# Quantifying biodiversity for valuation

Olga Lyashevska

MSc Environmental Sciences

A thesis submitted in fulfilment of the requirement for the degree of
**Doctor of Philosophy**

School of Biological Sciences
Faculty of Medicine, Health and Life Science

Queen's University Belfast

November 2011

# Acknowledgements

This research could not have been completed without support of a number of people, to which I am sincerely grateful.

First and foremost, I am greatly indebted to Dr. Keith Farnsworth for being a supervisor who believes, inspires, and gives freedom. His guidance towards objective science has helped me to shape and refine my thinking not only about biodiversity, but also life in general. Our mind-broadening discussions over a cup of coffee were particularly invaluable to me. Thank you, Keith, for keeping me on track, putting up with my obscure drafts, and all my technical inclinations. This thesis would never be completed without you!

I am also grateful to all my friends for making me feel here in Belfast like at home. Many thanks to everyone in 6.16 (and beyond) for a stimulating working atmosphere and our regular lunch breaks that would help me coping with mid-afternoon dips and boost my productivity. Very special thanks go to Tak Fung for being my first (and very pedantic) reader.

Discipline and dedication are essential to a martial artist. Being one, I dare to say that it is a life-changing experience which gave me resilience that I desperately needed on a daily basis to work away on my thesis and being able to complete it. For this, I want to thank all my coaches – both physical and mental for their encouragements to become a 'fighter'. You taught me how to achieve what I want and stay strong while going beyond my comfort zone.

Last and most significantly, I wish to express my love and appreciation for my family. Without your support and belief in me throughout all these years being away from home I would never get that far.

This research was supported by Science, Technology, Research and Innovation for the Environment (STRIVE) grant from the Environmental Protection Agency of the Republic of Ireland: 2007-PhD-SD-3. Many thanks to BioMar and BIOTIC teams for kindly offering biodiversity source data that I could use for simulations.

# Abstract

Biodiversity, arising at multiple levels, is known as a multi-dimensional and complex concept, but is also has a rather loose definition. Imprecise definitions are not very suitable for objective quantification or the rigour of economic valuation.

Therefore, to construct a more substantial definition of value for biodiversity, a theoretical argument aiming to link biodiversity and functional (meaningful) information needs to be developed. A working hypothesis is that biodiversity is a measure of the total difference within a biological system, which can be summarised in terms of the system's total information content, of which functional information is a subset. Since functional information has systematic (non-random) patterns, it coincides with the scientific meaning of biological complexity, thus providing the foundation of value in biodiversity.

The theory presented sets the goal of estimating biological complexity from the potentially valuable information derived from empirical biodiversity metric data (ecological measures). To achieve this, the ecological properties of a system, as they are measured by ecologists, were translated into a simply defined single valued property. This led to a conclusion that if there exists a systematic relationship among empirical biodiversity metrics, then there must be a unifying property underlying intrinsic value of biodiversity.

Then, an advantage of a representation of biodiversity as information was demonstrated by comparing it with the most commonly used metric – species richness. It was shown that species richness missed a large proportion of diversity, emphasising the importance of additional ecological properties and the need for species databases to record functional traits, presence, and abundances in communities, as well as phylogenetic information.

Finally, by providing intellectual foundations and developing an analytical tool for biodiversity quantification, this study sets the goal for further research. An advantage of the approach in this study to economic valuation is that value is based on real, measurable, and intrinsic properties of systems, such that it is objective in contrast with present opinion-based economic methods applied to biodiversity.

# Contents

List of Tables v

List of Figures vi

List of Symbols ix

List of Acronyms and Abbreviations xi

1 Why another Study of Biodiversity? 1

1.1 Biodiversity: important but confusing . . . 2

1.2 Sources of value . . . 6

1.3 Biodiversity as information? . . . 9

2 What (really) is biodiversity? 12

2.1 Setting out the problem . . . 12

2.2 Methods . . . 14

2.3 The nuts and bolts of biodiversity . . . 16

2.4 Decomposition of biodiversity . . . 25

2.5 A formal structure . . . 31

2.6 Diversity in space . . . 32

2.7 Discussion and Conclusions . . . 35

3 Knowledge engineering biodiversity 38

3.1 Introduction . . . 38

3.2 Methods . . . 42

3.3 A worked example: of the database . . . 47

3.4 Biodiversity data mining: Results . . . 52

3.5 Discussion and Conclusions . . . 61

Appendices 65

3.A Documentation of studies included in database of measures . . . 65

3.B Biodiversity data model . . . 74

3.C Database biodivDB (SQL dump) . . . 75

3.D Data dictionary . . . 79

3.E Relational cardinalities . . . . . . . . . . . . . . . . . . . . . . . . . . . . 80
3.F Biodiversity data mining . . . . . . . . . . . . . . . . . . . . . . . . . . . . 81

