	Dillon et al., 2007	North Temperate Lakes LTER: Snail Survey in Northern Wisconsin Lakes 2006	Terrestrial invertebrates	Land Snails	Freshwater	Temperate	Temperate lake	Use all dataset	21	17772	Local	7	1PLN
84	Carpenter et al., 2006b	Biocomplexity at North Temperate Lakes LTER; Coordinated Field Studies: Riparian Plots 2001 - 2004 - Live Tree Counts	Terrestrial plants	Woody plants	Terrestrial	Temperate	Riparian habitat	Use all dataset	42	18300	Regional	11	Logser

85	Ernest et al., 2009	Long-term monitoring and experimental manipulation of a Chihuahuan Desert ecosystem near Portal, Arizona, USA: Portal_plant_summer_annual_19892002	Terrestrial plants	Desert herbaceous plants	Terrestrial	Temperate	Desert	Used last year sampled - 2002	35	11513	Local	14	1PLN
86	Ernest et al., 2009	Long-term monitoring and experimental manipulation of a Chihuahuan Desert ecosystem near Portal, Arizona, USA: Portal_plant_summer_annual_19892002 & Portal_plant_summer_perennial_19892002	Terrestrial plants	Desert herbaceous plants	Terrestrial	Temperate	Desert	Used last year sampled - 2002	69	16211	Local	21	Logser
87	Ernest et al., 2009	Long-term monitoring and experimental manipulation of a Chihuahuan Desert ecosystem near Portal, Arizona, USA: Portal_plant_winter_annual_19892002	Terrestrial plants	Desert herbaceous plants	Terrestrial	Temperate	Desert	Used last year sampled with enough records - 2001	36	16412	Local	14	1PLN
88	Ernest et al., 2009	Long-term monitoring and experimental manipulation of a Chihuahuan Desert ecosystem near Portal, Arizona, USA: Portal_plant_winter_annual_19892002 & Portal_plant_winter_perennial_19892002	Terrestrial plants	Desert herbaceous plants	Terrestrial	Temperate	Desert	Used last year sampled with enough records - 2001	57	16649	Local	20	1PLN
89	Ernest et al., 2009	Long-term monitoring and experimental manipulation of a Chihuahuan Desert ecosystem near Portal, Arizona, USA: Portal_ant_bait_19882002	Terrestrial invertebrates	Ants	Terrestrial	Temperate	Desert	Used last year sampled with enough records - 1991	15	10397	Local	1	Logser

90	Muldavin a	Pinon Juniper Net Primary Production Quadrat Data from the Sevilleta National Wildlife Refuge, New Mexico: 1999-2001_juniper_savannah_woodland	Terrestrial plants	Woodland vegetation	Terrestrial	Temperate	Desert/Grassland	Used last year sampled - 2001; only data for juniper savannah woodland (J)	83	11293	Local	25	1PLN
91	Muldavin b	Pinon-Juniper (Core Site) Quadrat Data for the Net Primary Production Study at the Sevilleta National Wildlife Refuge, New Mexico (2003-)	Terrestrial plants	Woodland vegetation	Terrestrial	Temperate	Desert/Grassland	Used last year sampled with enough records - 2008	84	16012	Local	26	1PLN
92	Lauenroth, 2013	SGS-LTER Disturbance intensity and above- and belowground herbivory effects on long-term recovery of shortgrass steppe on the Central Plains Experimental Range, Nunn, Colorado, USA 1977-1990	Terrestrial plants	Shortgrass steppe vegetation	Terrestrial	Temperate	Shortgrass steppe	Used last year sampled - 1990; only data for treatment 'Ungrazed'	24	42322	Local	12	Multimodal
93	Thomas et al., 2003; Chuyong et al., 2004; Kenfack et al., 2007	Korup Forest Dynamics Plot, Cameroon	Terrestrial plants	Woody plants	Terrestrial	Tropical	Tropical forest	Used last year sampled - 1998; data for '>10cm'	282	24591	Local	48	1PLN
94	Sukumar; Sukumar et al., 2004	Mudumalai Forest Dynamics Plot, India	Terrestrial plants	Woody plants	Terrestrial	Tropical	Tropical forest	Used last year sampled - 2000; data for '>10cm'	61	12574	Local	24	1PLN
95	Tan et al.; Lee et al., 2002, 2005	Lambir Hills Forest Dynamics Plot, Malaysia	Terrestrial plants	Woody plants	Terrestrial	Tropical	Tropical forest	Used last year sampled - 1997; data for '>10cm'	984	32350	Local	76	Multimodal
96	Fletcher & Kassim; Manokaran et al., 2004	Pasoh Forest Dynamics Plot, Malaysia	Terrestrial plants	Woody plants	Terrestrial	Tropical	Tropical forest	Used last year sampled - 2000; data for '>10cm'	671	28279	Local	69	Multimodal
97	Gunatilleke & Gunatilleke; Gunatilleke et al., 2004	Sinharaja Forest Dynamics Plot, Sri Lanka	Terrestrial plants	Woody plants	Terrestrial	Tropical	Tropical forest	Used last year sampled - 2001; data for '>10cm'	171	16776	Local	41	Logser

98	Sun & Hsieh; Su et al., 2007	Fushan Forest Dynamics Plot, Taiwan	Terrestrial plants	Woody plants	Terrestrial	Temperate	Tropical forest	Used last year sampled - 2002; data for '>10cm'	77	19270	Local	32	Logser
99	Bunyavejchewin; Bunyavejchewin et al., 1998, 2001, 2009	Huai Kha Khaeng Forest Dynamics Plot, Thailand	Terrestrial plants	Woody plants	Terrestrial	Tropical	Tropical forest	Used last year sampled - 1999; data for '>10cm'	241	21874	Local	55	Multimodal
100	Brockelman & Nathalang	Mo Singto Forest Dynamics Plot, Thailand	Terrestrial plants	Woody plants	Terrestrial	Tropical	Tropical forest	Used last year sampled - 2004; data for '>10cm'	262	131009	Local	67	Logser
101	Oliveira	Ilha do Cardoso, Brasil	Terrestrial plants	Woody plants	Terrestrial	Temperate	Tropical forest	Used last year sampled - 2005	116	15040	Local	46	Multimodal
102	Alvarez; Vallejo et al., 2004	La Planada Forest Dynamics Plot, Colombia	Terrestrial plants	Woody plants	Terrestrial	Tropical	Tropical forest	Used last year sampled - 2003; data for '>10cm'	173	15013	Local	50	Multimodal
103	Valencia; Valencia et al., 2004	Yasuni Forest Dynamics Plot, Ecuador	Terrestrial plants	Woody plants	Terrestrial	Tropical	Tropical forest	Used last year sampled - 2003; data for '>10cm'	817	17428	Local	65	1PLN
104	Condit, 1998; Hubbell et al., 1999, 2010	Barro Colorado Island Forest Dynamics Plot, Panama	Terrestrial plants	Woody plants	Terrestrial	Tropical	Tropical forest	Used last year sampled - 2005; data for '>10cm'	227	20848	Local	50	1PLN
105	Condit	Sherman Forest Dynamics Plot, Panama - Abundance of all tree species in the entire plot, 1996-1998 (saplings and trees)	Terrestrial plants	Woody plants	Terrestrial	Tropical	Tropical forest	Used last year sampled - 1999	228	21911	Local	56	Logser
106	Zimmerman et al.; Zimmerman et al., 2010	Luquillo Forest Dynamics Plot, Puerto Rico	Terrestrial plants	Woody plants	Terrestrial	Tropical	Tropical forest	Used last year sampled - 1995; data for '>10cm'	86	14001	Local	37	Logser
107	Paquette et al., 2007	Lac Croche understory vegetation data set	Terrestrial plants	Understory vegetation	Terrestrial	Temperate	Northern temperate forest	Used last year sampled - 2006	12	138160	Local	6	1PLN

108	Day et al., 2004; Day, 2010	Long-term N-fertilized vegetation plots on Hog Island, Virginia Coastal Barrier Islands	Terrestrial plants	Dune vegetation	Marine	Temperate	Barrier Island	Used last year sampled with enough records - 2005; only data for treatment 'Control'	25	17194	Local	12	Multimodal
109	Beck, 1996	Nesting seabird census of Hog Island and Cobb Island of the Virginia Coast Reserve 1991	Birds	Seabirds	Marine	Temperate	Barrier Island	Use all dataset	104	83055	Local	25	Logser
110	Balslev	Aarhus University Palm Transect Database	Terrestrial plants	Woody plants (Arecaceae - The palm family)	Terrestrial	Tropical	Tropical forest	Used last year sampled - 2011	135	97129	Regional	1	1PLN
111	Bird Studies Canada, 2012a	BC Coastal Waterbird Survey	Birds	Coastal waterbirds	Marine	Temperate	Coastal habitats	Used last year sampled - 2012 - data retrieved from Nature Counts	124	547893	Regional	19	Logser
112	Stevens, 2010	Trekvis - Migratory fishes in the river Scheldt	Fish	Diadromous fish	Marine	Temperate	Estuarine	Used last year sampled with enough records - 2007	51	20338	Regional	27	Logser
113	The Swedish University of Agriculture Sciences	National Forest Inventory (SLU)	Terrestrial plants	Terrestrial plants	Terrestrial	Polar / Temperate	Temperate forest	Used last year sampled - 1999	228	42220	Regional	71	1PLN
114	Williams, 1999	Pelagic Fish Observations 1968-1999	Fish	Fish	Marine	Polar / Temperate	Pelagic	Used last year sampled with enough records - 1993	77	12407	Continental	22	1PLN
115	de Abreu et al., 2003	Planktic foraminifera counts of sediment core MD95-2040	Chromista	Foraminifera	Marine	Temperate	Ocean sediments	Use all dataset	29	480230	Local	3	Logser
116	USDA Forest Service, 2007	USDA Forest Service, Redwood Sciences Laboratory - Lamna Point Count	Birds	Landbirds	Terrestrial	Temperate	Mixed	Used last year sampled with enough records - 2005	108	13686	Regional	37	1PLN
117	Sarnthein et al., 2003	Distribution of foraminifera of sediment core GIK23258-2	Chromista	Foraminifera	Marine	Polar	Ocean sediments	Use all dataset	12	276201	Local	3	Logser

Appendix II

References for Data Sources for the Multimodality analysis

de Abreu, L., Shackleton, N.J., Schönfeld, J., Hall, M. & Chapman, M. (2003) “Planktic foraminifera counts of sediment core MD95-2040”. doi:10.1594/PANGAEA.66714, In Supplement to: de Abreu *et al.* (2003): Millennial-scale oceanic climate variability off the Western Iberian margin during the last two glacial periods. *Marine Geology*, 196(1-2), 1-20, doi:10.1016/S0025-3227(03)00046-X. *PANGAEA*. Available at: <http://www.gbif.org/dataset/662510f4-f762-11e1-a439-00145eb45e9a>, accessed 2012.

Alvarez, M. “La Planada Forest Dynamics Plot, Colombia.” *The Center for Tropical Forest Science. Smithsonian Tropical Research Institute*. Available at: <http://www.ctfs.si.edu/site/La+Planada/>, accessed 2013.

Bakker, C., Herman, P. & Vink, M. (1994) A new trend in the development of the phytoplankton in the Oosterschelde (SW Netherlands) during and after the construction of a storm-surge barrier. *Hydrobiologia*, 282-283, 79–100.

Bakker, K. & Herman, P. (1990) “Phytoplankton in the Oosterschelde before, during and after the storm-surge barrier (1982-1990).” *Netherlands Institute of Ecology, Centre for Estuarine and Marine Ecology, Netherlands*. Available at: <http://www.emodnet-biology.eu/data-catalog?module=dataset&david=1646>, accessed 2013.

Ballard, G., Herzog, M., Fitzgibbon, M., Moody, D., Jongsomjit, D. & Stralberg, D. (2008) “PRBO Conservation Science - Point Counts.” *The California Avian Data Center. Petaluma, California*. Available at: www.prbo.org/cadc, accessed 2012.

Balslev, H. “Aarhus University Palm Transect Database.” *Department of Bioscience, Aarhus University*. Available at: <http://www.gbif.org/dataset/a9e763c8-f674-4492-94a8-4fd4eb9342a5>, accessed 2013.

Battles, J.J., Fahey, T. & Cleavitt, N. (2003a) “Forest Inventory of a Whole Tree Harvest: Hubbard Brook Experimental Forest Watershed 5, 1982, pre-harvest.” *The Hubbard Brook Ecosystem Study LTER Program*. Available at: <http://www.hubbardbrook.org/data/dataset.php?id=36>, accessed 2012.

Battles, J.J. & Fahey, T.J. “Tree inventory data for the Hubbard Brook Valley Plots, baseline data collected 1995 - 1998.” *The Hubbard Brook Ecosystem Study LTER Program*. Available at: <http://www.hubbardbrook.org/data/dataset.php?id=125>, accessed 2012.

Battles, J.J., Johnson, C., Hamburg, S., Fahey, T., Driscoll, C. & Likens, G. (2003b) “Forest Inventory of a Northern Hardwood Forest: Watershed 6 2002.” The Hubbard Brook Ecosystem Study LTER Program. Available at: <http://www.hubbardbrook.org/data/dataset.php?id=35>, accessed 2012.

Beck, R. (1996) “Nesting seabird census of Hog Island and Cobb Island of the Virginia Coast Reserve 1991.” *Virginia Coast Reserve Long-Term Ecological Research Project*. Available at: <http://www.vcrlter.virginia.edu/cgi-bin/showDataset.cgi?docid=knb-lter-vcr.18>, accessed 2013.

Bird Studies Canada (2012a) “BC Coastal Waterbird Survey (2004).” *NatureCounts, a node of the Avian Knowledge Network*. Available at: <http://www.birdscanada.org/birdmon/>, accessed 2012.

Bird Studies Canada (2012b) “Marsh Monitoring Program.” *NatureCounts, a node of the Avian Knowledge Network*. Available at: <http://www.birdscanada.org/birdmon/>, accessed 2012.

Brockelman, W.Y. & Nathalang, A. “Mo Singto Forest Dynamics Plot, Thailand.” *The Center for Tropical Forest Science. Smithsonian Tropical Research Institute*. Available at: <http://www.ctfs.si.edu/site/Mo+Singto/>, accessed 2013.

Brown, S.K.R., Zwanenburg, K. & Branton, R. (2005) “East Coast North America Strategic Assessment Groundfish Atlas - ECNASAP.” *OBIS Canada, Bedford Institute of Oceanography, Dartmouth, Nova Scotia, Canada*. Available at: <http://iobis.org/>, accessed 2013.

Brush, G. (2007) “Permanent Plot Vegetation Sampling, 2003 Shrub and Vine Data.” *Baltimore Ecosystem Study LTER Program. Baltimore, MD, USA*. Available at: http://www.beslter.org/metacat_harvest_attribute_level_eml/html_metadata/bes_414.asp, accessed 2012.

Bunyavejchewin, S. “Huai Kha Khaeng Forest Dynamics Plot, Thailand.” *The Center for Tropical Forest Science. Smithsonian Tropical Research Institute*. Available at: <http://www.ctfs.si.edu/site/Huai+Kha+Khaeng/>, accessed 2013.

Bunyavejchewin, S., Baker, P.J., LaFrankie, J. V. & Ashton, P.S. (2001) Stand structure of a seasonal dry evergreen forest at Huai Kha Khaeng Wildlife Sanctuary, western Thailand. *Natural History Bulletin of the Siam Society*, 49, 89–106.

Bunyavejchewin, S., LaFrankie, J. V., Baker, P.J., Davies, S.J. & Ashton, P.S. (2009) Forest trees of Huai Kha Khaeng Wildlife Sanctuary, Thailand: Data from the 50-hectare Forest Dynamic Plot. *The National Parks, Wildlife and Plant Conservation Department*.

Bunyavejchewin, S., LaFrankie, J. V., Pattapong, P., Kanzaki, M., Itoh, A., Yamakura, T. & Ashton, P.S. (1998) Topographic analysis of a large-scale research plot in seasonal dry evergreen forest at Huai Kha Khaeng Wildlife Sanctuary, Thailand. *Tropics*, 8, 45–60.

Carpenter, S., Kitchell, J., Kratz, T. & Magnuson, J. (2006a) “Biocomplexity at North Temperate Lakes LTER; Coordinated Field Studies: Fish / Crayfish Abundance 2001 - 2004.” *North Temperate Lakes Long Term Ecological Research Program, Center for Limnology, University of Wisconsin-Madison*. Available at: <http://lter.limnology.wisc.edu/dataset/biocomplexity-north-temperate-lakes-lter-coordinated-field-studies-fish-crayfish-abundance-2>, accessed 2013.

Carpenter, S., Kratz, T., Cronon, W., Provencher, R. & Turner, M. (2006b) “Biocomplexity at North Temperate Lakes LTER; Coordinated Field Studies: Riparian Plots 2001 - 2004.” *North Temperate Lakes Long Term Ecological Research Program, Center for Limnology, University of Wisconsin-Madison*. Available at: <http://lter.limnology.wisc.edu/dataset/biocomplexity-north-temperate-lakes-lter-coordinated-field-studies-riparian-plots-2001-2004>, accessed 2013.

Cavender-Bares, J. & Reich, P.B. (2012) Shocks to the system: community assembly of the oak savanna in a 40-year fire frequency experiment. *Ecology*, 93, S52–S69.

CCE LTER (a) "Bird and mammal counts for National Marine Fisheries cruises as part of the Rockfish Recruitment Survey (ID117). *CalCOFI - Scripps Institution of Oceanography. California Current Ecosystem (CCE) Long Term Ecological Research (LTER)*. Available at: <http://oceaninformatics.ucsd.edu/datazoo/data/ccelter/datasets?action=summary&id=117>, accessed 2012.

CCE LTER (b) “Bird and mammal counts for CalCOFI cruises off the west coast of the United States (ID112).” *CalCOFI - Scripps Institution of Oceanography. California Current Ecosystem (CCE) Long Term Ecological Research (LTER)*. Available at: <http://oceaninformatics.ucsd.edu/datazoo/data/ccelter/datasets?action=summary&id=112>, accessed 2012.

CCE LTER (c) “Bird and mammal counts for continuous plankton recorder cruises on the great circle route from Vancouver, BC to Tokyo (ID118).” *CalCOFI - Scripps Institution of Oceanography. California Current Ecosystem (CCE) Long Term Ecological Research (LTER)*. Available at:

<http://oceaninformatics.ucsd.edu/datazoo/data/ccelter/datasets?action=summary&id=118>, accessed 2012.

Chuyong, G.B., Condit, R., Kenfack, D., Losos, E., Sainge, M., Songwe, N.C. & Thomas, D.W. (2004) *Korup Forest Dynamics Plot, Cameroon. Forest Diversity and Dynamism: Findings from a Large-Scale Plot Network* (ed. by E.C. Losos and E.G.J. Leigh), pp. 506–516. University of Chicago Press, Chicago.

Clark, D. & Branton, B. (2007) “DFO Maritimes Research Vessel Trawl Surveys Fish Observations.” *OBIS Canada Digital Collections. OBIS Canada, Bedford Institute of Oceanography, Dartmouth, Nova Scotia, Canada*. Available at: <http://iobis.org/>, accessed 2013.

Condit, R. “Sherman Forest Dynamics Plot, Panama.” *The Center for Tropical Forest Science. Smithsonian Tropical Research Institute*. Available at: <http://www.ctfs.si.edu/site/Sherman/>, accessed 2013.

Condit, R. (1998) *Tropical forest census plots*, Springer-Verlag and R. G. Landes Company, Berlin, Germany and Georgetown, Texas.

Coral Reef Ecosystem Division (2011) “CRED Rapid Ecological Assessments of Fish Belt Transect Surveys and Fish Stationary Point Count Surveys in the Pacific Ocean 2000-2010.” *Coral Reef Ecosystem Division (CRED), Pacific Island Fisheries Sciences Center, NOAA National Marine Fisheries Service. Coral Reef Ecosystem Division, Honolulu, HI*. Available at: <http://www.usgs.gov/obis-usa/>, accessed 2013.

CSIRO “CSIRO Marine Data Warehouse - OBIS Australia.” *CSIRO Division of Marine and Atmospheric Research (CMAR), Australia*. Available at: <http://iobis.org/>, accessed 2013.

DATRAS (2010a) “Fish trawl survey: ICES Beam Trawl Survey for commercial fish species. ICES Database of trawl surveys (DATRAS).” *The International Council for the Exploration of the Sea, Copenhagen*. Available at: <http://www.emodnet-biology.eu/data-catalog/?%3Fmodule=dataset&daside=2761>, accessed 2013.

DATRAS (2010b) “Fish trawl survey: ICES French Southern Atlantic Bottom Trawl Survey for commercial fish species. ICES Database of trawl surveys (DATRAS).” *The International Council for the Exploration of the Sea, Copenhagen*. Available at: <http://www.emodnet-biology.eu/data-catalog/?%3Fmodule=dataset&daside=2759>, accessed 2013.

DATRAS (2010c) “Fish trawl survey: ICES North Sea International Bottom Trawl Survey for commercial fish species. ICES Database of trawl surveys (DATRAS).” *The International Council for the Exploration of the Sea, Copenhagen*. Available at: <http://www.emodnet-biology.eu/data-catalog?module=dataset&dasid=2763>, accessed 2013.

DATRAS (2010d) “Fish trawl survey: Irish Ground Fish Survey for commercial fish species. ICES Database of trawl surveys (DATRAS).” *The International Council for the Exploration of the Sea, Copenhagen*. Available at: <http://www.emodnet-biology.eu/data-catalog?module=dataset&dasid=2762>, accessed 2013.

DATRAS (2010e) “Fish trawl survey: Northern Irish Ground Fish Trawl Survey. ICES Database of trawl surveys (DATRAS).” *The International Council for the Exploration of the Sea, Copenhagen*. Available at: <http://www.emodnet-biology.eu/data-catalog?module=dataset&dasid=2764>, accessed 2013.

DATRAS (2010f) “Fish trawl survey: Scottish Rockall Survey for commercial fish species. ICES Database of trawl surveys (DATRAS).” *The International Council for the Exploration of the Sea, Copenhagen*. Available at: <http://www.emodnet-biology.eu/data-catalog?module=dataset&dasid=2767>, accessed 2013.

DATRAS (2010g) “Fish trawl survey: Scottish West Coast Survey for commercial fish species. ICES Database of trawl surveys (DATRAS).” *The International Council for the Exploration of the Sea, Copenhagen*. Available at: <http://www.emodnet-biology.eu/data-catalog?module=dataset&dasid=2766>, accessed 2013.

Day, F. (2010) “Long-term N-fertilized vegetation plots on Hog Island, Virginia Coastal Barrier Islands, 1992-2014.” *Virginia Coast Reserve Long-Term Ecological Research Project*. Available at: <http://www.vcrlter.virginia.edu/cgi-bin/showDataset.cgi?docid=knb-lter-vcr.106>, accessed 2013.

Day, F.P., Conn, C., Crawford, E. & Stevenson, M. (2004) Long-term effects of nitrogen fertilization on plant community structure on a coastal barrier island dune chronosequence. *Journal of Coastal Research*, 20, 722–730.

Degraer, S., Wittoeck, J., Appeltans, W., Cooreman, K., Deprez, T., Hillewaert, H., Hostens, K., Mees, J., Vanden Berghe, E. & Vincx, M. (2006) “Macrobenthos: Long term trends in the macrobenthos of the Belgian Continental Shelf.” *Oostende, Belgium*. Available at: <http://www.emodnet-biology.eu/data-catalog?module=dataset&dasid=145>, accessed 2013.

Dillon, R., Johnson, P., Olden, J., Solomon, C. & Zanden, J. Vander (2007) "North Temperate Lakes LTER: Snail Survey in Northern Wisconsin Lakes 2006." *North Temperate Lakes Long Term Ecological Research Program, Center for Limnology, University of Wisconsin-Madison*. Available at: <http://lter.limnology.wisc.edu/dataset/north-temperate-lakes-lter-snail-survey-northern-wisconsin-lakes-2006>, accessed 2013.

Driscoll, C., Bailey, S., Blum, J., Buso, D., Eagar, C., Fahey, T., Fisk, M., Groffman, P., Johnson, C., Likens, G., Hamburg, S. & Siccama, T.G. (2003) "Forest Inventory of a Calcium Amended Northern Hardwood Forest: Watershed 1, 1996." *The Hubbard Brook Ecosystem Study LTER Program*. Available at: <http://www.hubbardbrook.org/data/dataset.php?id=40>, accessed 2012.

Ellison, A. & Gotelli, N. (2009) "Ant Distribution and Abundance in New England since 1990. Harvard Forest Data Archive: HF147." *The Harvard Forest Long Term Ecological Research Program*. Available at: <http://harvardforest.fas.harvard.edu:8080/exist/xquery/data.xq?id=hf147>, accessed 2013.

Ernest, S.K.M., Valone, T.J. & Brown, J.H. (2009) Long-term monitoring and experimental manipulation of a Chihuahuan Desert ecosystem near Portal, Arizona, USA. *Ecology*, 90, 1708.

Fletcher, C. & Kassim, A.R. "Pasoh Forest Dynamics Plot, Malaysia." *The Center for Tropical Forest Science. Smithsonian Tropical Research Institute*. Available at: <http://www.ctfs.si.edu/site/Pasoh/>, accessed 2013.

Foster, D., Francis, D. & Fuller, J. (2006a) "Influence of Little Ice Age on New England Vegetation from 2000 BP to Present. Harvard Forest Data Archive: HF078." *The Harvard Forest Long Term Ecological Research Program*. Available at: <http://harvardforest.fas.harvard.edu:8080/exist/xquery/data.xq?id=hf078>, accessed 2013.

Foster, D., Holle, B. Von & Parshall, T. (2006b) "Land Use on the Southern New England and New York Coasts 1600-2001. Harvard Forest Data Archive: HF044." *The Harvard Forest Long Term Ecological Research Program*. Available at: <http://harvardforest.fas.harvard.edu:8080/exist/xquery/data.xq?id=hf044>, accessed 2013.

Foster, D. & Motzkin, G. (2003) "Land Use and Forest Dynamics at Harvard Forest 1937-1995. Harvard Forest Data Archive: HF015." *The Harvard Forest Long Term Ecological Research Program*. Available at: <http://harvardforest.fas.harvard.edu:8080/exist/xquery/data.xq?id=hf015>, accessed 2013.

- Gaiser, E. (2010) “Macrophyte count data collected from Northeast Shark Slough, Everglades National Park (FCE) from September 2006 to Present”. <http://dx.doi.org/10.6073/pasta/effd9e98134913af21b670febebd6233>. *Florida Coastal Everglades LTER Program*. Available at: http://fcelter.fiu.edu/data/core/metadata/EML/?datasetid=LT_PP_Gaiser_001, accessed 2012.
- Gido, K.B. “Fish population on selected watersheds at Konza Prairie - CFP01.” *Konza Prairie LTER Program*. Available at: <http://www.konza.ksu.edu/KNZ/pages/data/Knzdsdetail.aspx?datasetCode=CFP01>, accessed 2012.
- Gould, E. & Foster, D. (2000) “Vegetation Inventory of Harvard Forest 1986-1993. Harvard Forest Data Archive: HF037.” *The Harvard Forest Long Term Ecological Research Program*. Available at: <http://harvardforest.fas.harvard.edu:8080/exist/xquery/data.xq?id=hf037>, accessed 2012.
- Grimm, N., Hope, D., Gries, C., Martin, C. & Burns, E. (2005) “Survey 200 long term study of multiple sites in central Arizona-Phoenix.” *Central Arizona-Phoenix Long-Term Ecological Research. Global Institute of Sustainability, Arizona State University*. Available at: <https://caplter.asu.edu/data/data-catalog/?id=278>, accessed 2012.
- Gunatilleke, C.V.S., Gunatilleke, I.A.U.N., Ashton, P.S., Ethugala, A.U.K., Weerasekera, N.S. & Esufali, S. (2004) *Sinharaja Forest Dynamics Plot, Sri Lanka. Forest Diversity and Dynamism: Findings from a Large-Scale Plot Network* (ed. by E.C. Losos and E.G.J. Leigh), pp. 599–608. University of Chicago Press, Chicago.
- Gunatilleke, N. & Gunatilleke, S. “Sinharaja Forest Dynamics Plot, Sri Lanka.” *The Center for Tropical Forest Science. Smithsonian Tropical Research Institute*. Available at: <http://www.ctfs.si.edu/site/Sinharaja/>, accessed 2013.
- Hale, S.S., Hughes, M.M., Strobel, C.J., Buffum, H.W., Copeland, J.L. & Paul, J.F. (2002) Coastal ecological data from the Virginian Biogeographic Province, 1990–1993. *Ecology*, 83, 2942–2942.
- Harmon, M. & Franklin, J. (2012) “Long-term growth, mortality and regeneration of trees in permanent vegetation plots in the Pacific Northwest, 1910 to present.” *Long-Term Ecological Research. Forest Science Data Bank, Corvallis*. Available at: <http://andrewsforest.oregonstate.edu/data/abstract.cfm?dbcode=TV010>, accessed 2012.
- HMANA “Hawk Migration Association of North America (HMANA).” Available at: <http://www.hmana.org/>, accessed 2012.

- Hosley, N. (2003) "Vegetation Inventory of Harvard Forest 1937. Harvard Forest Data Archive: HF039." *The Harvard Forest Long Term Ecological Research Program*. Available at: <http://harvardforest.fas.harvard.edu:8080/exist/xquery/data.xq?id=hf039>, accessed 2012.
- Hubbell, S.P., Condit, R. & Foster, R.B. (2010) "Barro Colorado Island Forest Dynamics Plot, Panama." *The Center for Tropical Forest Science. Smithsonian Tropical Research Institute*. Available at: <http://www.ctfs.si.edu/site/Barro+Colorado+Island/>, accessed 2013.
- Hubbell, S.P., Foster, R.B., O'Brien, S.T., Harms, K.E., R., C., Wechsler, B., S.J. Wright & Lao, S.L. de (1999) Light gap disturbances, recruitment limitation, and tree diversity in a neotropical forest. *Science*, 283, 554–557.
- Johnson, S. & Farrand, A. (2014) "Aquatic insect sampling in Lookout Creek at the H.J. Andrews Experimental Forest, 2001." *Long-Term Ecological Research. Forest Science Data Bank, Corvallis*. Available at: <http://andrewsforest.oregonstate.edu/data/abstract.cfm?dbcode=SA017>, accessed 2012.
- Jones, J. & Miller, J. (2005) "Spatial and temporal distribution and abundance of moths in the Andrews Experimental Forest, 1994 to 2008." *H. J. Andrews Experimental Forest. Forest Science Data Bank, Corvallis*. Available at: <http://andrewsforest.oregonstate.edu/data/abstract.cfm?dbcode=SA015>, accessed 2012.
- Kenfack, D., Thomas, D.W., Chuyong, G. & Condit, R. (2007) Rarity and abundance in a diverse African forest. *Biodiversity and Conservation*, 16, 2045–2074.
- Kittredge, D., Foster, D. & McDonald, R. (2009) "Timber Harvesting Field Study in Western Massachusetts 2004-2005. Harvard Forest Data Archive: HF127." *The Harvard Forest Long Term Ecological Research Program*. Available at: <http://harvardforest.fas.harvard.edu:8080/exist/xquery/data.xq?id=hf127>, accessed 2013.
- Lathrop, R. (2000) "Madison Wisconsin Lakes Zooplankton 1976 - 1994." *North Temperate Lakes Long Term Ecological Research Program, Center for Limnology, University of Wisconsin-Madison*. Available at: <http://lter.limnology.wisc.edu/dataset/madison-wisconsin-lakes-zooplankton-1976-1994>, accessed 2013.
- Lauenroth, W.K. (2013) "SGS-LTER Disturbance intensity and above- and belowground herbivory effects on long-term recovery of shortgrass steppe on the Central Plains Experimental Range, Nunn, Colorado, USA 1977-1990." *Shortgrass Steppe (SGS) Long Term Ecological Research Program*. Available at: http://sgslter.colostate.edu/dataset_view.aspx?id=grubr, accessed 2013.

Lee, H., Davies, S.J., LaFrankie, J. V., Tan, S., Itoh, A., Yamakura, T. & Ashton, P.S. (2002) Floristic and structural diversity of 52 hectares of mixed dipterocarp forest in Lambir Hills National Park, Sarawak, Malaysia. *Journal of Tropical Forest Science*, 14, 379–400.

Lee, H.S., Ashton, P.S., Yamakura, T., Tan, S., Davies, S.J., Itoh, A., Chai, E.O.K., Okhubo, T. & LaFrankie, J. V. (2005) *The 52-hectare Forest Research Plot at Lambir Hills, Sarawak, Malaysia: Tree distribution maps, diameter tables and species documentation*. Forest Department Sarawak, The Arnold Arboretum-CTFS Asia Program, Smithsonian Tropical Research Institute, Kuching, Sarawak, Malaysia.

Manokaran, N., Seng, Q.E., Ashton, P.S., LaFrankie, J. V., Noor, N.S.M., Ahmad, W.M.S. & Okuda, T. (2004) *Pasoh Forest Dynamics Plot, Peninsular Malaysia. Forest Diversity and Dynamism: Findings from a Large-Scale Plot Network* (ed. by E.C. Losos and E.G.J. Leigh), pp. 585–598. University of Chicago Press, Chicago.

Marine Resources Research Institute (2011) “Southeast Area Monitoring and Assessment Program (SEAMAP) South Atlantic.” *SCDNR SEAMAP-SA Program. U.S. Geological Survey*. Available at: <http://www.usgs.gov/obis-usa/>, accessed 2013.

McLarney, W.O., Meador, J. & Chamblee, J. (2010) “Upper Little Tennessee River Biomonitoring Program Database.” *Coweeta Long Term Ecological Research Program*. Available at: http://coweeta.uga.edu/dbpublic/dataset_details.asp?accession=LTWA_2010_06_01, accessed 2012.

Muldavin, E. (a) “Pinon Juniper Net Primary Production Quadrat Data from the Sevilleta National Wildlife Refuge, New Mexico: 1999-2001.” *Sevilleta Long Term Ecological Research Program*. Available at: <http://sev.lternet.edu/data/sev-187>, accessed 2013.

Muldavin, E. (b) “Pinon-Juniper (Core Site) Quadrat Data for the Net Primary Production Study at the Sevilleta National Wildlife Refuge, New Mexico (2003-Present).” *Sevilleta Long Term Ecological Research Program*. Available at: <http://sev.lternet.edu/node/1718>, accessed 2013.

NatureCounts (a) “Maritimes Breeding Bird Atlas (2006-2010): point count data.” *NatureCounts, a node of the Avian Knowledge Network. Bird Studies Canada*. Available at: <http://www.birdscanada.org/birdmon/>, accessed 2012.

NatureCounts (b) “Ontario Breeding Bird Atlas (2001-2005): point count data.” *NatureCounts, a node of the Avian Knowledge Network. Bird Studies Canada*. Available at: <http://www.birdscanada.org/birdmon/>, accessed 2012.

Nilon, C. (2010) "Biodiversity - Fauna - Bird Survey - Table 1 of 4 - Birds." *Baltimore Ecosystem Study LTER Program. Baltimore, MD, USA.* Available at: http://beslter.org/metacat_harvest_attribute_level_eml/html_metadata/bes_543.asp, accessed 2012.

NIWA "South Western Pacific Regional OBIS Data Specify Subset (South Western Pacific OBIS)." *National Institute of Water and Atmospheric Research.* Available at: <http://iobis.org/>, accessed 2013.

NOAA/NOS/NCCOS/CCMA (2007a) "La Parguera, Puerto Rico Fish Assessment and Monitoring Data (2002 - Present)." *National Oceanic and Atmospheric Association (NOAA)/National Ocean Service (NOS)/National Centers for Coastal Ocean Science (NCCOS)/Center for Coastal Monitoring and Assessment (CCMA) - Biogeography Team. Silver Spring, MD.* Available at: <http://iobis.org/>, accessed 2013.

NOAA/NOS/NCCOS/CCMA (2007b) "St. Croix, USVI Fish Assessment and Monitoring Data (2002 - Present)." *National Oceanic and Atmospheric Association (NOAA)/National Ocean Service (NOS)/National Centers for Coastal Ocean Science (NCCOS)/Center for Coastal Monitoring and Assessment (CCMA) - Biogeography Team. Silver Spring, MD.* Available at: <http://iobis.org/>, accessed 2013.

NOAA/NOS/NCCOS/CCMA (2007c) "St. John, USVI Fish Assessment and Monitoring Data (2002 - Present)." *National Oceanic and Atmospheric Association (NOAA)/National Ocean Service (NOS)/National Centers for Coastal Ocean Science (NCCOS)/Center for Coastal Monitoring and Assessment (CCMA) - Biogeography Team. Silver Spring, MD.* Available at: <http://iobis.org/>, accessed 2013.

Northeast Fisheries Science Center (2005) "Northeast Fisheries Science Center Bottom Trawl Survey Data (OBIS-USA)." *NOAA's National Marine Fisheries Service (NMFS) Northeast Fisheries Science Center. Woods Hole, Massachusetts, USA.* Available at: <http://iobis.org/>, accessed 2013.

NTL LTER (2012) "North Temperate Lakes LTER: Fish Abundance 1981 - current." *North Temperate Lakes Long Term Ecological Research Program, Center for Limnology, University of Wisconsin-Madison.* Available at: <http://lter.limnology.wisc.edu/dataset/north-temperate-lakes-lter-fish-abundance-1981-current>, accessed 2012.

NTL LTER (2011) "North Temperate Lakes LTER: Zooplankton - Madison Lakes Area 1997 - current." *North Temperate Lakes Long Term Ecological Research Program, Center for Limnology, University of Wisconsin-Madison.* Available at: <http://lter.limnology.wisc.edu/dataset/north-temperate-lakes-lter-zooplankton-madison-lakes-area-1997-current>, accessed 2013.

- Ohmart, R., Pearson, D., Hostetler, M., Katti, M. & Hulen, T. (2003) "Transect bird survey with data synthesis from multiple transects in the central Arizona-Phoenix area: period 1998 to 2000." *Central Arizona-Phoenix Long-Term Ecological Research. Global Institute for Sustainability, Arizona State University*. Available at: <https://caplter.asu.edu/data/data-catalog/?id=43>, accessed 2012.
- Oliveira, A. de "Ilha do Cardoso, Brasil." *The Center for Tropical Forest Science. Smithsonian Tropical Research Institute*. Available at: <http://www.ctfs.si.edu/site/Ilha+do+Cardoso/>, accessed 2013.
- Paquette, A., Laliberté, E., Bouchard, A., Blois, S. de, Legendre, P. & Brisson, J. (2007) Lac Croche understory vegetation data set (1998–2006). *Ecology*, 88, 3209–3209.
- Pardieck, K.L., Ziolkowski Jr., D.J. & Hudson, M.-A.R. (2015) "North American Breeding Bird Survey Dataset 1966 - 2014, version 2013.0." *U.S. Geological Survey, Patuxent Wildlife Research Center*. Available at: <https://www.pwrc.usgs.gov/bbs/RawData/>, accessed 2012.
- Ramesh, B.R., Swaminath, M.H., Patil, S. V., Dasappa, Pélissier, R., Venugopal, P.D., Aravajy, S., Elouard, C. & Ramalingam, S. (2010) Forest stand structure and composition in 96 sites along environmental gradients in the central Western Ghats of India. *Ecology*, 91, 3118–3118.
- Reich, P., Wedin, D., Hobbie, S. & Davis, M. "Experiment 133 - Effect of Burning Patterns on Vegetation in the Fish Lake Burn Compartments - Shrub Survey." *Cedar Creek Ecosystem Science Reserve*. Available at: <http://www.cedarcreek.umn.edu/research/data/experiment?e133>, accessed 2012.
- Reichert, M. (2009) "MARMAP Chevron Trap Survey 1990-2009." *SCDNR/NOAA MARMAP Program, SCDNR MARMAP Aggregate Data Surveys, The Marine Resources Monitoring, Assessment, and Prediction (MARMAP) Program, Marine Resources Research Institute, South Carolina Department of Natural Resources USA*. Available at: <http://www.usgs.gov/obis-usa/>, accessed 2013.
- Reichert, M. (2010a) "MARMAP Fly Net 1990-2009." *SCDNR/NOAA MARMAP Program, SCDNR MARMAP Aggregate Data Surveys, The Marine Resources Monitoring, Assessment, and Prediction (MARMAP) Program, Marine Resources Research Institute, South Carolina Department of Natural Resources USA*. Available at: <http://www.usgs.gov/obis-usa/>, accessed 2013.
- Reichert, M. (2010b) "MARMAP Yankee Trawl 1990-2009." *SCDNR/NOAA MARMAP Program, SCDNR MARMAP Aggregate data surveys, The Marine Resources Monitoring, Assessment, and*

Prediction (MARMAP) Program, Marine Resources Research Institute, South Carolina Department of Natural Resources USA. Available at: <http://www.usgs.gov/obis-usa>, accessed 2013.