4 A simulation study **85**
4.1 Why simulation? . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 85
4.2 Empirical distributional properties . . . . . . . . . . . . . . . . . . . . . . 89
4.3 Building a model for simulating communities . . . . . . . . . . . . . . . . 98
4.4 Discussion and Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . 103

Appendices **108**
4.A Simulation of hypothetical community . . . . . . . . . . . . . . . . . . . . 108

5 Biodiversity metric: multum in parvo **110**
5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 110
5.2 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 112
5.3 Modelling results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 120
5.4 Discussion and Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . 128

Appendices **130**
5.A Community structure and composition indices . . . . . . . . . . . . . . . . 130
5.B Taxonomic diversity indices . . . . . . . . . . . . . . . . . . . . . . . . . . 132
5.C Functional diversity indices . . . . . . . . . . . . . . . . . . . . . . . . . . 133

6 Implications: from theory to practice **134**
6.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 134
6.2 The history of Global action on biodiversity . . . . . . . . . . . . . . . . . 135
6.3 Summarising the contribution of this study . . . . . . . . . . . . . . . . . 141
6.4 De-emphasising abundance . . . . . . . . . . . . . . . . . . . . . . . . . . 145
6.5 Returning to the economic context . . . . . . . . . . . . . . . . . . . . . . 147
6.6 Further Progress of the idea . . . . . . . . . . . . . . . . . . . . . . . . . . 151

References **155**

# List of Tables

2.1 A nine-level hierarchy of biocomplexity. Left column names the level of organization and right column gives examples of the complex interactions and processes that take place at that level, contributing to biocomplexity. Complexity is also added by interactions among levels, both upwards and downwards, producing feedback circuits. Note: at the top level, one may place the planetary bio-geochemical regulatory system. . . . . . . . . . . . . 21

3.1 Types and sources of variation in biodiversity estimate. Adapted from Osenberg et al. (1999) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 40

3.2 Online databases being searched for biodiversity literature . . . . . . . . . . 42

3.3 Classification, type, and description of the ERD entities in biodivDB database 49

3.4 Relationship cardinalities in biodiversity model . . . . . . . . . . . . . . . . . 50

3.5 Summary of the database of measures of biodiversity (1995-2009) . . . . . . 53

3.6 Summary of biodiversity estimates across studies included in biodivDB . . . 53

3.7 Summary of the variables used in GLM . . . . . . . . . . . . . . . . . . . . . 57

4.1 "Hillis" taxonomic sampling schemes. Adapted from Hillis (1998) . . . . . . 88

4.2 An example of functional traits and their possible values obtained to calculate functional diversity indices . . . . . . . . . . . . . . . . . . . . . . . . . 90

4.3 The unit variation at different levels in taxonomically structured species database (ITIS) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 91

4.4 Models parameters for rank abundance distribution plot . . . . . . . . . . . 103

5.1 Groups of indices and measures of biodiversity . . . . . . . . . . . . . . . . . 113

5.2 Correlation between measures of biodiversity calculated on 1000 simulated ecological communities. Three groups of measures/indicators can be distinguished: (A) community structure and composition – Richness, Pielou, Jaccard, Sorensen, Chao-Jaccard, Chao-Sorensen, Shannon, Turnover, and Abundance; (B) taxonomic diversity – D, Dstar, Lambda, Dplus, SDPlus; and (C) functional diversity – FEve, FDiv, FDis, RaoQ . . . . . . . . . . . 121

5.3 Rotation matrix: correlations of each of the variables with the first principal axes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 123

# List of Figures

1.1 Theoretical models of relationship between biodiversity and ecosystem functioning. Adapted from Danovaro and Pusceddu (2007) . . . . . . . . . . . . 8

2.1 Conceptual analysis of biodiversity. Sequences of inferences (represented by arrows) applied to an information-based definition of biodiversity suggest a link between biodiversity and ecosystem services . . . . . . . . . . . . . . . 15

2.2 The notion of difference in the concept of biodiversity illustrated using the digital image analogy . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17

2.3 Decomposition of biodiversity as a metric into empirical descriptors. Empty boxes mean that potentially more descriptors can be considered . . . . . . . 27

2.4 An intersection between empirical ($\mathbf{D}_E$) and orthogonal ($\mathbf{D}_O$) descriptors illustrates those empirical descriptors that are also orthogonal . . . . . . . . 28

2.5 Decomposition of biodiversity as a metric into empirical levels. Here "trophic n" for consumers refer to a group with a specified trophic level . . . . . . . 30

3.1 The number of replicates ($N(\mathcal{E}_{im})$) plotted against the biodiversity estimate ($\mathcal{E}_{im}$) on logarithmic scale and using species level as a grouping factor . . . 55

3.2 The number of replicates ($N(\mathcal{E}_{im})$) plotted against the biodiversity estimate ($\mathcal{E}_{im}$) on logarithmic scale and using descriptor richness as a grouping factor 56