Sarnthein, M., van Kreveld, S.A., Erlenkeuser, H., Grootes, P.M., Kucera, M., Pflaumann, U. & Schulz, M. (2003) "Distribution of foraminifera of sediment core GIK23258-2". doi:10.1594/PANGAEA.114682, In Supplement to: Sarnthein et al. (2003): Centennial-to-millennial-scale periodicities of Holocene climate and sediment injections off western Barents shelf, 75°N. *Boreas*, 32(3), 447-461, doi:10.1111/j.1502-3885.2003.tb01227.x. *PANGAEA*. Available at: <http://www.gbif.org/dataset/8750386c-f762-11e1-a439-00145eb45e9a>, accessed 2012.

Shochat, E., Katti, M. & Warren, P. (2004) "Point count bird censusing: long-term monitoring of bird distribution and diversity in central Arizona-Phoenix: period 2000 to 2011." *Central Arizona-Phoenix Long-Term Ecological Research. Global Institute for Sustainability, Arizona State University*. Available at: <https://caplter.asu.edu/data/data-catalog/?id=46>, accessed 2012.

Silveira, F.L. & Lopes, R.M. (2011) "Previous fisheries REVIZEE Program." *WSAOBIS*. Available at: <http://iobis.org/>, accessed 2013.

Sipe, T., Bowden, R. & McClaugherty, C. (2009) "Tree Seedlings in CRUI Land Use Project at Harvard Forest 1996. Harvard Forest Data Archive: HF141." *The Harvard Forest Long Term Ecological Research program*. Available at: <http://harvardforest.fas.harvard.edu:8080/exist/xquery/data.xq?id=hf141>, accessed 2012.

Stevens, M. (2010) "Trekvis - Migratory fishes in the river Scheldt." *Research Institute for Nature and Forest (INBO)*. Available at: <http://www.gbif.org/dataset/b2d0f29e-4614-4001-93c8-f651878a86d2>, accessed 2014.

Su, S., Chang-Yang, C., Lu, C., Tsui, C., Lin, T., Lin, C., Chiou, W., Kuan, L., Chen, Z. & Hsieh, C. (2007) *Fushan subtropical forest dynamics plot: tree species characteristics and distribution patterns*, Taipei.

Sukumar, R. "Mudumalai Forest Dynamics Plot, India." *The Center for Tropical Forest Science. Smithsonian Tropical Research Institute*. Available at: <http://www.ctfs.si.edu/site/Mudumalai/>, accessed 2013.

Sukumar, R., Suresh, H.S., Dattaraja, H.S., John, R. & Joshi, N. V. (2004) *Mudumalai Forest Dynamics Plot, India. Forest Diversity and Dynamism: Findings from a Large-Scale Plot Network* (ed. by E.C. Losos and E.G.J. Leigh), pp. 551–563. University of Chicago Press, Chicago.

Sun, I.F. & Hsieh, C.-F. “Fushan Forest Dynamics Plot, Taiwan.” *The Center for Tropical Forest Science. Smithsonian Tropical Research Institute*. Available at: <http://www.ctfs.si.edu/site/Fushan/>, accessed 2013.

Tan, S., Davies, S., Yamakura, T. & Itoh, A. “Lambir Hills Forest Dynamics Plot, Malaysia.” *The Center for Tropical Forest Science. Smithsonian Tropical Research Institute*. Available at: <http://www.ctfs.si.edu/site/Lambir/>, accessed 2013.

The Swedish University of Agriculture Sciences “National Forest Inventory (SLU).” *GBIF-Sweden*. Available at: <http://www.gbif.org/dataset/c46708d0-12aa-11dd-9ff0-b8a03c50a862>, accessed 2013.

Thomas, D., Kenfack, D. & Chuyong, G. “Korup Forest Dynamics Plot, Cameroon.” *The Center for Tropical Forest Science. Smithsonian Tropical Research Institute*. Available at: <http://www.ctfs.si.edu/site/Korup/>, accessed 2013.

Thomas, D.W., Kenfack, D., Chuyong, G.B., Sainge, N.M., Losos, E.C., Condit, R.S. & Songwe, N.C. (2003) Tree Species of Southwestern Cameroon: Tree distribution maps, diameter tables and species documentation of the 50-ha Korup Forest Dynamics Plot. Center for Tropical Forest Science, Washington, D.C.

Trexler, J. (2007) “Consumer Stocks: Fish, Vegetation, and other Non-physical Data from Everglades National Park (FCE), South Florida from February 2000 to Present.” *Florida Coastal Everglades LTER Program*. Available at: http://fcelter.fiu.edu/data/core/metadata/EML/?datasetid=LT_CD_Trexler_001, accessed 2012.

USDA Forest Service (2007) “USDA Forest Service, Redwood Sciences Laboratory - Lamna Point Count.” *Avian Knowledge Network*. Available at: <http://www.gbif.org/dataset/864da4c2-f762-11e1-a439-00145eb45e9a>, accessed 2012.

USFS “Landbird Monitoring Program (UMT-LBMP).” *US Forest Service*. Available at: <http://www.avianknowledge.net/>, accessed 2012.

Valencia, R. “Yasuni Forest Dynamics Plot, Ecuador.” *The Center for Tropical Forest Science. Smithsonian Tropical Research Institute*. Available at: <http://www.ctfs.si.edu/site/Yasuni/>, accessed 2013.

Valencia, R., Condit, R., Foster, R.B., Romoleroux, K., Muñoz, G.V., Svenning, J.-C., Magård, E., Bass, M., Losos, E.C. & Balslev, H. (2004) *Yasuni Forest Dynamics Plot, Ecuador. Forest Diversity*

and Dynamism: Findings from a Large-Scale Plot Network (ed. by E.C. Losos and E.G.J. Leigh), pp. 609–620. University of Chicago Press, Chicago.

Vallejo, M.I., Samper, C., Mendoza, H. & Otero, J.T. (2004) *La Planada Forest Dynamics Plot, Colombia. Forest Diversity and Dynamism: Findings from a Large-Scale Plot Network* (ed. by E.C. Losos and E.G.J. Leigh), pp. 517–526. University of Chicago Press, Chicago.

Viereck, L.A., Van Cleve, K., Chapin, F.S.S., Ruess, R.W. & Hollingsworth, T.N. (2005) “Vegetation Plots of the Bonanza Creek LTER Control Plots: Species Count (1975 - 2004).” *Bonanza Creek LTER - University of Alaska Fairbanks*. Available at: http://www.lter.uaf.edu/data_detail.cfm?datafile_pkey=175, accessed 2012.

Wade, E.J. (2011) “Snow crab research trawl survey database (Southern Gulf of St. Lawrence, Gulf region, Canada) from 1988 to 2010.” *OBIS Canada Digital Collections. OBIS Canada, Bedford Institute of Oceanography, Dartmouth, Nova Scotia, Canada*. Available at: <http://iobis.org/>, accessed 2013.

Walker, J., Shochat, E., Katti, M. & Warren, P. (2004) “Point Count Bird Censusing Data Subset for Paper ‘Effects of land use and vegetation cover on bird communities’ Walker et. al.” *Central Arizona-Phoenix Long-Term Ecological Research. Global Institute for Sustainability, Arizona State University*. Available at: <https://caplter.asu.edu/data/data-catalog/?id=394>, accessed 2012.

Warren, P., Kinzig, A., Martin, C. & Machabee, L. (2005) “Ecological and social Interactions in urban parks: bird surveys in local parks in the central Arizona-Phoenix metropolitan area.” *Central Arizona-Phoenix Long-Term Ecological Research. Global Institute for Sustainability, Arizona State University*. Available at: <https://caplter.asu.edu/data/data-catalog/?id=256>, accessed 2012.

Williams, D. “Pelagic Fish Observations 1968-1999.” *Australian Antarctic Data Centre*. Available at: <http://www.gbif.org/dataset/85b0a82a-f762-11e1-a439-00145eb45e9a>, accessed 2012.

Woehler, E. (1999, updated 2015) “Distribution and abundance of seabirds in the Southern Indian Ocean, 1978/1979+”. *Australian Antarctic Data Centre* - doi:10.4225/15/5643E8C0743C2. Available at <http://www.iobis.org>, accessed 2013.

Woods, K.D. (2009) Multi-decade, spatially explicit population studies of canopy dynamics in Michigan old-growth forests. *Ecology*, 90, 3587–3587.

Zimmerman, J. “LFDP phenology plot seedlings – 16 ha plot.” *Luquillo Long Term Ecological Research Program*. Available at: <http://luq.lternet.edu/data/luqmetadata175>, accessed 2014.

Zimmerman, J. & Brokaw, N. "Census of species, diameter and location at the Luquillo Forest Dynamics Plot (LFDP), Puerto Rico." *Luquillo Long Term Ecological Research Program*. Available at: <http://luq.lternet.edu/data/luqmetadata119>, accessed 2013.

Zimmerman, J., Thompson, J. & Brokaw, N. "Luquillo Forest Dynamics Plot, Puerto Rico." *The Center for Tropical Forest Science. Smithsonian Tropical Research Institute*. Available at: <http://www.ctfs.si.edu/site/Luquillo/>, accessed 2013.

Zimmerman, J.K., Comita, L.S., Thompson, J., Uriarte, M. & Brokaw, N. (2010) Patch dynamics and community metastability of a subtropical forest: compound effects of natural disturbance and human land use. *Landscape Ecology*, 25, 1099–1111.

References for (additional) Data Sources for the Scaling analysis

Edgar, G.J. & Stuart-Smith, R.D. (2008) “Reef Life Survey (RLS): Invertebrates.” *Institute for Marine and Antarctic Studies (IMAS)*. Available at: <http://catalogue-rls.imas.utas.edu.au/geonetwork/srv/en%0A/metadata.show?uuid=60978150-1641-11dd-a326-00188b4c0af>, accessed 2016.

Edgar, G.J. & Stuart-Smith, R.D. (2014a) “Reef Life Survey (RLS): Global reef fish dataset.” *Institute for Marine and Antarctic Studies (IMAS)*. Available at: <http://catalogue-rls.imas.utas.edu.au/geonetwork/srv/en%0A/metadata.show?uuid=9c766140-9e72-4bfb-8f04-d51038355c5>, accessed 2016.

Edgar, G.J. & Stuart-Smith, R.D. (2014b) Systematic global assessment of reef fish communities by the Reef Life Survey program. *Scientific Data*, **1**, 140007.

North Pacific Groundfish Observer Program. “North Pacific Groundfish Observer”. *Alaska Fisheries Science Center*. Available at: <http://iobis.org/>, accessed 2013.

Pardieck, K.L., Ziolkowski, D.J.Jr., Hudson, M.-A.R. & Campbell, K. (2016) “North American Breeding Bird Survey Dataset 1966 - 2015, version 2015.1”. *U.S. Geological Survey, Patuxent Wildlife Research Center*. Available at: www.pwrc.usgs.gov/BBS/RawData/, accessed 2016. doi 10.5066/F7C53HZN.

USDA Forest Service (2010) Forest inventory and analysis national core field guide (Phase 2 and 3). Version 4.0. USDA Forest Service, Forest Inventory and Analysis, Washington, D.C., USA. Accessed 2016.

Woudenberg, S.W., Conkling, B.L., O’Connell, B.M., LaPoint, E.B., Turner, J.A. & Waddell, K.L. (2010). The Forest Inventory and Analysis Database: Database description and users manual version 4.0 for Phase 2. Gen. Tech. Rep. RMRS-GTR-245. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fort Collins, CO., U.S.

Acknowledgements for the Data Sources

Australian Antarctic Data Centre; EurOBIS; NOAA/NCCOS/CCMA; OBIS-Australia; OBIS-Canada; OBIS-USA; South Western Pacific OBIS; Tropical and Subtropical Western South Atlantic OBIS; the Long Term Ecological Research (LTER) Network; HJ Andrews Experimental Forest (NSF LTER Grant DEB 08-23380, US Forest Service Pacific Northwest Research Station, and Oregon State University); Avian Knowledge Network; Point Blue Conservation Science (PRBO Conservation Science); The Hawk Migration Association of North America (HMANA, HawkCount database - www.hawkcount.org), the site coordinators and hawk watchers; NatureCounts; Bird Studies Canada; Baltimore Ecosystem Study LTER (NSF Grant 0423476 and Cary Institute of Ecosystem Studies); Bonanza Creek LTER (NSF Grants DEB-0620579, DEB-0423442, DEB-0080609, DEB-9810217, DEB-9211769, DEB-8702629, the USDA Forest Service, Pacific Northwest Research Station (Agreement # RJVA-PNW-01-JV-11261952-231)); Breeding Bird Survey of North America; California Current Ecosystem LTER (Division of Ocean Sciences, NSF Grants OCE-0417616 and OCE-10-26607); Cedar Creek Ecosystem Science Reserve and the University of Minnesota (NSF LTER Grants DEB-0620652 and DEB-1234162); Central Arizona-Phoenix LTER (NSF Grants BCS-1026865, DEB-0423704 and DEB-9714833); Coweeta LTER (NSF Grants DEB-1440485, DEB-0823293, DEB-9632854 and DEB-0218001); Florida Coastal Everglades LTER Program (NSF Grants DEB-1237517, DBI-0620409, and DEB-9910514) and the Everglades National Park; Harvard Forest LTER; Hubbard Brook Ecosystem Study (NSF DEB-1114804); Konza Prairie LTER (NSF Grants DEB-0218210, DEB-0823341); Luquillo Experimental Forest LTER (NSF Grants BSR-8811902, DEB 9411973, DEB 0080538, DEB 0218039, DEB 0620910, DEB 0963447 and DEB-129764, the University of Puerto Rico, and the International Institute of Tropical Forestry (IITF)); North Temperate Lakes LTER; Sevilleta LTER (NSF Grants BSR 88-11906, DEB 9411976, DEB 0080529 and DEB 0217774); Shortgrass Steppe LTER (NSF DEB-1027319); Center for Tropical Forest Science of the Smithsonian Tropical Research Institute (NSF Grants DEB-0640386, DEB-0425651, DEB-0346488, DEB-0129874, DEB-00753102, DEB-9909347, DEB-9615226, DEB-9615226, DEB-9405933, DEB-9221033, DEB-9100058, DEB-8906869, DEB-8605042, DEB-8206992, DEB-7922197, the John D. and Catherine T. MacArthur Foundation, the Mellon Foundation, the Small World Institute Fund); the principal investigators of the Korup Forest Dynamics Plot; Indian Institute of Science; Arnold Arboretum of Harvard University (NSF DEB-9107247, DEB-9629601 and DEB-0075334, grants from USAID and the Rockefeller Foundation); Forest Department of Sarawak (Malaysia) and Osaka City, Ehime & Kyoto Universities (Monbusho Grants 06041094, 08NP0901 and 09NP0901); Forest Research Institute

Malaysia; National Institute of Environmental Studies (Japan); University of Peradeniya and the Forest Department of Sri Lanka; Taiwan Forestry Research Institute and Tunghai University; Royal Thai Forest Department; National Parks Wildlife and Plant Conservation Department; Thai National Park, Wildlife and Plant Conservation Department; Thai Ministry of Natural Resources and Environment; National Center for Genetic Engineering and Biotechnology (Thailand); National Science and Technology Development Agency (Thailand); Universidade de São Paulo; Instituto de Investigación de Recursos Biológicos "A. Von Humboldt"; Pontificia Universidad Católica del Ecuador, Estación Biológica Yasuni, Herbario QCA (Ecuador), University of Aarhus (Denmark); Virginia Coast Reserve LTER (NSF Grant 1237733); Danish Biodiversity Information Facility; Research Institute for Nature and Forest (INBO); GBIF-Sweden; PANGAEA - Publishing Network for Geoscientific and Environmental Data; the Institute for Marine and Antarctic Studies at the University of Tasmania, Australia.

Appendix III

R Code for fitting mixtures of 1, 2 and 3 PLN distributions (1PLN, 2PLN and 3PLN, respectively) and calculate maximum likelihood estimates (MLE)

```
##need to load package 'poilog'
```

```
##The species abundance data should be vector called 'counts'; e.g. counts<- c(1,3,5,1,25)  
##represents a community with two species with abundance 1, and another three species with  
##abundances 3, 5 and 25
```

```
##R optimization routine nlminb can be used to estimate the best-fit parameters
```

```
##The parameter searches should be initialized from several starting points
```

```
##have to define max.abund= max(counts)
```

```
##example of performing the fitting for 1PLN:
```

```
this.fit <- try(nlminb(inipar1, loglike.1, lower = c(0.000001,0.000001), upper = c(max.abund,10)),  
TRUE)
```

```
##estimating parameter mu between (0.000001, max.abund)
```

```
##estimating parameter sig between (0.000001, 10)
```

```
##Users of these functions should cite (a) the manuscript Antão et al. (2016), (b) the R software  
program, and also (c) the package poilog
```

1PLN function

```
loglike.1 <- function(params) {  
  
  # params[1] is the mean of log-abundance (backtransformed to arithmetic scale)  
  
  # params[2] is the standard deviation of log-abundance  
  
  
  # First, make the vector of probabilities for all possible states:  
  
  probs <- rep(NA,max.abund)  
  
  
  # Next, get the specific abundance values that we need to calculate, i.e., the observed values  
  
  abund <- sort(unique(counts))  
  
  
  # Use dpoilog to calculate probability for each observed abundance value  
  
  # and place in corresponding place in the probs vector  
  
  probs[abund] <- dpoilog(abund,mu=log(params[1]),params[2])  
  
  
  # NOTE that in poilog, "mu" is the mean of log-abundance (not back-  
  
  # transformed), so params[1] needs to be logged before getting passed  
  
  
  # probability species is absent  
  
  p0 <- dpoilog(0,mu=log(params[1]),params[2])  
  
  
  # zero-truncated probabilities:  
  
  totprob <- probs/(1-p0)
```

```

# Finally, calculate the -log likelihood for the data set:

sum(-log(totprob[counts]),na.rm=T)

}

```

2PLN function

```

loglike.2 <- function(params) {

# Mixture of two PLN distributions

# params[1] is the mean of log-abundance (backtransformed to arithmetic scale) for distribution 1
# params[2] is the standard deviation of log-abundance for distribution 1
# params[3:4] are the corresponding parameters for distribution 2
# params[5] is the probability that an observed species is from distribution 1

# as per loglike.1b, but need to initialize a vector for the probabilities for each distribution
probs1 <- rep(NA,max.abund)

probs2 <- probs1

abund <- sort(unique(counts))

# First, make the vector of probabilities for the two distributions individually,
# and the associated probability that a species from
# each distribution is not present in the sample.

```

```

probs1[abund] <- dpoilog(abund,mu=log(params[1]),params[2])
p01 <- dpoilog(0,mu=log(params[1]),params[2])
probs2[abund] <- dpoilog(abund,mu=log(params[3]),params[4])
p02 <- dpoilog(0,mu=log(params[3]),params[4])

# Then produce mixture distribution (probability=params[5] that
# a species observed in the sample is chosen from distribution 1):
totprob <- (1-params[5])*probs2/(1-p02) + params[5]*probs1/(1-p01)

# Then, calculate the -log likelihood for the data set:
return(sum(-log(totprob[counts]),na.rm=T))
}

```

3PLN function

```

loglike.3 <- function(params) {
# Mixture of three PLN distributions

# params[1:6] are the mu and sig parameters of the constituent PLN distributions
# params[7] is the probability that an observed species is from distribution 1.
# params[8] is the probability that a species is from distribution 2,
# conditional on the species not being from distribution 1.
# (this parameterization guarantees that the mixture probabilities

```

```

# (f1,f2,f3, below) will sum to unity if params[7] and params[8]
# are constrained to lie between 0 and 1.

# as per loglike.1b, but need to initialize a vector for the probabilities for each distribution
probs1 <- rep(NA,max.abund)

probs2 <- probs1

probs3 <- probs1

abund <- sort(unique(counts))

# First, make the vector of probabilities for the three distributions individually,
# and the associated probability that a species from
# each distribution is not present in the sample.

probs1[abund] <- dpoilog(abund,mu=log(params[1]),params[2])
p01 <- dpoilog(0,mu=log(params[1]),params[2])
probs2[abund] <- dpoilog(abund,mu=log(params[3]),params[4])
p02 <- dpoilog(0,mu=log(params[3]),params[4])
probs3[abund] <- dpoilog(abund,mu=log(params[5]),params[6])
p03 <- dpoilog(0,mu=log(params[5]),params[6])

# Then produce mixture distribution (probability=params[7] that
# a species observed in the sample is chosen from distribution 1 and
# params[8] is the conditional prob that species observed is from distribution 2):

```

```

# Calculate mixture probabilities so that f1 + f2 + f3 = 1

f1 <- params[7] # probability that species is from distr 1

f2 <- (1-params[7])*params[8] # probability that species is from distr 2

f3 <- (1-params[7])*(1-params[8]) # probability that species is from distr 3

totprob <- f3*probs3/(1-p03) + f2*probs2/(1-p02) + f1*probs1/(1-p01)

# Then, calculate the -log likelihood for the data set:

return(sum(-log(totprob[counts]),na.rm=T))

}

```

Appendix IV

A. Simulation Study

The pseudo-code to perform the simulations was:

1. Generate 100 simulated SAD samples for logseries and the PLN mixtures with known parameters.
2. For each combination of parameters, run optimization routines performing the fit of the alternative log-likelihood functions.
3. For each simulated SAD sample and for AICc and BIC calculate:
 - $AICdiff = AICc \text{ true model} - \min(AICc \text{ remaining models})$ (similarly for BIC). Negative AICdiff or BICdiff indicate that the true model was successfully selected;
 - The frequency of selecting the true underlying distribution (out of 100).

This piece of code produces sampled abundance data from a 2PLN mixture for a community with 100 species.

```
S<- 100      ## number of species to simulate – species pool

#set parameters

mutrue1<-4   ###mu for distribution 1

mutrue2<-200 ###mu for distribution 2 (fixed octave 8)

sigtrue1<-1  ###sig value for distribution 1

sigtrue2<-1  ###sig value for distribution 2

ptrue<-0.5   ###p proportion (probability species comes from distribution 1)

S1<- round(S*ptrue) #proportion of the species pool from distribution 1

S2<- S-S1 #proportion of the species pool from distribution 2
```

```

#generates a sample from distribution 1

cnt1<-rpoilog(S1,mu=log(mutrue1),sig=sigtrue1)

#generates a sample from distribution 2

cnt2<-rpoilog(S2,mu=log(mutrue2),sig=sigtrue2)

#combines both abundance vectors as single sample

counts<- c(cnt1,cnt2)

```

B. Parametric bootstrap analysis for empirical data

Parametric bootstrap analysis consisted of generating abundance values from a 1PLN probability distribution parameterized using the maximum likelihood parameters $\hat{\mu}$ and $\hat{\sigma}$ (the mean and standard deviation of the log-abundances, respectively) estimated from the empirical data. For instance, for dataset ID4, estimated parameters were $\hat{\mu}=19.21$ and $\hat{\sigma}=5.31$, and the number of species is $S=39$. Using these parameter values and S as sample size, 100 parametric bootstrap samples were generated, and the PLN mixture distributions were fitted.

Code details:

```

#### using the estimated 1PLN best fit parameters to simulate the communities

#### empirical abundance values are in a vector named 'counts'

#### example for dataset ID4:

```

```
mutrue <- 19.21
```

```
sigtrue <- 5.31
```

```
S1 <- 39      ##number of species
```

```
maxabund <- max (counts)      ##get max abund to set the range for pfd
```

```
#Step1: create a pdf using the estimated parameters
```

```
range <- seq (1, maxabund)
```

```
empdist <- dpoilog (range, mu=log(mutrue), sig=sigtrue)
```

```
#Step2: generate 100 sampled distributions from the pdf created above
```

```
dists <- array(NA, dim= c(S1,100))      ####to save sps abundances for each distribution
```

```
for (k in 1:100) {
```

```
  cnt <- sample (range, S1, prob= empdist, replace=T)
```

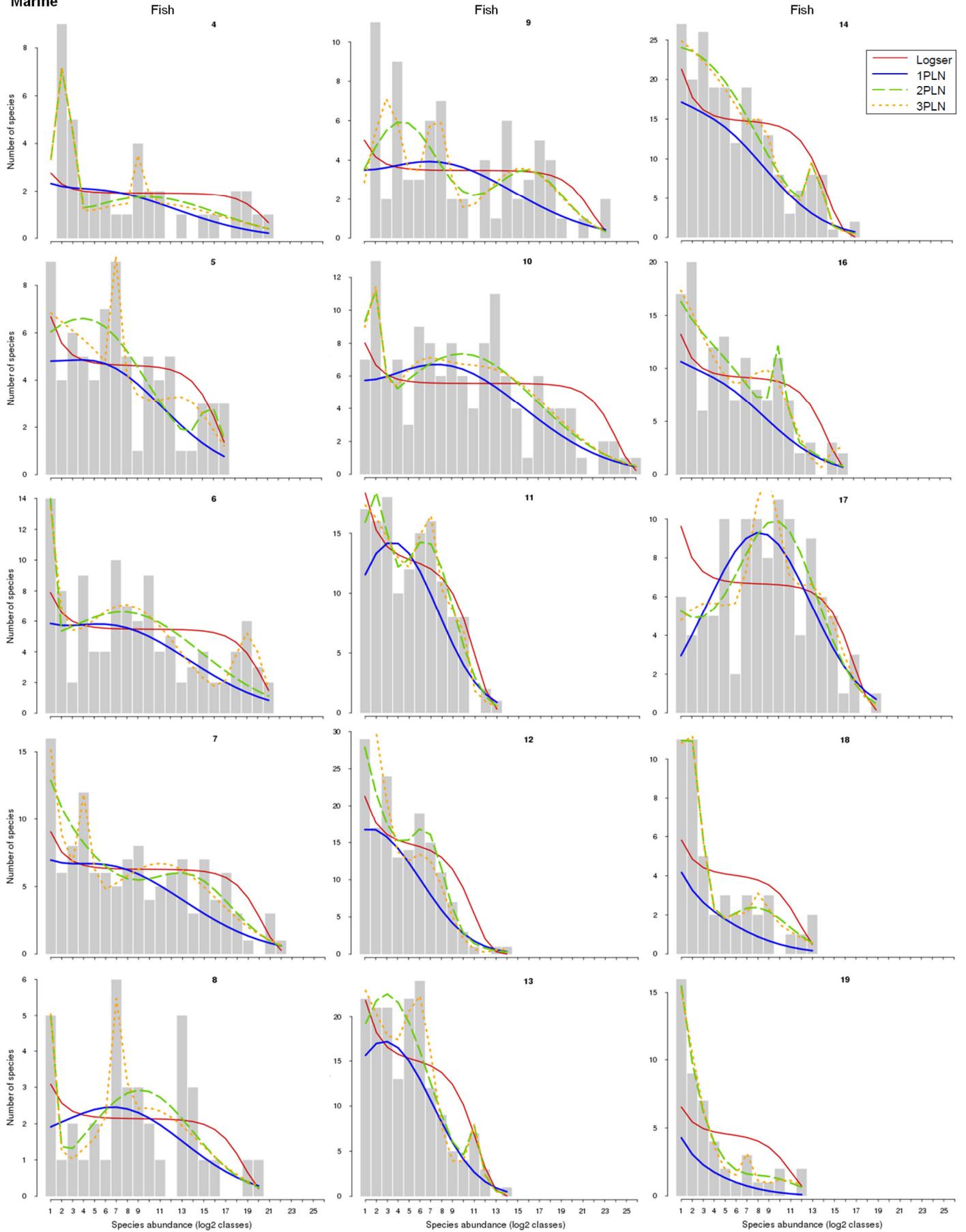
```
  dists [1:length(cnt), k] <- cnt
```

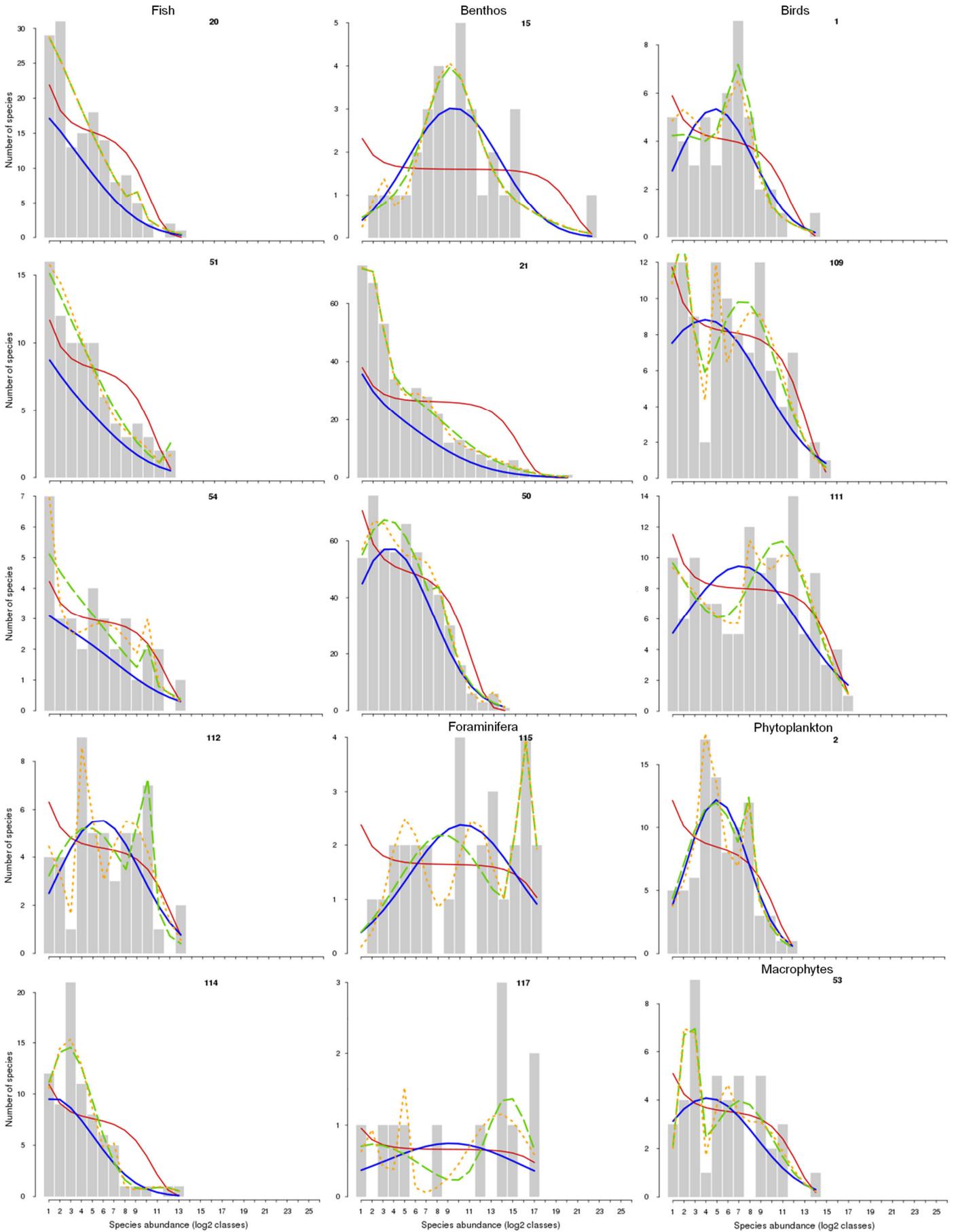
```
}
```

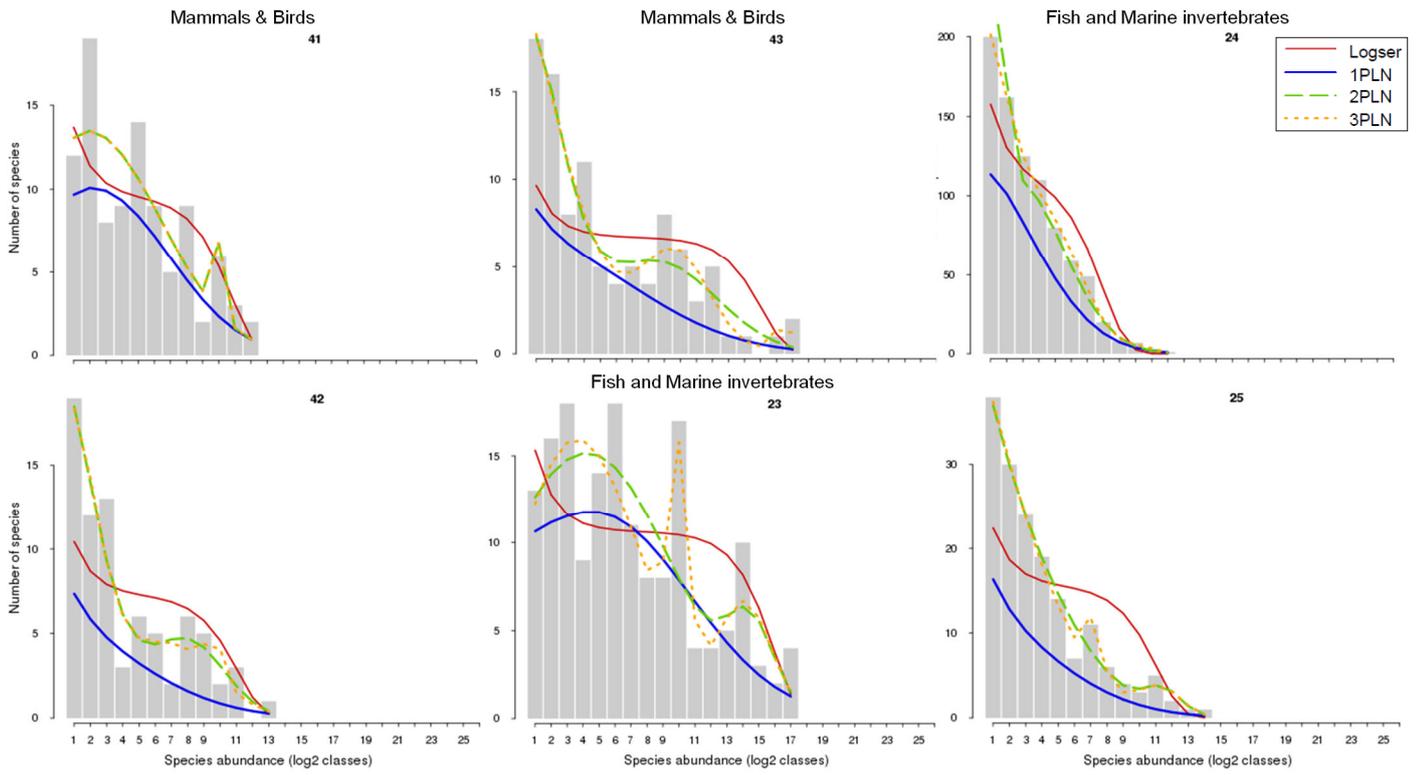

Appendix V

Figure V.1 Non-multimodal empirical species abundance distributions, identified by the corresponding ID. All the fitting routines were run on non-binned data. SADs were plotted with bins representing true doubling classes of abundance, following Gray *et al.* (2006). For all SADs the y-axis is the number of species and the x-axis is the species abundance in log₂ classes (the first bar represents species with abundance 1, the second one species with abundances 2-3, then 4-7, 8-15, etc). The fitted curves are red line for the logseries, bold blue line for 1PLN, dashed green line for 2PLN and dotted orange line for 3PLN. The taxon for each SAD can be identified at the top of the columns in the panel (all the same taxa), or for each individual SAD.