3.3 Diagnostic plots – General Unrestricted Model . . . . . . . . . . . . . . . . . 58

3.4 Diagnostic plots – General Restricted Model . . . . . . . . . . . . . . . . . . 60

4.1 The unit variation in distributional properties of taxonomic richness at each taxonomic resolution . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 92

4.2 The taxon delimitation variation: species richness per genera on a logarithmic scale across different kingdoms. Distributions in Protozoa, Plantae, and Animalia, as it is shown by quantiles, have similar medians. All distributions are skewed to the right to some degree, with Plantae and Animalia having the heaviest tails. Points which are plotted at the end of the whiskers are potential outliers . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 93

4.3 The grouped kernel density plot of distribution of species richness on a logarithmic scale across different taxonomic levels superposed in a single graph. The Gaussian kernel is used to compute the estimated density. Species richness distribution at all taxonomic levels in Integrated Taxonomic Information System (ITIS) dataset has positive skew. This indicates that the bulk of the values lies to the left . . . . . . . . . . . . . . . . . . . . . . . 94
4.4 The ECDF for the log of species richness for each of the taxonomic level in ITIS dataset . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 95
4.5 Normal Q-Q plot of species richness for different taxonomic levels in ITIS dataset for the first 100 quantiles . . . . . . . . . . . . . . . . . . . . . . . 96
4.6 The grouped kernel density plot of distribution of species richness on a logarithmic scale across different taxonomic levels superposed in a single graph. The Gaussian kernel is used to compute the estimated density . . . . . . . . 97
4.7 ECDF plot of species richness across taxonomic level in Irish Benthic Marine Database (BioMar) dataset . . . . . . . . . . . . . . . . . . . . . . . . . 98
4.8 Normal Q-Q plot of Species Richness Distribution (SRD) for different taxonomic levels in BioMar dataset. Quantiles from the standard normal distribution on the x-axis are plotted against quantiles of SRD in BioMar on the y-axis. Due to the large number of data points only 100 quantiles have been used for data visualisation . . . . . . . . . . . . . . . . . . . . . . . . 99
4.9 Two-sample Q-Q plot to compare SRD in ITIS (x-axis) and BioMar (y-axis) after conditioning on taxonomic level. The Q-Q plots are non-linear and their curved pattern suggests that the quantiles in BioMar are more closely spaced than in ITIS . . . . . . . . . . . . . . . . . . . . . . . . . . 100
4.10 Plate notation: the process of stratified resampling with replacement from ITIS to instantiate the topology $\mathbf{W_i}$ as a species list $N_S(i, j, k, m, g)$ . . . . 102
4.11 Rank abundance distribution plot (Whittaker plot). Different models fitted to BioMar data (Null), following Wilson (1991), show logarithmic species abundances against species rank orders . . . . . . . . . . . . . . . . . . . . 104

5.1 Spearman correlation between different biodiversity indices in a two-dimensional space represented as a heatmap. Dark shade indicates negative correlation, light shade – positive correlation . . . . . . . . . . . . . . . . . . . . . . . . 122
5.2 Screeplot shows the percentage of variance explained (y-axis) versus the principal components (x-axis). It indicates a clear separation in fraction of total variance. The point of separation is referred to as ”elbow”. Broken stick method suggests to retain first three to five axes . . . . . . . . . . . . 124
5.3 Variance explained by each variable in three dimensional space (61.7%). Different segments on the bar show contribution of the variables to corresponding principal components . . . . . . . . . . . . . . . . . . . . . . . . . 125

5.4 Hierarchical clustering of biodiversity indices using Manhattan distance. This dendrogram is calculated using the rotation matrix of variables, the columns of this matrix contain eigenvectors (principal components). Depending on a cut off point several clusters can be observed: Cluster 1 – taxonomic diversity; Cluster 2 – functional diversity; and Cluster 3 – community structure and composition . . . . . . . . . . . . . . . . . . . . . . . . 126

5.5 The y-axis is scalar distance from the origin in the largest three PC's ($M^*$), plotted against species richness (RICH) for 1000 synthetic communities shown here rank-ordered by species richness from left to right. Dashed line shows the contribution of species richness alone on the y-axis . . . . . . . . 127