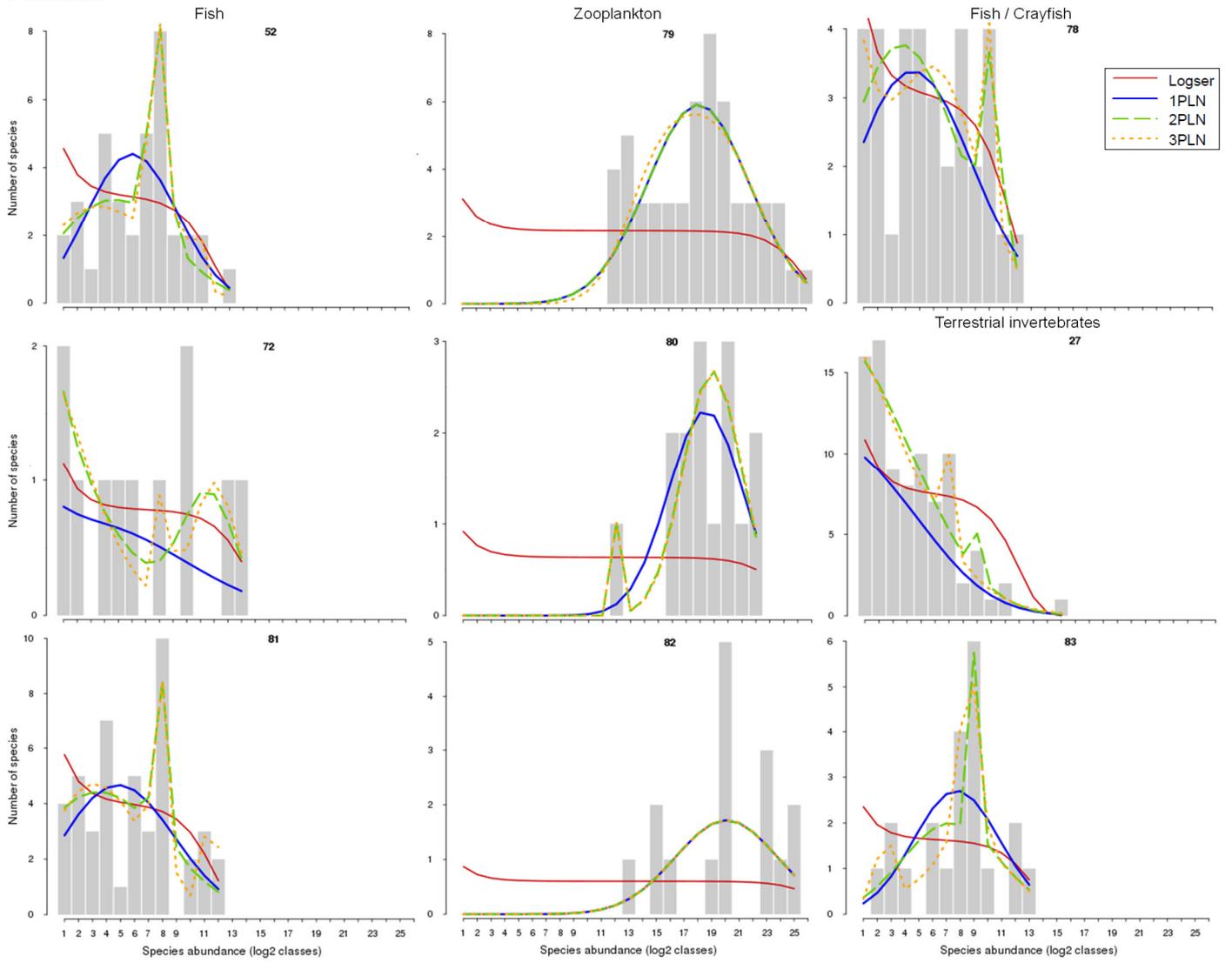
Marine



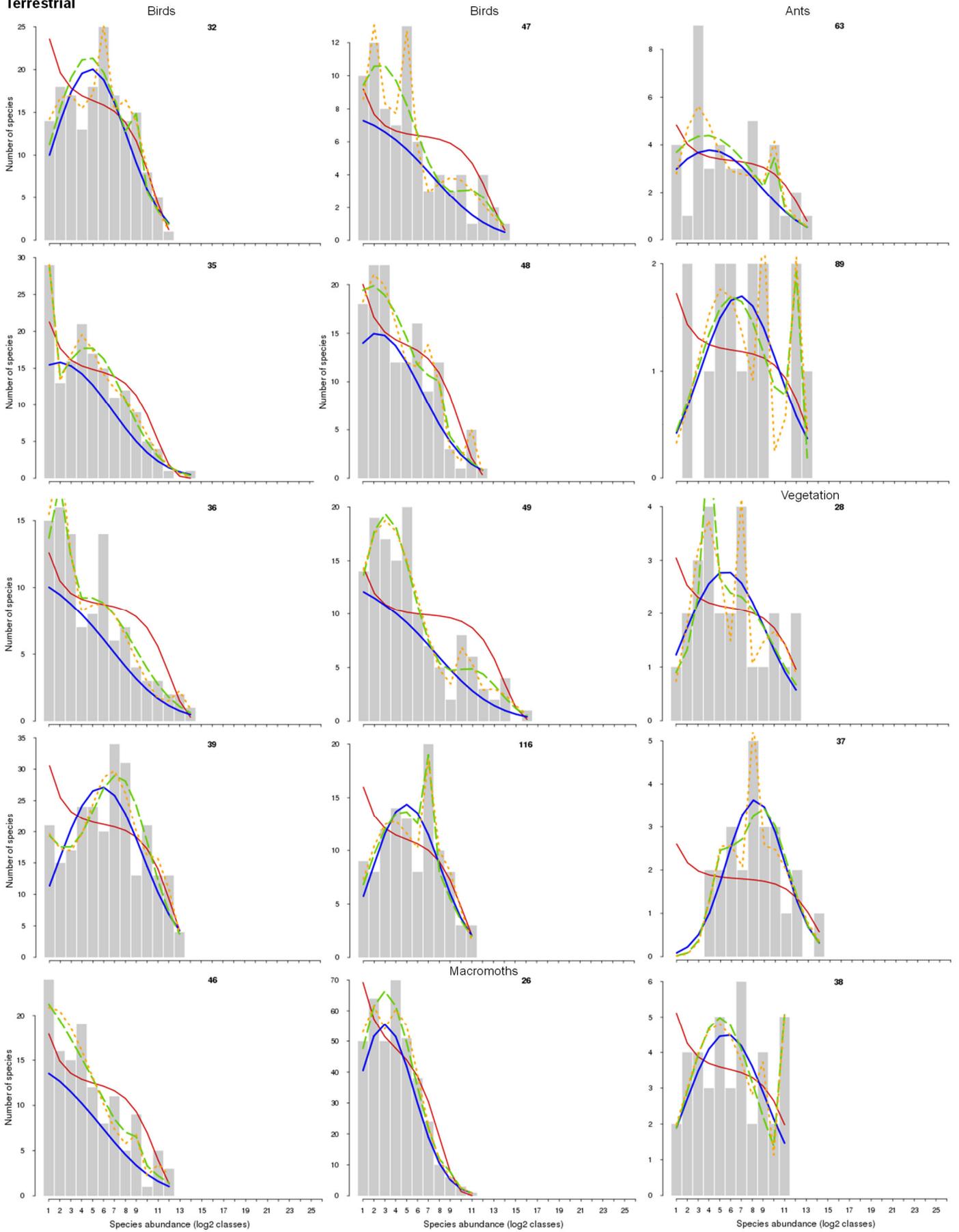


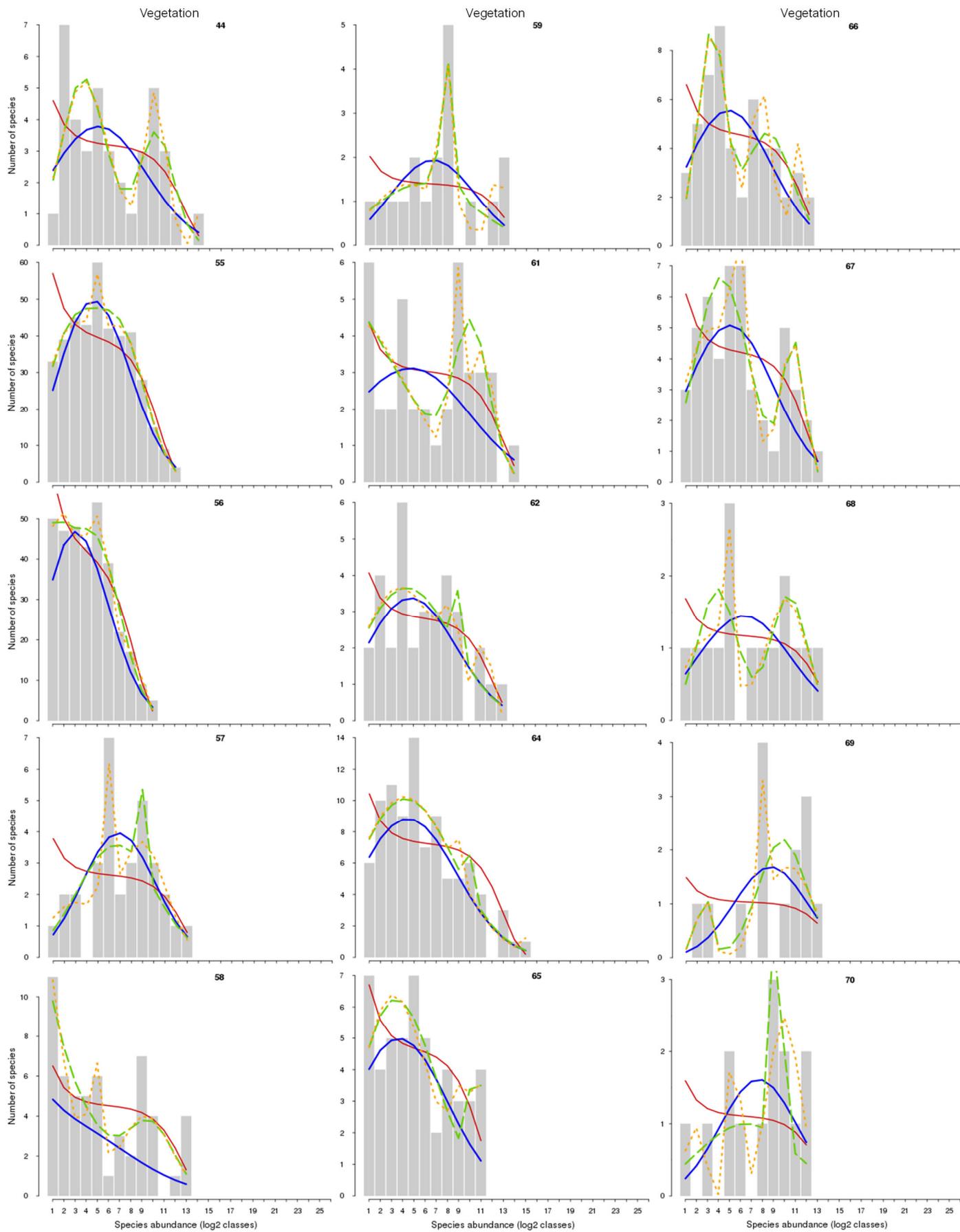


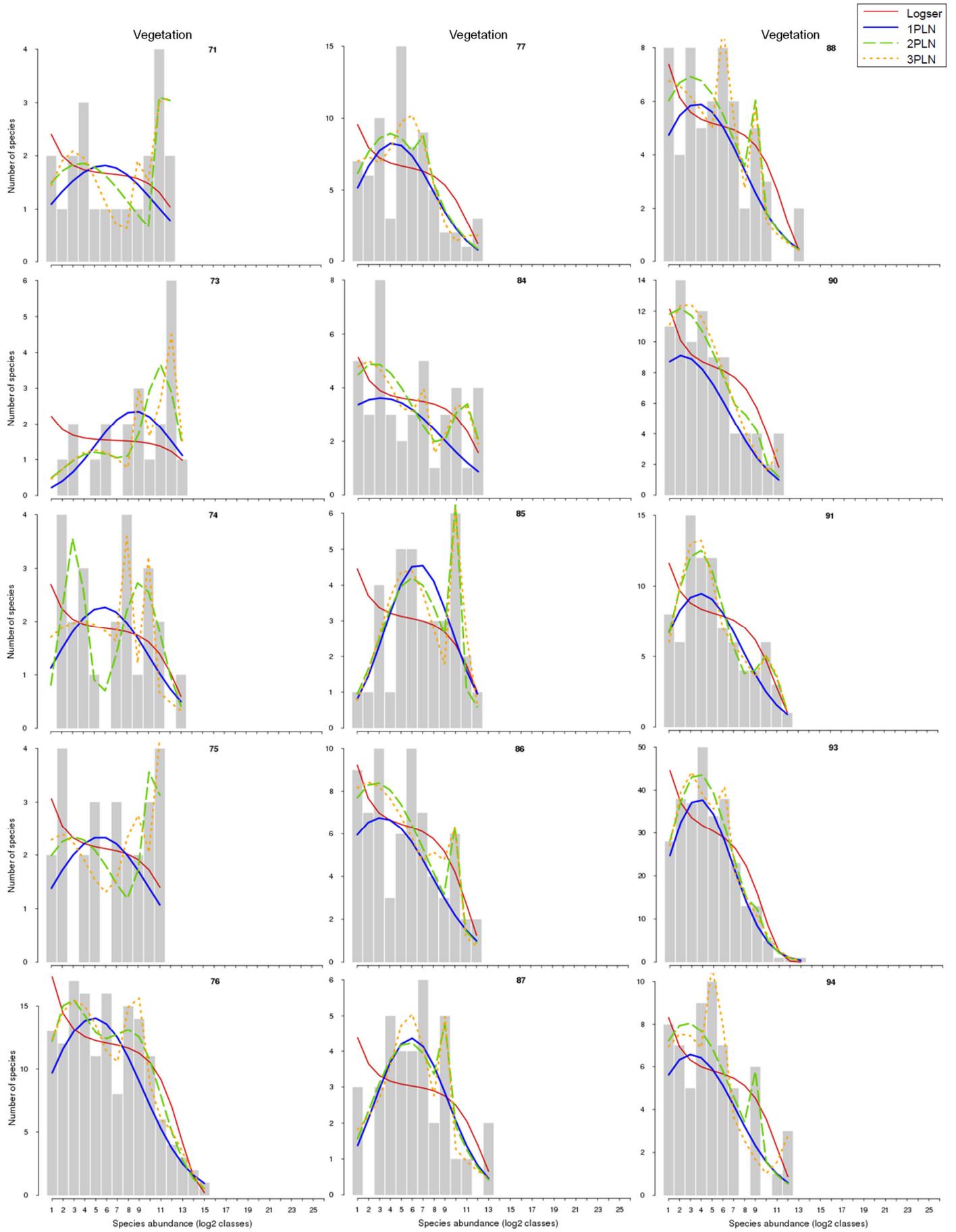
Freshwater

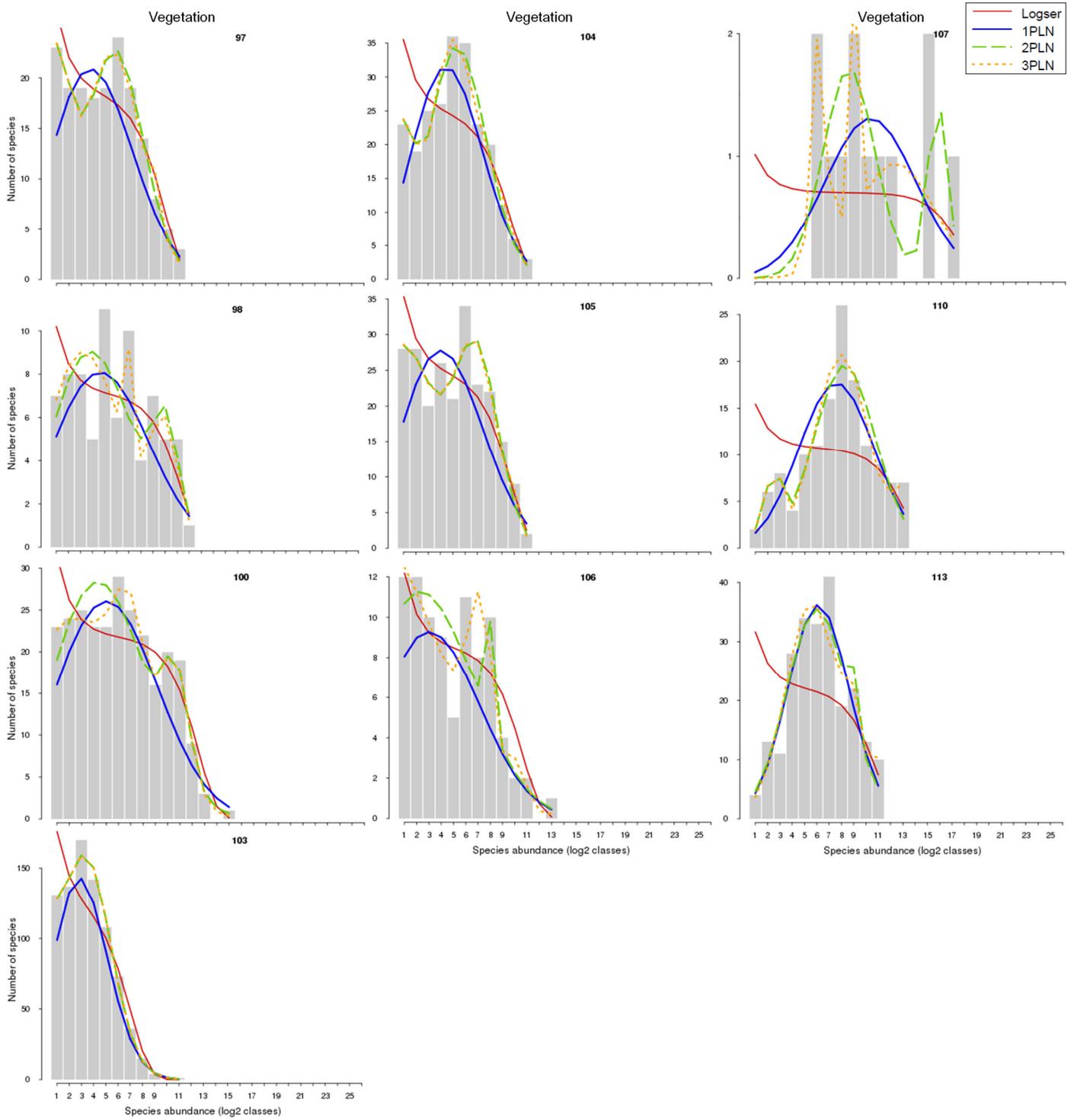


Terrestrial









Appendix VI

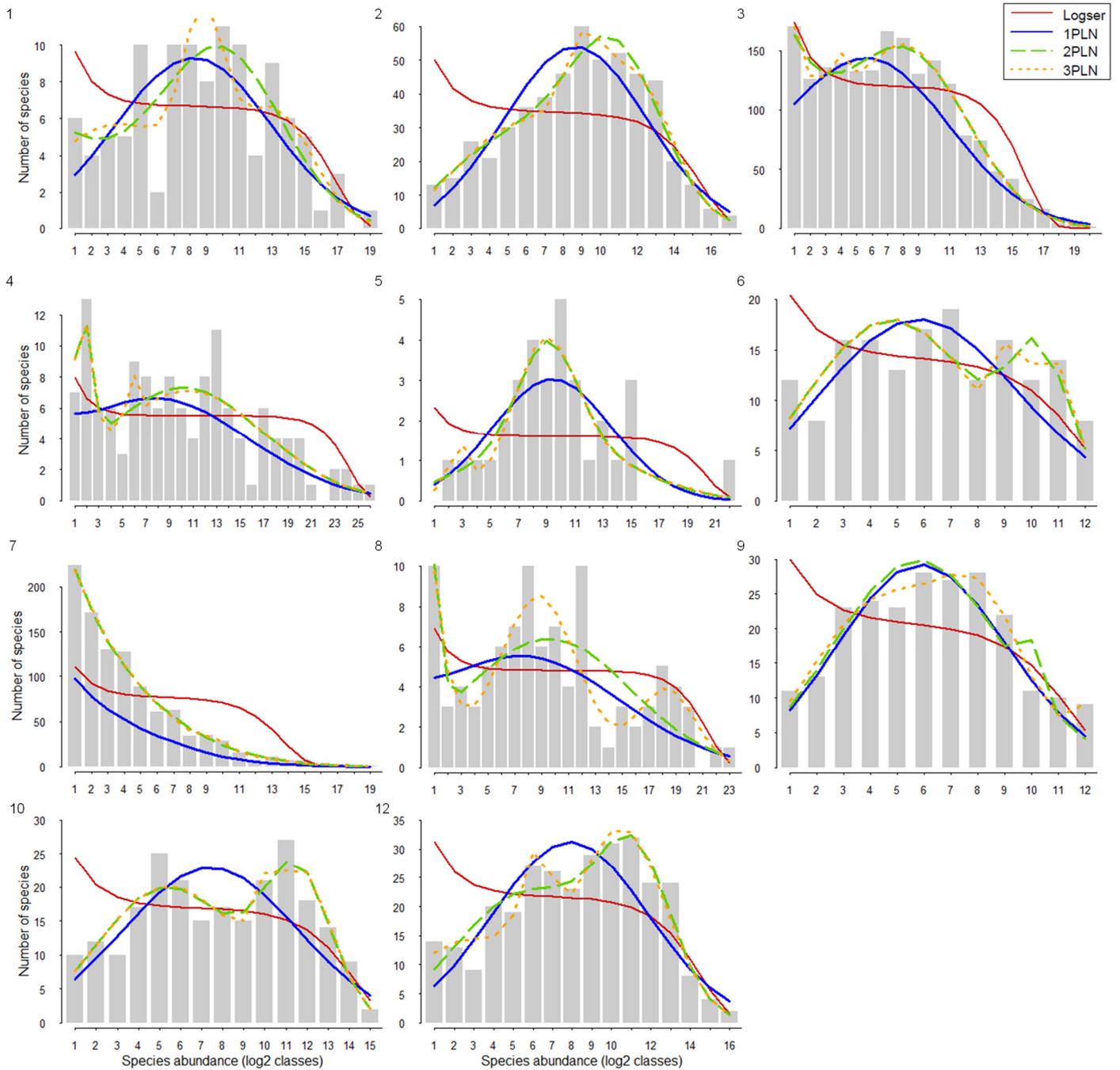
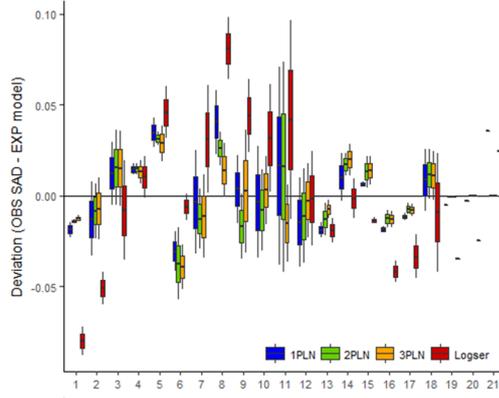
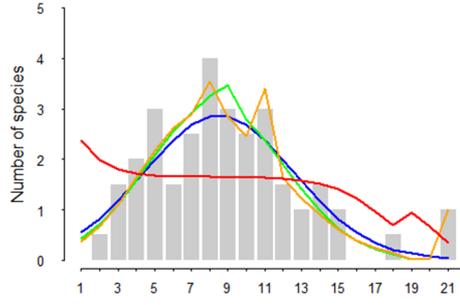


Figure VI.1 Total extent SADs of the empirical datasets analysed in chapter 5, identified by the corresponding ID. The best fitted curves are red line for the logseries, blue line for 1PLN, dashed green line for 2PLN and dotted orange line for 3PLN. All the SADs were better fit by 2PLN or 3PLN at the total extent, except IDs 1, 4, 5 and 7, which were better fit by 1PLN.

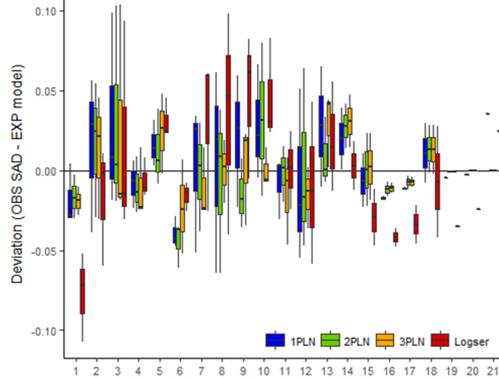
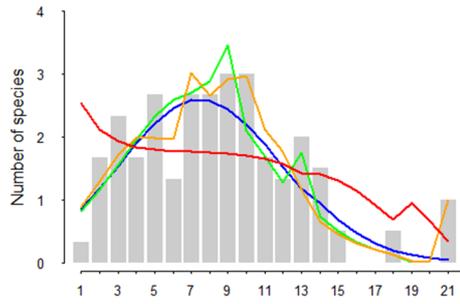
Figure VI.2 Left panel – Comparison of the mean empirical SADs (histograms) and the mean fitted models. Right panel – Deviation between the each empirical SAD and the best fit parameterization of each alternative model; deviations are calculated as the difference between the proportion of species observed and the predicted by each model for each octave of abundance. In both plots, 1PLN is represented in blue, 2PLN in green, 3PLN in orange, and logseries in red. Each community is identified by the corresponding ID (next pages).

ID5

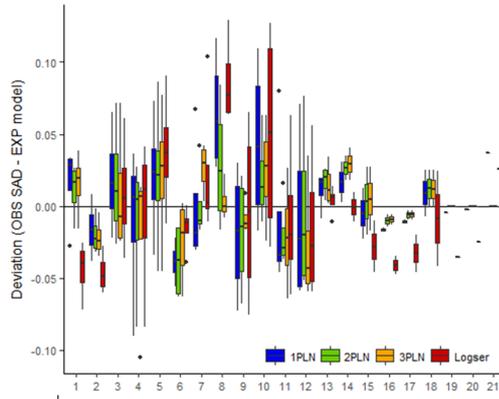
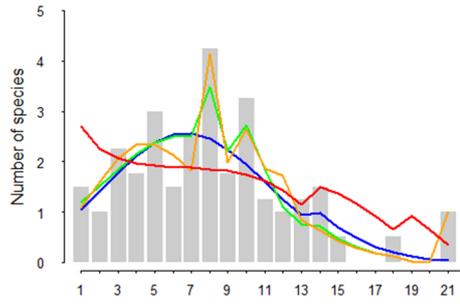
B
i
s
e
c
t
i
o
n



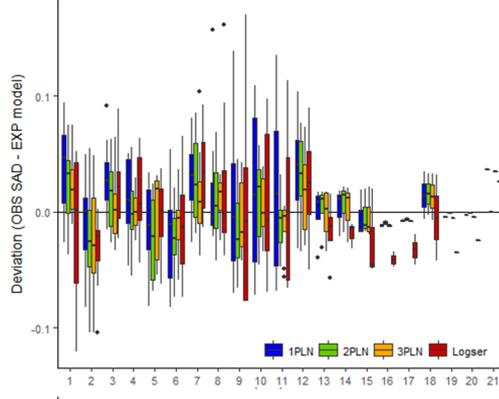
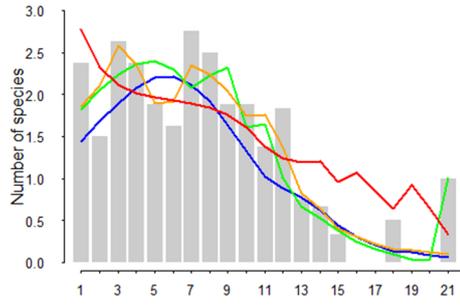
T
h
i
r
d
s



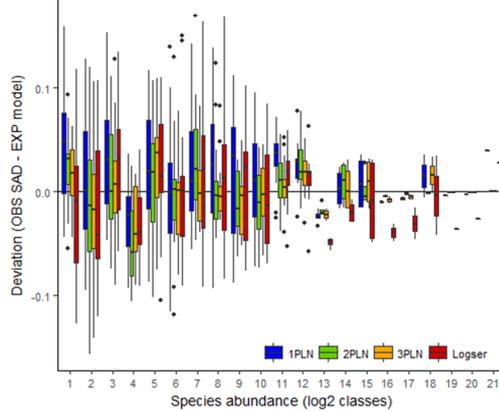
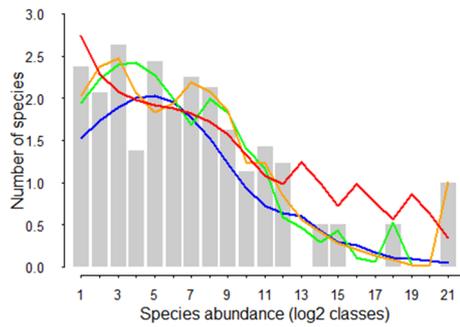
Q
u
a
r
t
e
r
s



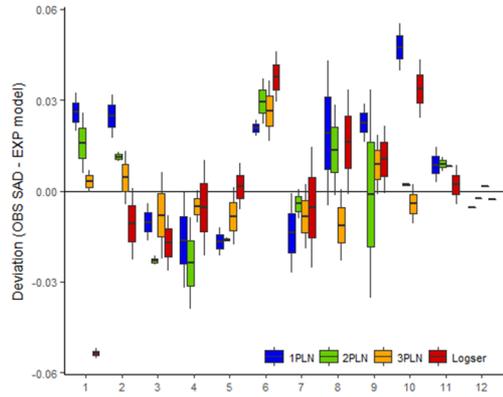
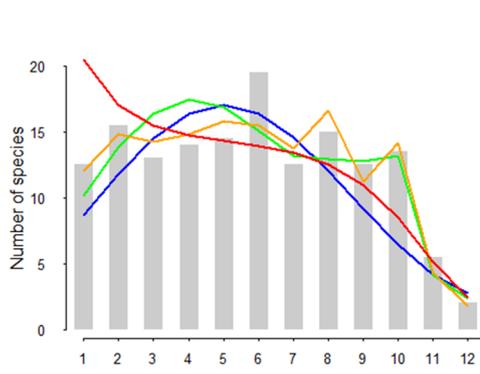
E
i
g
h
t
s



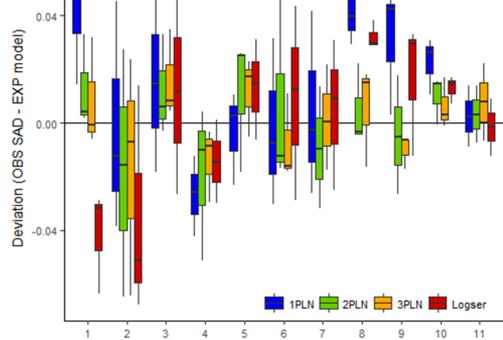
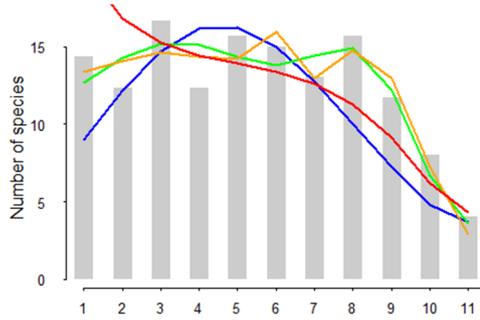
S
i
x
t
e
e
n
t
h
s



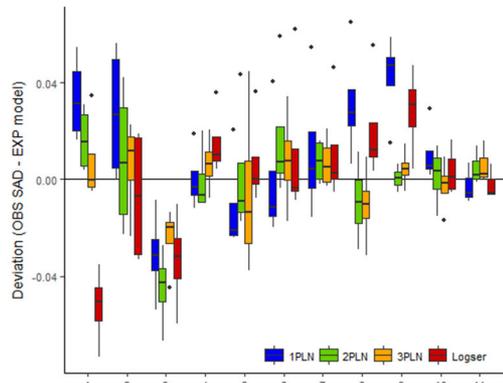
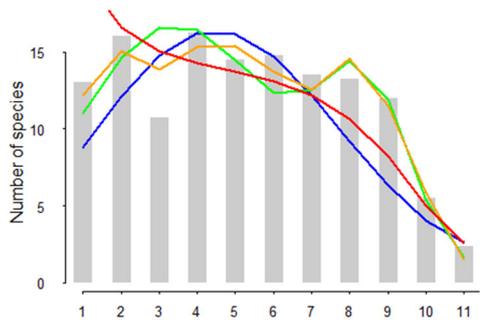
ID6
B
i
s
e
c
t
i
o
n



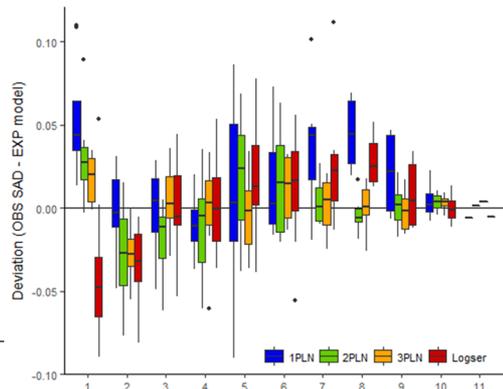
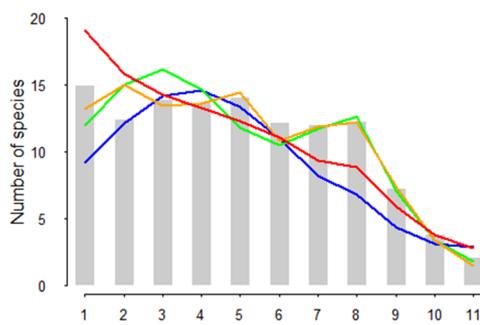
T
h
i
r
d
s



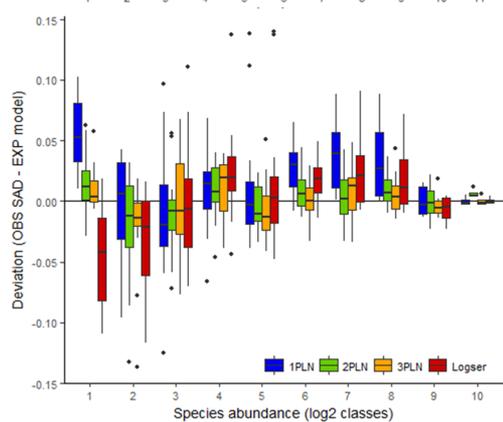
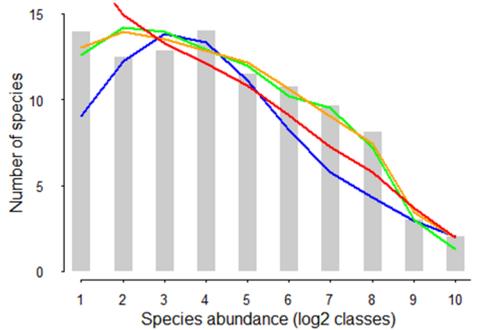
Q
u
a
r
t
e
r
s



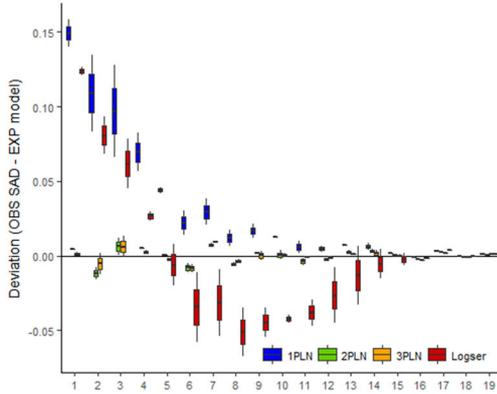
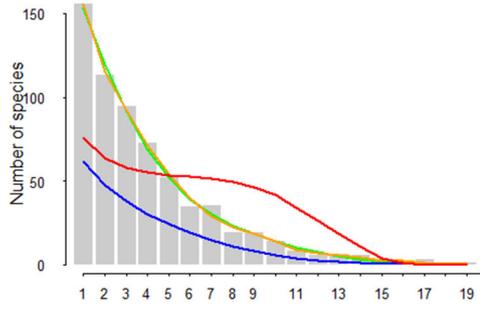
F
i
g
h
t
s



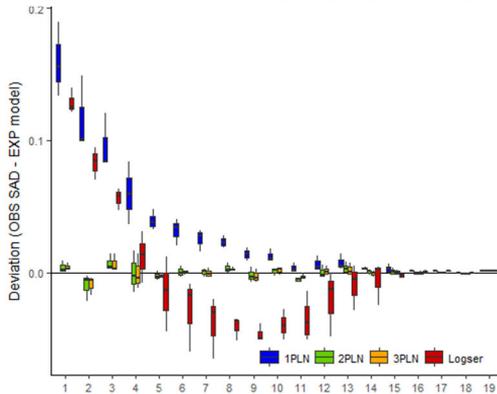
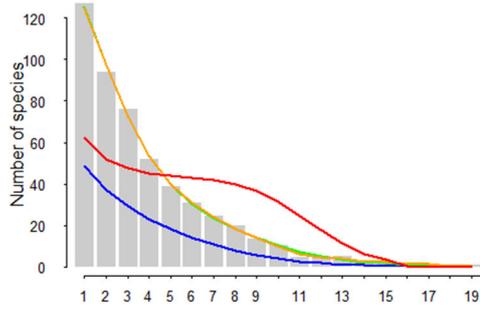
S
i
x
t
e
n
t
h
s



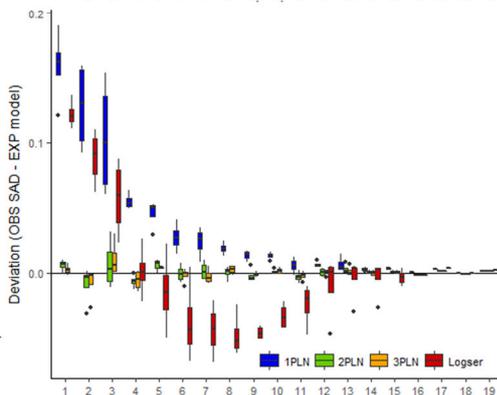
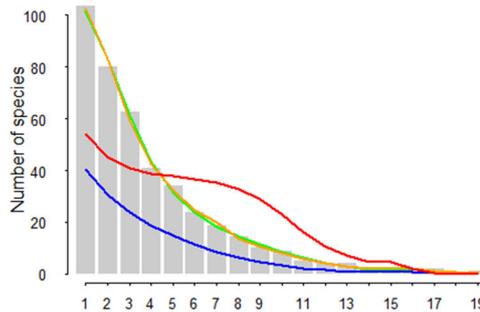
ID7
B
i
s
e
c
t
i
o
n



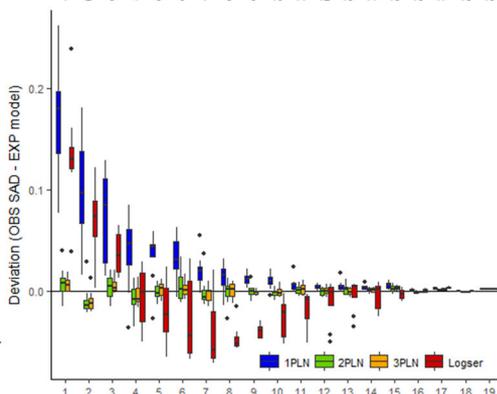
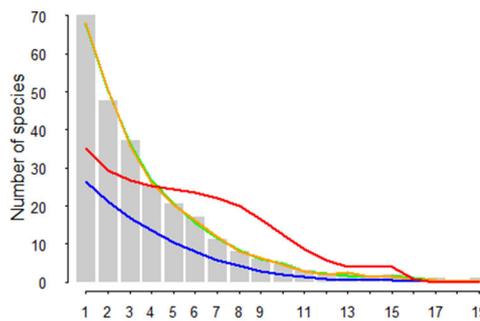
T
h
i
r
d
s



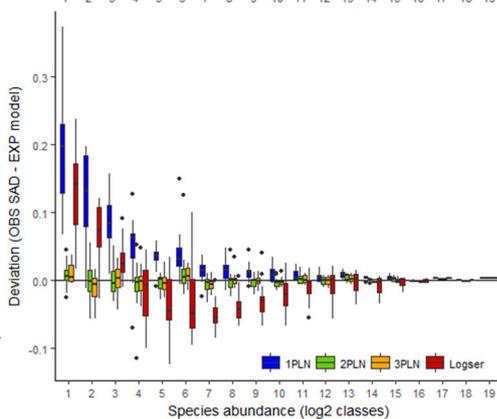
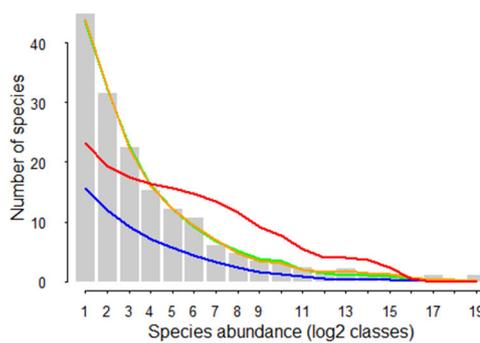
Q
u
a
r
t
e
r
s



E
i
g
h
t
s

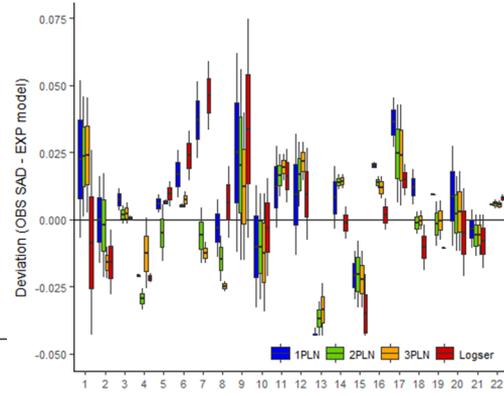
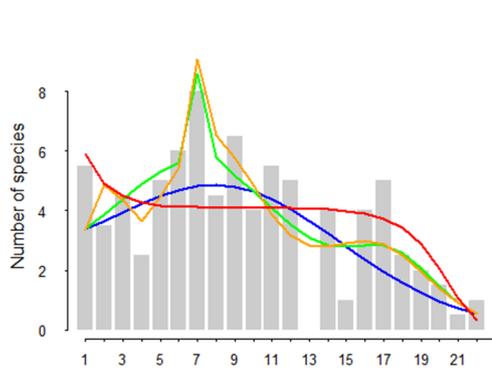


S
i
x
t
e
n
t
h
s

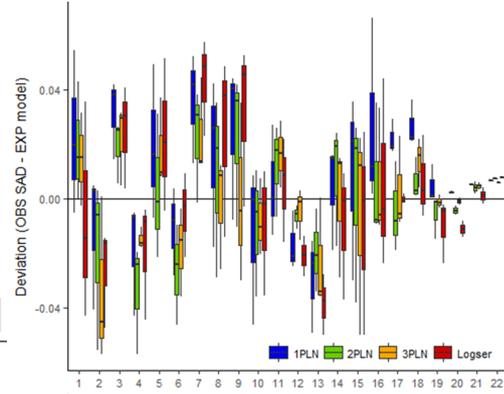
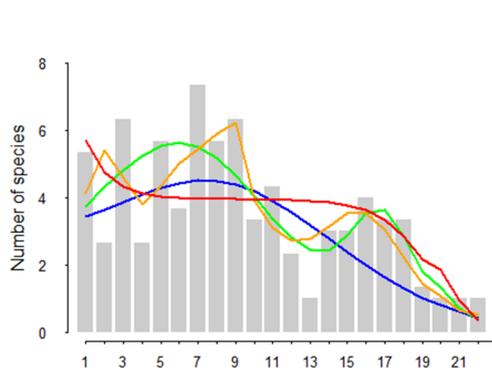


ID8

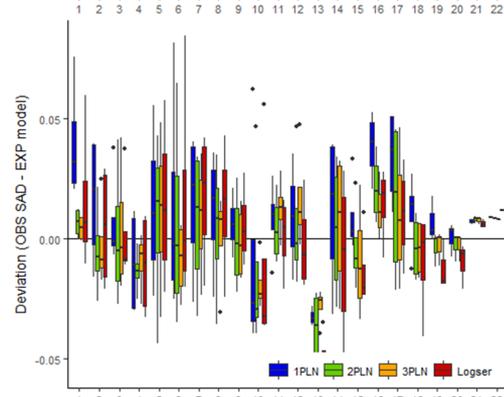
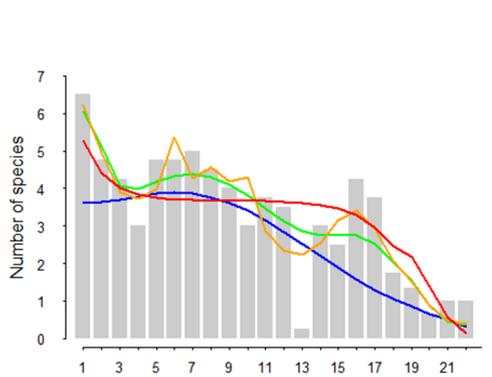
B
i
s
e
c
t
i
o
n