# List of Symbols

| | |
|---|---|
| $\mathcal{I}$ | a biodiversity index (see also $\mathbf{M}'$) ix, 31, 48, 74 |
| $\mathbf{M}'$ | a subset of permutation matrix of biodiversity measures (see also $\mathcal{I}$) ix, 31, 48 |
| $\mathbf{D}\vert\mathbf{L}$ | a permutation matrix of biodiversity measures (see also $\mathbf{M}$) ix, 15, 25, 36 |
| D\|L | a descriptor D at level L that forms a measure of biodiversity (see also M) ix, 15, 31, 42–46, 48, 52, 54, 60, 62, 64, 83 |
| M | a measure of biodiversity (see also D\|L) ix, 31, 48, 83 |
| $\mathbf{M}$ | a permutation matrix of biodiversity measures (see also $\mathbf{D}\vert\mathbf{L}$) ix, 31, 48 |
| $\Delta$ | taxonomic diversity index (see also DELT) xii, 113, 115 |
| $\Delta^*$ | taxonomic distinctness index (see also DSTR) xii, 113, 115 |
| $\Lambda^+$ | variation in taxonomic distinctness (see also LPLU) xii, 113, 115, 116 |
| $\Delta^+$ | taxonomic diversity for presence/absence (see also DPLU) xii, 113, 115 |
| D | a particular biodiversity descriptor 15, 25, 26, 31, 43–45, 48, 50, 63, 64, 74, 83, 120 |
| L | a particular biodiversity level 15, 25, 29, 31, 43–45, 48, 50, 63, 64, 74, 83, 120 |
| $I_S$ | semiotic “meaning” information 19 |
| $I_E$ | entropic random information 19 |
| $\mathbf{D}_E$ | the set of empirical biodiversity descriptors 26, 28 |
| $\mathbf{D}_O$ | the set of orthogonal biodiversity descriptors 26, 28 |
| $\mathbf{D}$ | the set of biodiversity descriptors 31 |
| $\mathbf{L}$ | the set of biodiversity levels 31 |
| $\mathcal{E}_{im}$ | biodiversity estimate, where $i$ is a set of $i = 1, \ldots, k$ studies on $m_i = 1, \ldots, m$ measures of biodiversity 52 |
| $\mathbf{C_T}$ | classes in BioMar 99 |
| $\mathbf{W_i}$ | the i-th community 99, 100 |
| $N_c(i)$ | number of classes in i-th community 99 |
| $\mathbf{C_W(i)}$ | the set of classes in community 100 |
| $N_O(i, j)$ | number of orders in the j-th class of the i-th community 100 |

| | |
|---|---|
| $C_j(i)$ | the j-th class of the i-th community 100 |
| $O_k(i,j)$ | the k-th order of the j-th class of i-th community 100 |
| $\mathbf{O_W}(\mathbf{i})$ | the set of orders in i-th community 100 |
| $N_F(i,j,k)$ | number of families in k-th order of j-th class in i-th community 100 |
| $F_m(i,j,k)$ | m-th family in k-th order of j-th class in i-th community 100 |
| $\mathbf{F_W}(\mathbf{i})$ | the set of families in -th community 100 |
| $N_G(i,j,k,m))$ | number of genus in m-th family of the k-th order of j-th class in i-th community 100 |
| $G_g(i,j,k,m)$ | g-th genus in m-th famility of k-th order of j-th class in i-th community 101 |
| $\mathbf{G_W}(i)$ | the set of genus in i-th community 101 |
| $N_S(i,j,k,m,g)$ | number of species in g-th genus of m-th family of the k-th order of j-th class in i-th community 101 |
| n | a total number of individuals in a community 115 |
| $\omega_{ij}$ | the matrix of species pairwise taxonomic distances 115, 116 |
| $L^C$ | the classes level 117 |
| $L^S$ | the species level 117 |
| $L^S \times t$ | the species-by-traits matrix 117 |
| $L^C \times t$ | the classes-by-traits matrix 117 |
| t | the list of functional traits 117 |
| $d_{ij}$ | the pairwise functional difference between species 119 |

# List of Acronyms and Abbreviations

| | |
|---|---|
| ITIS | Integrated Taxonomic Information System vii, 89–91, 94, 96–101, 106 |
| BioMar | Irish Benthic Marine Database vii, 89, 94–100, 104–106 |
| SRD | Species Richness Distribution vii, 97, 99, 100 |
| CBD | Convention on Biological Diversity 4, 136, 137, 152 |
| GDI | General Definition of Information 16 |
| A(IC) | Algorithmic Information Content 21–24 |
| F(IC) | Functional Information Content 22–24 |
| RDB | Relational Database 40–47, 51, 52, 61, 64 |
| ERD | Entity Relationship Diagram 40–42, 44, 45, 47, 48, 50, 51, 74 |
| biodivDB | relational database of measures of biodiversity 43, 47, 74, 81 |
| GLM | General Linear Modelling 47, 55, 62 |
| OLS | Ordinary-Least Squares 47, 56 |
| GUM | General Unrestricted Model 55, 57–60 |
| GRM | General Restricted Model 59, 60 |
| BIOTIC | Benthic Biological Traits Information Catalogue 89, 102, 103, 105 |
| OiC | Orders in Classes 90, 92, 99–101 |
| FiO | Families in Orders 90, 92, 99, 100 |
| GiF | Genera in Families 90, 92, 100 |
| SiG | Species in Genera 90, 92, 101 |
| PCA | Principal Component Analysis 111, 120, 122, 123 |
| SIMP | Simpson Index 113, 114, 120 |
| PIEL | Pielou Index 113, 114, 120 |
| JACC | Jaccard Index 113, 120 |
| SORE | Sorensen Index 113, 114, 120 |
| CHJA | Chao-Jaccard Index 113 |
| CHSO | Chao-Sorensen Index 113 |
| SHAN | Shannon Index 113, 114, 120 |
| TURN | Turnover Index 113, 120 |
| ABUN | Abundance 113, 114 |
| RICH | Richness 113, 114, 122 |