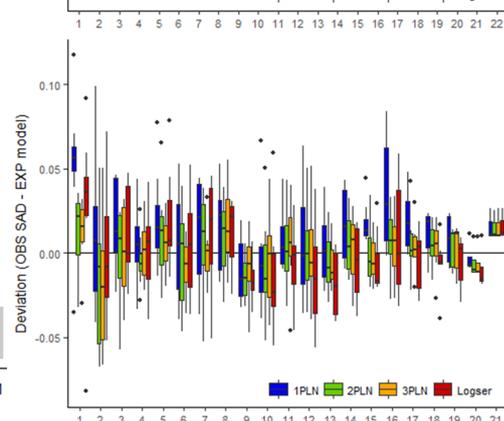
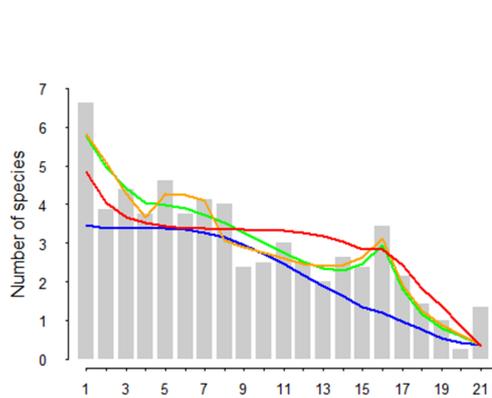
T
h
i
r
d
s



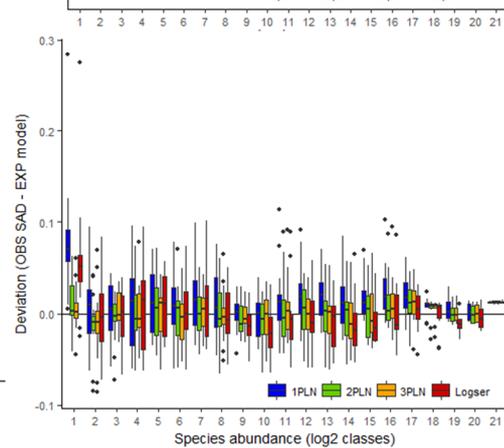
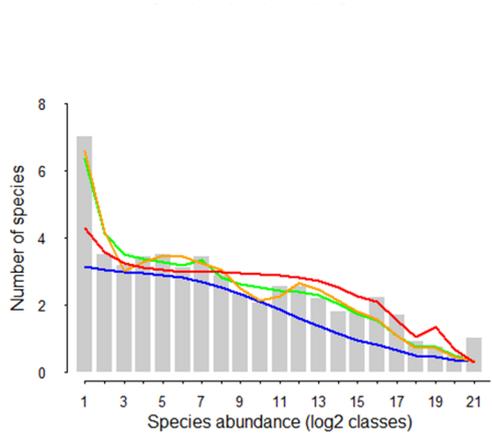
Q
u
a
r
t
e
r
s



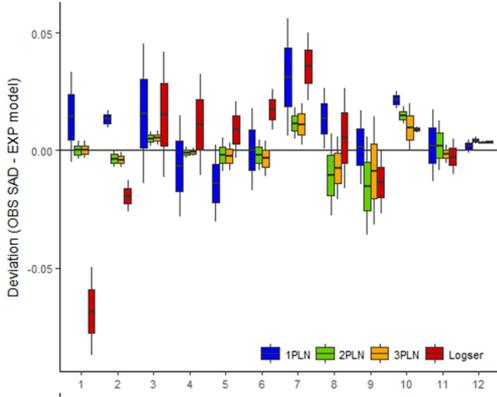
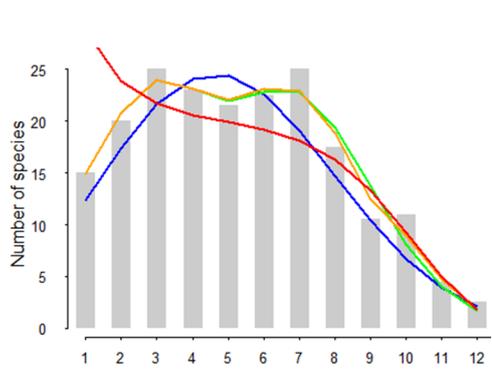
E
i
g
h
t
s



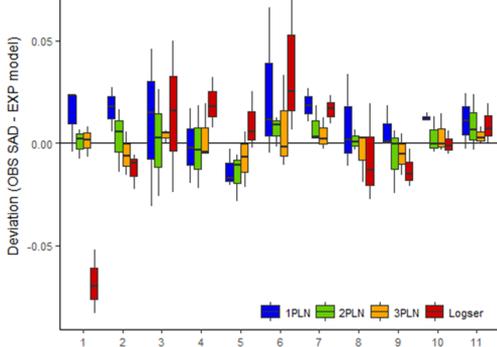
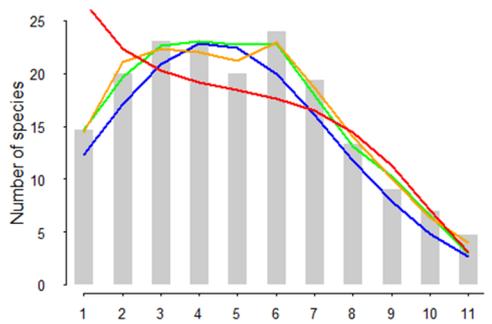
S
i
x
t
e
e
n
t
h
s



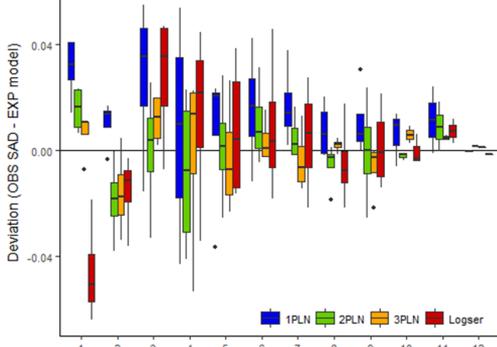
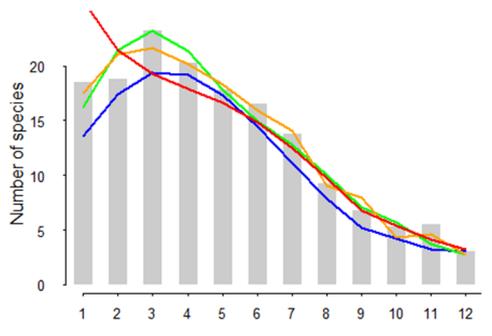
ID9
B
i
s
e
c
t
i
o
n



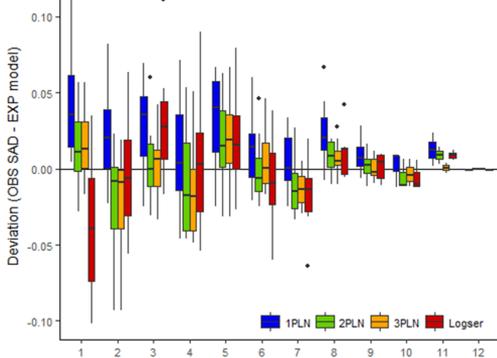
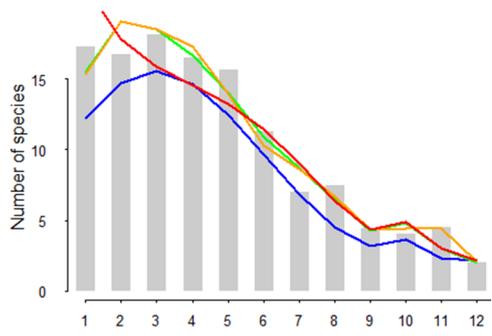
T
h
i
r
d
s



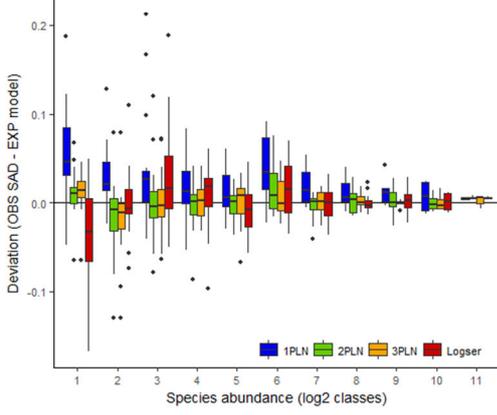
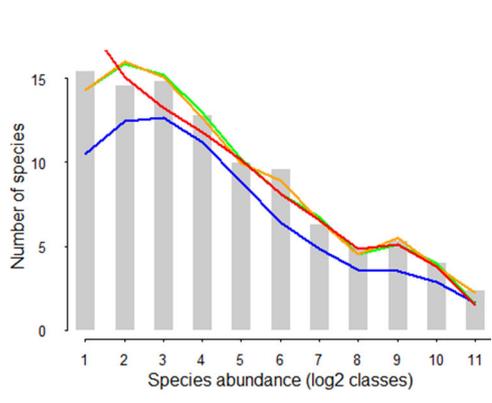
Q
u
a
r
t
e
r
s



F
i
f
t
h
s

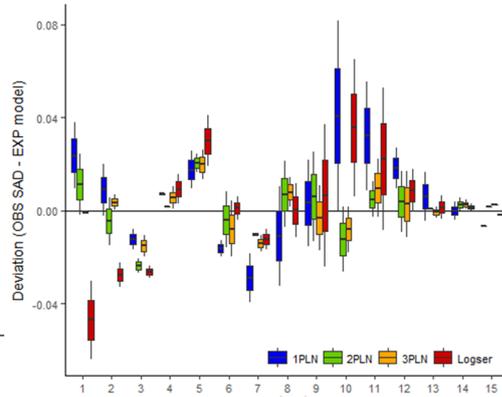
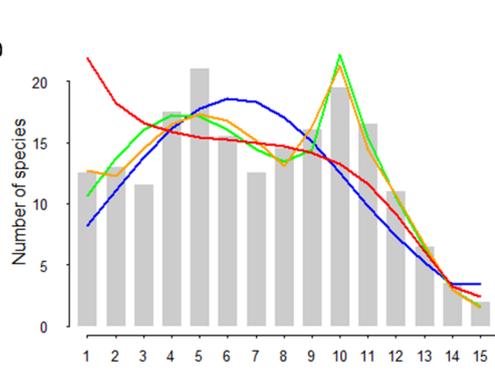


S
i
x
t
e
n
t
h
s

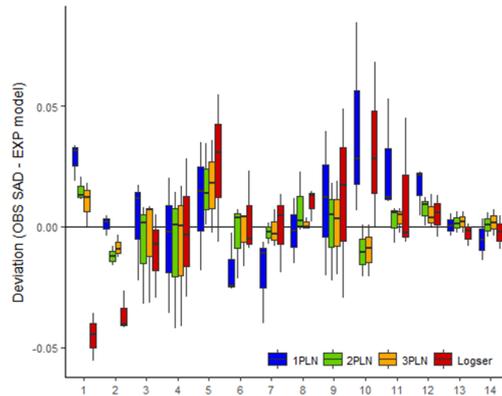
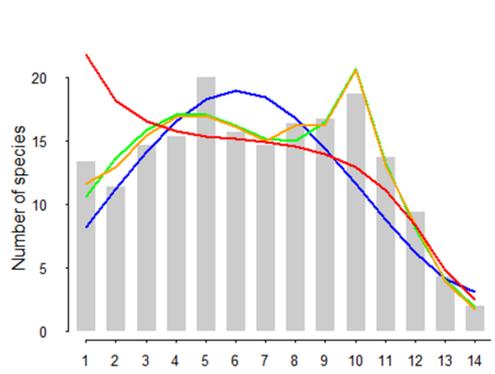


ID10

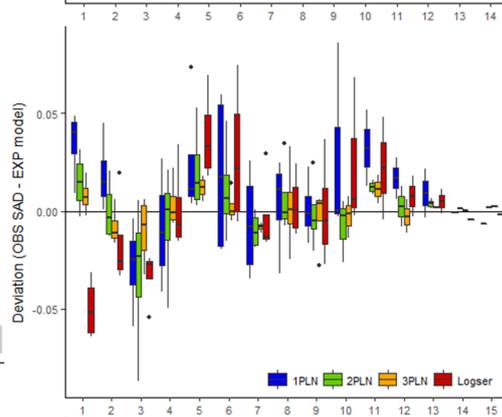
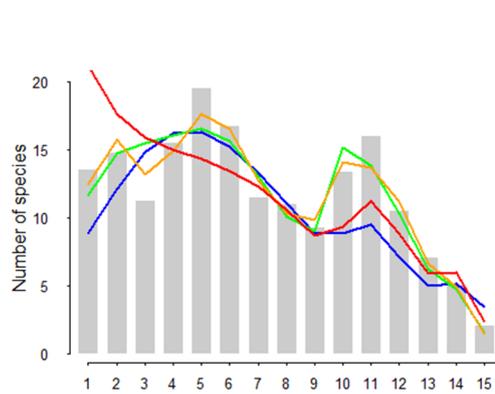
B
i
s
e
c
t
i
o
n



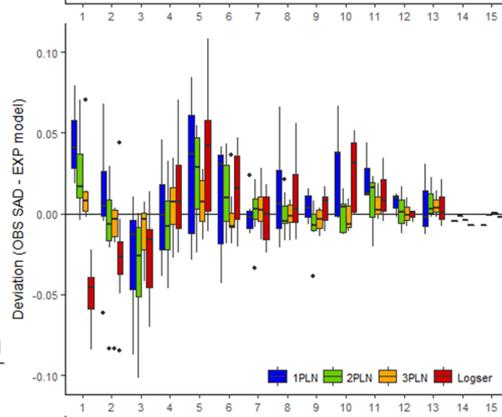
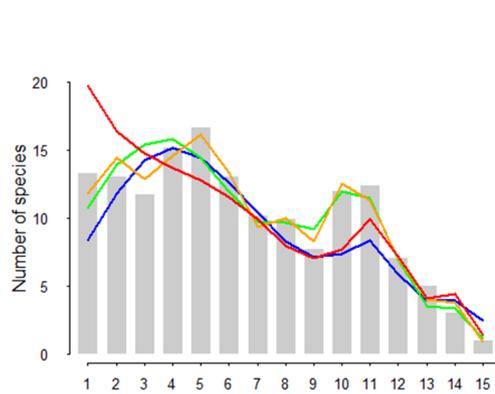
T
h
i
r
d
s



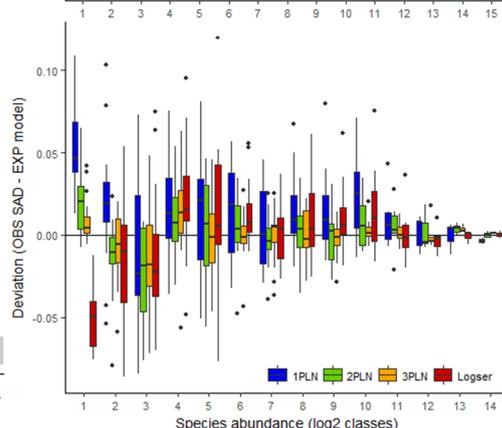
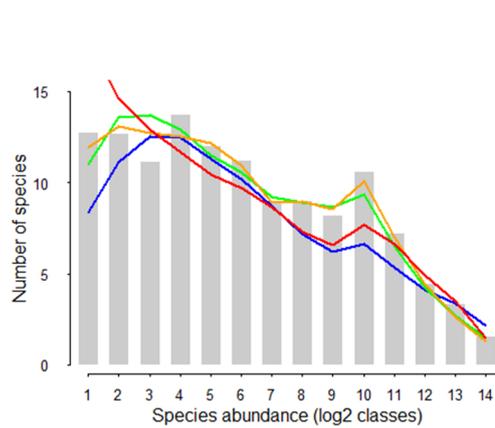
Q
u
a
r
t
e
r
s



E
i
g
h
t
s

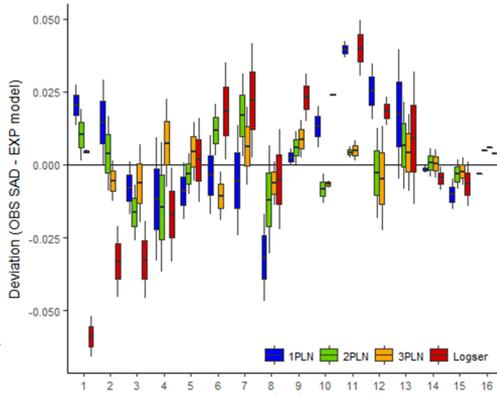
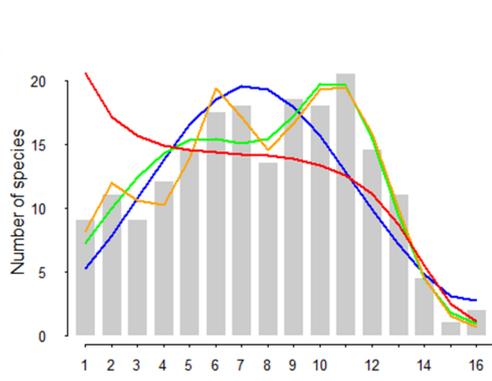


S
i
x
t
e
e
n
t
h
s

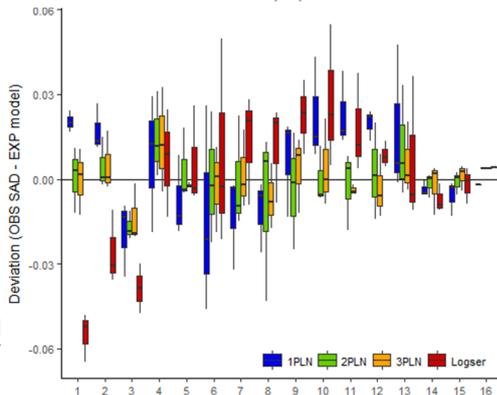
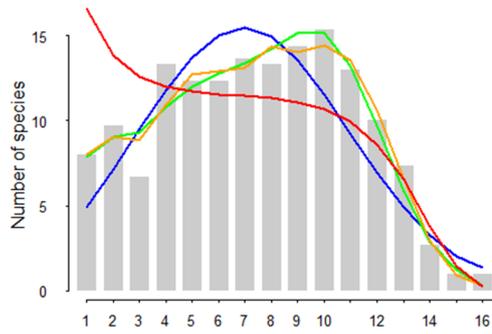


ID12

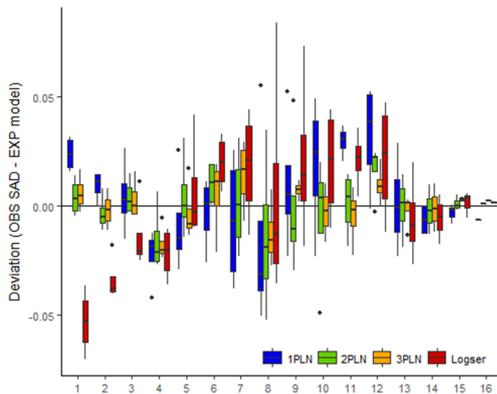
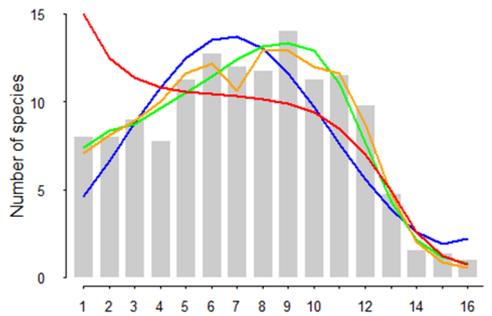
B
i
s
e
c
t
i
o
n



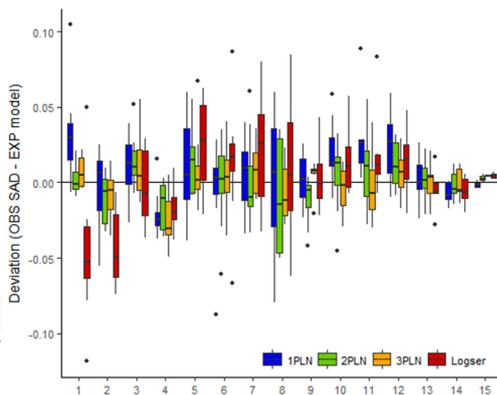
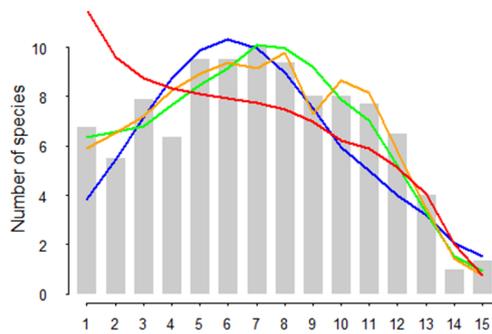
T
h
i
r
d
s



Q
u
a
r
t
e
r
s



F
i
g
h
t
s



S
i
x
t
e
n
t
h
s

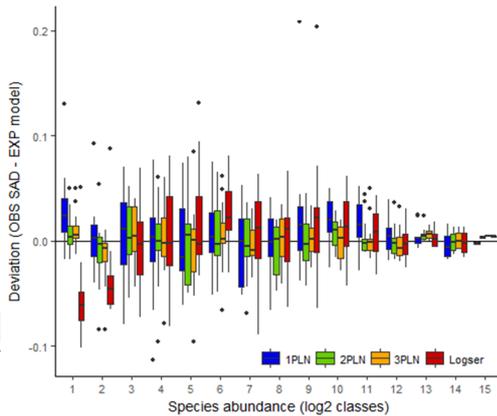
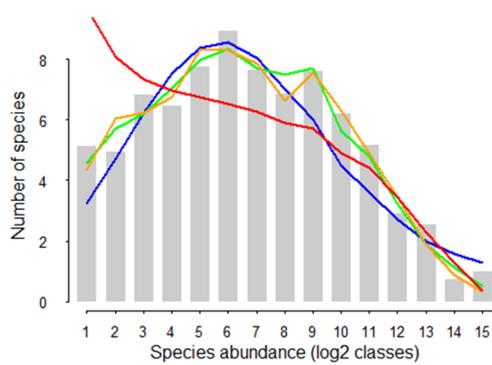


Table VI.1 Results for the linear model fitting of the Shannon index H' as a function of log₁₀(Area), with significant slopes indicated in bold.

Dataset	Slope	Adjusted R ²
1	0.3956	0.1922
2	0.46215	0.5931
3	0.7768	0.4648
4	0.17889	0.06479
5	-0.6784	0.2147
6	0.17315	0.5753
7	0.3145	0.1596
8	0.1878	0.06965
9	0.46838	0.6015
10	0.17273	0.5023
12	0.8702	0.4955

Appendix VII

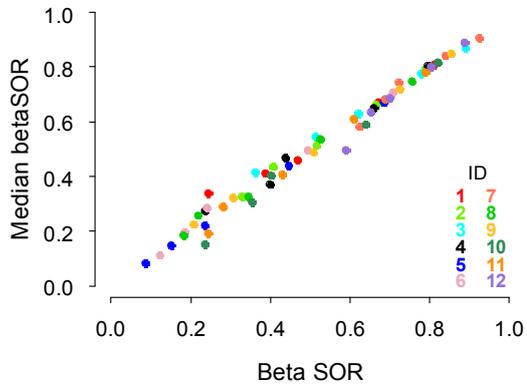
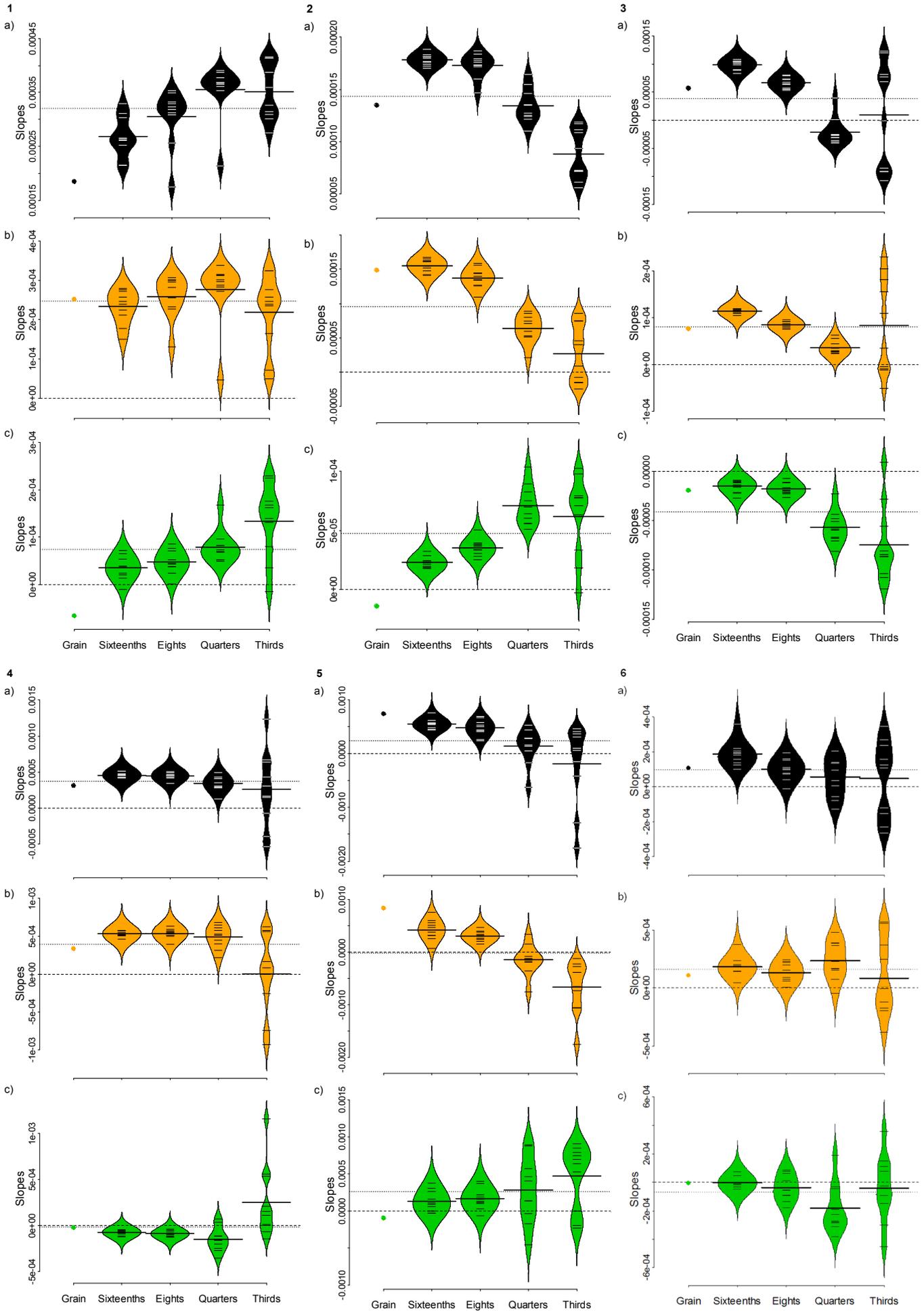


Figure VII.1 Median β_{SOR} values across all the splitting trials (excluding the last one) vs β_{SOR} from a single trial used in the analysis.

Figure VII.2 Variability of estimated DDS linear model slopes across all the random splitting trials, with the scale gradient level on the x axis. Each panel is identified by the corresponding dataset ID, with β_{SOR} on the top (black; a), β_{sim} in the centre (orange; b) and β_{nes} on the bottom (green; c) (next pages).



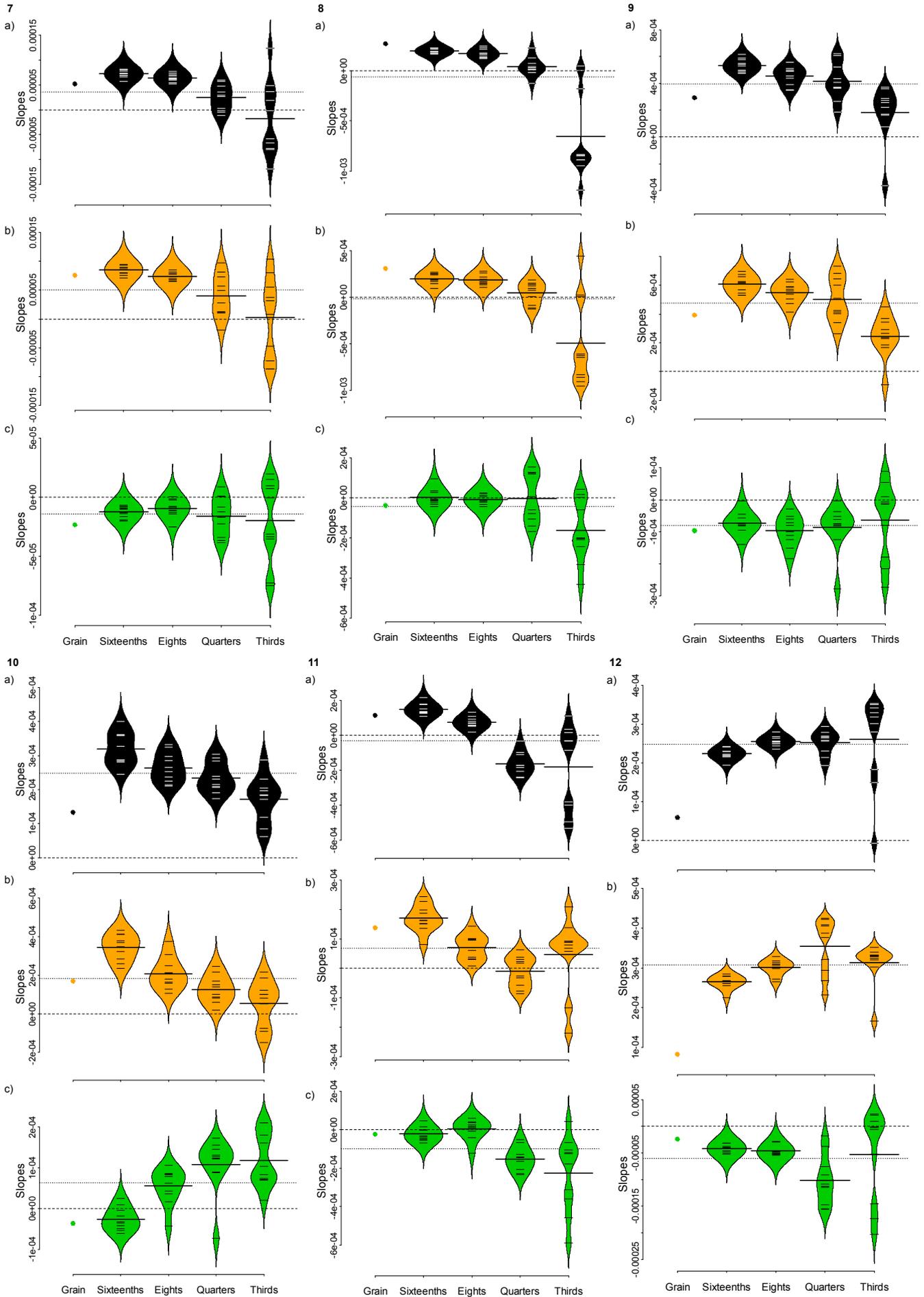


Table VII.1 Distance decay of similarity linear model coefficients from a single trial (used in the analysis) for each β metric; significant estimates are highlighted in bold (see also Figure 6.3).

Dataset ID	Scale level	β_{sor}		β_{sim}		β_{nes}	
		intercept	slope	intercept	slope	intercept	slope
1	Grain	5.98E-01	1.85E-04	4.28E-01	2.51E-04	1.70E-01	-6.62E-05
	1/16	2.96E-01	2.66E-04	2.32E-01	2.45E-04	6.48E-02	2.04E-05
	1/8	1.80E-01	3.09E-04	1.62E-01	2.33E-04	1.78E-02	7.61E-05
	1/4	-3.13E-02	3.78E-04	-2.21E-02	2.82E-04	-9.15E-03	9.52E-05
	1/3	-8.70E-02	4.14E-04	-3.05E-02	2.46E-04	-5.65E-02	1.68E-04
2	Grain	4.25E-01	1.35E-04	3.42E-01	1.49E-04	8.34E-02	-1.39E-05
	1/16	1.65E-01	1.74E-04	1.41E-01	1.42E-04	2.40E-02	3.23E-05
	1/8	1.31E-01	1.84E-04	1.08E-01	1.58E-04	2.21E-02	2.52E-05
	1/4	1.80E-01	1.30E-04	2.13E-01	6.44E-05	-3.36E-02	6.50E-05
	1/3	1.50E-01	1.11E-04	2.15E-01	8.70E-06	-6.54E-02	1.03E-04
3	Grain	7.88E-01	5.72E-05	7.13E-01	7.65E-05	7.48E-02	-1.94E-05
	1/16	5.80E-01	8.31E-05	4.59E-01	1.05E-04	1.21E-01	-2.16E-05
	1/8	5.02E-01	6.54E-05	3.40E-01	8.68E-05	1.63E-01	-2.14E-05
	1/4	6.76E-01	-3.86E-05	3.70E-01	2.97E-05	3.06E-01	-6.83E-05
	1/3	9.92E-02	1.23E-04	-2.77E-01	2.31E-04	3.76E-01	-1.08E-04
4	Grain	3.84E-01	3.12E-04	2.67E-01	3.39E-04	1.17E-01	-2.71E-05
	1/16	2.47E-01	4.24E-04	1.21E-01	5.21E-04	1.26E-01	-9.61E-05
	1/8	2.18E-01	4.22E-04	1.13E-01	5.13E-04	1.06E-01	-9.00E-05
	1/4	7.57E-02	4.92E-04	-4.45E-03	6.10E-04	8.01E-02	-1.18E-04
	1/3	1.18E-01	4.30E-04	3.82E-02	5.76E-04	7.98E-02	-1.46E-04
5	Grain	3.75E-01	7.34E-04	2.57E-01	8.32E-04	1.18E-01	-9.78E-05
	1/16	1.84E-01	5.72E-04	1.31E-01	3.13E-04	5.25E-02	2.59E-04
	1/8	1.37E-01	4.12E-04	9.65E-02	2.50E-04	4.01E-02	1.62E-04
	1/4	6.58E-02	5.27E-04	1.00E-01	-3.52E-04	-3.46E-02	8.79E-04
	1/3	1.74E-01	-4.23E-04	9.85E-02	-2.28E-04	7.52E-02	-1.95E-04
6	Grain	7.19E-01	1.06E-04	6.34E-01	1.12E-04	8.53E-02	-6.38E-06
	1/16	2.38E-01	1.94E-04	1.90E-01	1.96E-04	4.75E-02	-1.19E-06
	1/8	2.12E-01	4.23E-05	1.35E-01	2.22E-04	7.64E-02	-1.80E-04
	1/4	9.42E-02	2.03E-04	-1.83E-02	4.81E-04	1.13E-01	-2.78E-04
	1/3	1.92E-01	-2.66E-04	1.43E-01	-1.72E-04	4.84E-02	-9.42E-05
7	Grain	8.03E-01	5.16E-05	7.10E-01	7.53E-05	9.36E-02	-2.37E-05
	1/16	6.54E-01	7.41E-05	5.65E-01	9.35E-05	8.90E-02	-1.94E-05
	1/8	5.68E-01	6.79E-05	5.20E-01	7.02E-05	4.74E-02	-2.36E-06
	1/4	4.85E-01	5.70E-05	4.64E-01	5.66E-05	2.14E-02	3.85E-07
	1/3	5.52E-01	3.68E-05	5.25E-01	3.75E-05	2.76E-02	-7.00E-07

8	Grain	3.72E-01	2.69E-04	2.79E-01	3.08E-04	9.23E-02	-3.90E-05
	1/16	3.01E-01	1.74E-04	2.39E-01	1.85E-04	6.15E-02	-1.13E-05
	1/8	1.86E-01	2.25E-04	1.24E-01	2.73E-04	6.25E-02	-4.82E-05
	1/4	2.14E-01	-4.53E-05	1.52E-01	-4.57E-07	6.20E-02	-4.49E-05
	1/3	2.63E-01	-2.56E-04	1.46E-01	-1.63E-05	1.17E-01	-2.40E-04
9	Grain	7.30E-01	2.95E-04	6.41E-01	3.91E-04	8.88E-02	-9.59E-05
	1/16	4.11E-01	6.15E-04	3.17E-01	6.97E-04	9.40E-02	-8.18E-05
	1/8	3.98E-01	3.52E-04	3.06E-01	4.13E-04	9.17E-02	-6.07E-05
	1/4	2.43E-01	3.72E-04	2.00E-01	4.23E-04	4.35E-02	-5.15E-05
	1/3	1.20E-01	2.74E-04	1.10E-01	2.65E-04	1.02E-02	9.61E-06
10	Grain	7.58E-01	1.33E-04	6.87E-01	1.69E-04	7.10E-02	-3.61E-05
	1/16	3.44E-01	2.45E-04	3.07E-01	2.18E-04	3.70E-02	2.74E-05
	1/8	1.84E-01	1.87E-04	9.24E-02	2.00E-04	9.13E-02	-1.28E-05
	1/4	1.82E-01	1.23E-04	4.47E-02	1.90E-04	1.37E-01	-6.66E-05
	1/3	4.00E-01	-2.44E-04	3.81E-01	-2.89E-04	1.86E-02	4.48E-05
11	Grain	5.18E-01	1.13E-04	3.38E-01	1.37E-04	1.80E-01	-2.42E-05
	1/16	3.36E-01	1.54E-04	1.63E-01	2.07E-04	1.73E-01	-5.29E-05
	1/8	3.15E-01	5.97E-05	2.10E-01	5.16E-05	1.06E-01	8.06E-06
	1/4	4.20E-01	-1.87E-04	2.09E-01	-3.49E-05	2.11E-01	-1.52E-04
	1/3	2.90E-01	-8.53E-05	5.14E-02	8.96E-05	2.38E-01	-1.75E-04
12	Grain	8.38E-01	5.83E-05	7.70E-01	8.24E-05	6.80E-02	-2.42E-05
	1/16	4.02E-01	2.30E-04	2.71E-01	2.61E-04	1.32E-01	-3.11E-05
	1/8	2.51E-01	2.64E-04	5.46E-02	3.14E-04	1.97E-01	-5.05E-05
	1/4	1.36E-01	2.79E-04	-1.77E-01	3.87E-04	3.13E-01	-1.08E-04
	1/3	4.33E-02	2.81E-04	-1.16E-01	3.33E-04	1.59E-01	-5.24E-05

Table VII.2 Comparison of the DDS slopes and intercepts between the scaling levels, from the coefficients bootstrap distributions; significantly different comparisons are noted as ***.

ID	Levels compared	β_{sor}		β_{sim}		β_{nes}	
		slope	intercept	slope	intercept	slope	intercept
1	Grain > 1/16	---	***	---	***	---	***
	1/16 > 1/8	---	***	---	***	---	***
2	Grain > 1/16	---	***	---	***	---	***
	1/16 > 1/8	---	---	---	---	---	---
3	Grain > 1/16	---	***	---	***	---	---
	1/16 > 1/8	---	---	---	---	---	---
4	Grain > 1/16	---	***	---	***	***	---
	1/16 > 1/8	---	---	---	---	---	---
5	Grain > 1/16	---	***	***	***	---	***
	1/16 > 1/8	---	---	---	---	---	---
6	Grain > 1/16	---	***	---	***	---	***
	1/16 > 1/8	---	---	---	***	***	---
7	Grain > 1/16	---	***	---	***	---	---
	1/16 > 1/8	---	---	---	---	---	---
8	Grain > 1/16	---	***	***	---	---	***
	1/16 > 1/8	---	***	---	***	---	---
9	Grain > 1/16	---	***	---	***	---	---
	1/16 > 1/8	---	---	---	---	---	---
10	Grain > 1/16	---	***	---	***	---	***
	1/16 > 1/8	---	***	---	***	---	---
11	Grain > 1/16	---	***	---	***	---	---
	1/16 > 1/8	---	---	***	---	---	***
12	Grain > 1/16	---	***	---	***	---	---
	1/16 > 1/8	---	***	---	***	---	---