| | |
|---|---|
| DELT | Taxonomic Diversity (see also $\Delta$) 113, 115 |
| DSTR | Taxonomic Distinctness (see also $\Delta^*$) 113, 115 |
| LPLU | Variation in Taxonomic Distinctness (see also $\Lambda^+$) 113, 115 |
| DPLU | Taxonomic Diversity (for presence, absence)(see also $\Delta^+$) 113, 115 |
| SPLU | Taxonomic Diversity (accounting for species richness) 113, 122 |
| FRIC | Functional Richness 113, 117, 118 |
| FEVE | Functional Evenness 113, 117, 118, 122 |
| FDIV | Functional Divergence 113, 117, 118, 122 |
| FDIS | Functional Dispersion 113, 117, 119, 122 |
| RAOQ | Quadratic Entropy 113, 117, 119, 122 |
| GBO | Global Biodiversity Outlook 137 |
| IUV | Indirect Use Value 140 |

# Chapter 1

# Why another Study of Biodiversity?

> "I cannot help thinking of the deplorable fact that when the child has found out how its mechanical toy operates, there is no mechanical toy left."
>
> Chargaff 1978

This PhD thesis is about biodiversity and its value, or more precisely, its potential to be valued. A quick perusal of the published literature will show more than six thousand studies (Web of Science search term "Biodiversity & Value" in the Topic and the Title) which might fit that description, so why is yet another justified? Reading the titles of those publications (the first few hundred) builds an impression of a vast, varied, and unstructured subject – very nearly all concern a particular detail about a particular system, for which biodiversity is a relevant concept and for which value is worth mentioning.

Among these publications, one finds many different ways of defining and understanding what biodiversity is and what might constitute or give rise to its value. Sometimes diversity refers to specific species and their response to anthropogenic change, and sometimes it describes a property of a particular system, at a genetic or whole organism level. Biodiversity is often meant to describe the species variety of a particular guild or taxon, for example garden birds in a given location (Strohbach et al., 2009). Variety, however, is not a thing, so it is not clear how it can have a value: if value is recognised, surely it must be the value of things which compose the variety. Some recognise the benefit of genetic diversity as insurance against failures in agriculture (Tanksley and McCouch, 1997). Others may point to the store of yet to be found pharmaceuticals (see, e.g., Tan et al., 2006; Erwin et al., 2010) and some may concentrate on the psychological (Dean et al., 2011) or even religious (Bhagwat et al., 2011) benefits obtained from a diverse habitat. It is easy to conclude that biodiversity can mean different things to different people and as such a loose, ill-defined concept – it is not very suitable for the rigours of economic valuation (Ghilarov, 1996; Goldstein, 1999; Ricotta, 2005b; Mayer, 2006; Colyvan et al., 2009; Spangenberg and Settele, 2010; Meinard and Grill, 2011). However, economic valuation itself, in this context, has mostly rested on the entirely subjective public opinion assessment of stated preference surveys (with occasional additions where a market can be demonstrated) (see, e.g., Ressurreiçaõ et al., 2011). Neither loose definitions nor subjective estimates are compatible with a scientific approach, so, presently, it seems that biodiversity valuation is doomed to be unscientific. Unless an objective, well-defined yet comprehensive meaning can be found for biodiversity. To be scientific, it must have a definition that leads to unambiguous understanding, is readily quantifiable (even with units of measurement), and broad enough to capture what most people want to convey when they say biodiversity. Finding such a definition and exploring its potential as a tool for ecological economics is the overall goal of this thesis. The reason it is important is that current evidence suggests rapid loss of biodiversity, so the topic is urgent (May, 2011).

## 1.1 Biodiversity: important but confusing

Wilson (1988a) starts off his seminal book "Biodiversity" with the following words: "the diversity of life forms, so numerous that we have yet to identify most of them, is the greatest wonder of this planet". Indeed, the "tree of life", describing evolutionary diversification, has been growing for about 3.5 – 4 billion years, producing the present day diversity from a single common ancestor (Gaston, 1996). Approximately 1.4 billion years ago multicellular organisms began to diversify, and only after nearly 80% of the history of life had passed did multicellular animals appear, leading to at least a ten-fold expansion in the diversity of living forms. Present estimates of the number of species usually range from 3 million to 10 million, with figures as high as 100 million being reported (May, 2011). This large uncertainty is mainly due to our incomplete exploration and describing, though difficulties with defining species remain a problem in some taxa (He and Hubbell, 2011). The diversity of life at the genetic level may be even greater. Though the genome of organisms ranges from about 100 genes in bacteria to 40,000 in many plants and many genes are found in common among wide groups of organisms, the combination of them and the control networks that govern their behaviour can give rise to variation among individuals of the same species. For example, a typical mammal such as house mouse (*Mus musculus*) has about 10,000 genes, forms thousands of genetically distinct populations and millions of unique individuals within them.

We are now living in what has been termed by some the $6^{th}$ great mass extinction period. The rapid loss of diversity through extinctions is thought to be attributable to the behaviour of one species – *Homo sapiens* (Neumann et al., 2009). The intensity of extinction has varied markedly over time, but the current loss of biodiversity seems to be the largest in the past 65 million years, according to (Wilson, 1988b), amounting to a great natural catastrophe. As species pass from the world usually unnoticed, genetic variation and ecological function are permanently lost.

Given that scientifically-based estimates of the rate of extinction "vary not only widely, but wildly" (Brown, 1988), we do not know how fast species are disappearing. The precautionary approach requires that we take action to arrest harm in the face of scientific uncertainty, it is therefore an urgent matter to conserve life's diversity. Concern of this kind set the scene for the concept of "biodiversity" to arise. "Biodiversity" was introduced for the first time at the first National Forum on BioDiversity, held in Washington, D.C., on September 21-24, 1986, intended as a simple reference to the diversity of the biological world. The term "biodiversity" quickly gained a hold in several arenas including political, management, and scientific.

Biodiversity decline was originally attributed to species extinction, as exemplified by US Endangered Species Act in 1973, which became one of the most influential legal documents in the field. Tacitly understood at a species level, species count became the unit for measuring biodiversity which in turn grew as a tool in species conservation. Hamilton (2005) reviewed the homology between species and biodiversity concepts and concluded that as long as there are substantial theoretical limitations, the value of the term "biodiversity" to ecologists will remain questionable. Crist (2002) criticised the use of species extinction rate as a measure of biodiversity loss, mainly on the grounds of the unstated assumptions. Crist's main argument was that as the baseline value of total species number is uncertain, estimated numbers of species lost is necessarily vague, diverting attention away from the "biodiversity crises". Swingland (2007) pointed out that species, as a fundamental unit of the living world, is commonly and incorrectly used as a synonym of biodiversity, claiming that this is very misleading since preserving species is not the same as preserving their diversity.

Other commentators have been less critical, finding biodiversity, a "useful, versatile concept" (Lecerf and Richardson, 2010) which is found at many levels of biological organisation, with broad and wide ranging implications. It can mean anything from genetic and phenotypic variability, to variability in species numbers, ecosystem properties, and patterns, as well as functional heterogeneity. This, however, leaves ecologists uncertain and uncoordinated as to which components of biodiversity should be quantified and reported to society at large (Feld et al., 2009). Although the use of indices based on species count and abundance is a firmly established tradition in ecology, another important level at which biodiversity may be considered is the genetic level. More recently, there emerged yet another strand in biodiversity research which is a shift from components to processes (see, e.g., Balvanera et al., 2006, for a meta-analysis of biodiversity-ecosystem functioning relationship) and use of indices as measurable indicators that reflect the nature of these processes. To unify as many components of biodiversity as possible, ecologists are in constant search of new indices of biodiversity. They argue that biodiversity is so complex, that there is no (and there never will be) single index of biodiversity. Clearly, with the understanding of this "widely used but rarely defined" concept (Hamilton, 2005), biodiversity is complex and ambiguous, therefore difficult to use in quantitative science.

The conceptual definition of biodiversity given by The Convention on Biological Diversity (CBD) is a good example of this. According to this founding treaty document, biodiversity is "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". It is a broad and all-inclusive definition, which in practice neither says what are exactly the components of biodiversity that need to be sustained, nor how to measure them. This is typical of many definitions in which the essential complexity of biodiversity is overlooked (Polski, 2005).

Further, it is claimed that to maintain a level of welfare acquired by societies, some threshold level of biodiversity is required (Perrings and Pearce, 1994; Huggett, 2005). Development that would improve welfare, whilst maintaining this biodiversity, is loosely referred to as sustainable development (see also Gowdy, 2000, for a comparison of "weak" and "strong" sustainability). Although an exact mechanism of interrelation between sustainable development and biodiversity is yet to be found, at this stage there seems to be consensus among scientists that biodiversity is, in some way, crucial for sustainability (e.g., Kim and Byrne, 2006; Suneetha, 2010). To examine this scientifically, we must be precise about what components or aspects of biodiversity are crucial and in what way.

The link between biodiversity and sustainability was formally established by the CBD, according to which: sustainability is "the use of components of biological diversity in a way and at a rate that does not lead to the long-term decline of biological diversity, thereby maintaining its potential to meet the needs and aspirations of present and future generations". This definition recognises only the effects of human activity on biodiversity, not the converse.

As in the CBD definition, it leaves unspecified what needs to be sustained – the components of biodiversity are not defined. The uncertainty has allowed ecosystems, biodiversity, our "way of life", or our "standard of living" to join the list of options; none of which have been given a precise and generally accepted definition. If one takes the scientifically justifiable position of treating human systems, such as society and its phenomenon – the economy, as a component of the wider ecological system of the earth, then sustaining all this amounts to maintaining the structural integrity of ecosystems. Whilst not quantitatively clear, it is still qualitatively obvious that we are failing to do this when habitats, species, and genes are disappearing form the world. Unfortunately, many habitats and species known to be lost are also known to have been irreplaceable using the prevailing and foreseeable technology (see, e.g., Barlow et al., 2007). It is therefore highly likely that the present Holocene (and anthropogenic) mass extinction constitutes a failure of sustainability in any meaningful sense.

Nevertheless, as a concept for policy vision, sustainability is useful, since it simultaneously attempts to address ecological and socio-economic systems. This is strongly advocated by the Brundtland report of the United Nations World Commission on Environment and Development (WCED, 1987) by defining "sustainable development" as "development that meets the needs of the present without compromising the ability of future generations to meet their own needs". This form of the concept is tightly linked to that of ecosystem services that addresses those properties of ecosystem that benefit humans, either directly or indirectly (Hooper et al., 2005). Then, having stated, hypothetically, that biodiversity is responsible in some way for providing ecosystem services by a system (see, e.g., Beaumont et al., 2007), sustainability is all about maintaining the use of that system within its elastic limits. With the goal of sustainable development to ensure that economic progress has a healthy ecological foundation, biodiversity has gained use as a tool to diagnose the health of the supporting system. In a thought experiment we might imagine a world with all biological diversity replaced by a few highly engineered organisms that are maintained in tightly controlled monocultures to feed a human population: is that sustainable? To answer this question first we need to know if there is enough variety to provide necessary ecosystem services to maintain the whole system – engineered and the engineers, representing human society. It seems reasonable to say that at present, humanity cannot persist without some minimal level of natural diversity, since no effort to produce a self-sustaining living system has yet succeeded, even on a very small scale. To this extent, biodiversity is a fundamental matter of human persistence.

Concerned with biodiversity decline, Pearce and Moran (1995) argue that "demonstrating the value of biodiversity is a fundamental step in conservation". In other words, it was accepted that modern public-level decision making requires an economic framework for justification. Value, placed on the diversity of life, has become a necessary pre-condition for its conservation (Edwards and Abivardi, 1998). Given the short time-horizons prevalent in such economic decision making (technically incorporated through discount rates), we are in danger of failing to meet our sustainability promise (recalling Keynes' famous statement that "in the long-run, we are all dead"). It was not always the case: a century ago, for instance, the diversity of life was considered as an integral part of life and valuing it would have been thought both "presumptuous and a waste of time" (Ehrenfeld, 1988).

We now see two extreme approaches to conservation of biodiversity – utilitarian and deep ecology. While utilitarian approach is based on the idea that biodiversity supports human well-being through ecosystem services (Millenium Ecosystem Assessment, 2005), deep ecology considers preservation of species as a moral responsibility based on the argument that species have an intrinsic value (Brown and Moran, 1993). The Convention on Biological Diversity (Glowka et al., 1994) supports both stances with a range of other arguments lying in between. By preserving biodiversity we do not only keep ecosystems within their elastic limits which allows us to enjoy the level of services they provide, but also keep the exploitable information which may be of direct benefit. By exploiting and depleting natural resources, we affect both the present and the future ability of ecosystems to provide value to humanity. This is because changes of ecosystems brought about by economic activity are often irreversible: once depleted, there is no way to recreate non-renewable resources in their original state.

An economic context of present decision making and the need to justify all conservation efforts economically leads us to the problem of value and to environmental ethics (Maclaurin and Sterelny, 2008). Setting aside the latter, policy makers feel compelled to place an economic value on diversity whenever the dominant economic realities (e.g., population growth, poverty, and wealth accumulation) are principally to blame for biodiversity loss. For example, McKee et al. (2004) quantified a model of relationship between human population density and the number of threatened species and made a strong conclusion that population growth would have to be limited in order to conserve biodiversity. A great overlap between severe poverty and key areas of global biodiversity was acknowledged by Fisher and Christopher (2007), illustrating the conflict of goals between conservation and economic development in so many places.

## 1.2 Sources of value

To formalise the mechanism by which conservation decisions are taken, biodiversity value (in whatever sense) must enter into cost-benefit analysis (Salles, 2011). This requires that ecological measures of biodiversity are translated into the economically defined value, most commonly interpreted as an average willingness to pay. Immediately, we see several problems with metrification: which measures of biodiversity are to be used? How will people's willingness to pay for these be estimated and combined? At the root, we must ask what aspects of biodiversity give rise to value?

Frequently, we find that it is not even biodiversity that is being valued. Consider here an example: ginseng is a herb for which many positive pharmacological claims are made (Attele et al., 1999) and for which it is highly valued. To justify costly conservation measures aimed at ending the decline of ginseng (McGraw, 2001), economists have attempted to attach a monetary value. This has typically been quantified by the price paid for the pharmaceutically active substances that ginseng provides. In this standard economic approach, the value is defined by a market price for goods, not a component of biodiversity. Thus, valuing a biological resource requires the dissection and abstraction of components which can each be valued for their utility to people. These components are, in isolation, not living; the life responsible for their production is quickly overlooked. We may ask what factors determine the survival and medical efficacy of ginseng? Could it be that some level of biodiversity among soil biota is important? This is not considered when the plant is reduced to a set of useful chemicals.

I use this example to suggest the following misunderstanding that currently prevails biodiversity valuation: very often we value the components of biodiversity (plants, birds, etc.) rather than biodiversity itself (that is the value of variation in the properties of the plants or birds). Though it is scientifically possible to quantify the latter, this is generally not considered for valuation. The reason is likely to be partly the difficulty in identifying a utility function for diversity and partly the common practice of confusing biodiversity for a collection of organisms.

The tendency to concentrate on the production by components of biodiversity was confirmed by de Groot et al. (2010) and Nunes and van den Bergh (2001) who reviewed economic approaches to valuing biodiversity and its ecological consequences. They showed that economists mainly value the services that biodiversity hypothetically provides to humanity, rather than attempting to value biodiversity itself. This approach tacitly assumes that biodiversity (however defined) can be substituted by another source of its services, following the key assumption of neo-classical economics that consumers maximise their utility whilst remaining indifferent as to its source. Both philosophical and instrumental objections to that assumption in relation to ecosystem services were identified by Spangenberg and Settele (2010), who argue that “the basic assumptions underlying economic valuation are far from realistic and represent rather a caricature of human behaviour”.

There are, then, three broad sources of value to consider: the philosophical existence value which enters economics only in the form of subjective reporting within opinion polls; the direct market value, almost always referring to abstracted components of the biological system whose diversity we measure; and the supposed effect of biodiversity on the functioning of all the ecological systems upon which we depend flourish, this being the source of *indirect use-value*. In this thesis, I explicitly exclude existence value on the grounds that it cannot be quantified objectively, hence scientifically. I also exclude the market value on the grounds that economics has largely solved that problem already: there is little role for science there. Instead, I will concentrate on the sources of indirect use value to be found specifically in biodiversity, as opposed to the component parts which make up the diversity. The principal connection between biodiversity and ecological supports for human activities is the hypothetical relationship between biodiversity and ecosystem services. This relationship become the primary focus of modern biodiversity research and is motivated by the

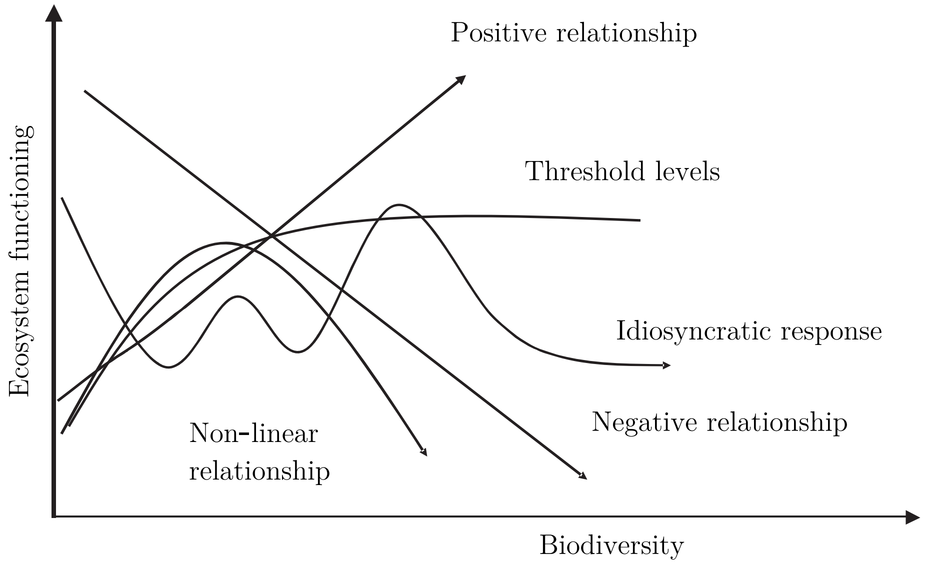


**Figure 1.1:** Theoretical models of relationship between biodiversity and ecosystem functioning. Adapted from Danovaro and Pusceddu (2007)

.

thrust of international and national agreements on biodiversity conservation at government level.

In company with the "main-stream", I shall adopt an anthropocentric stance for my work. The following chain of logical connections, therefore, will be assumed: living processes give rise to biodiversity, which I define as variety in life; this diversity supports ecosystem persistence and functions; these properties are necessary and quantitatively beneficial to human beings; for this reason we value them, in the sense of being willing to give something up in order to keep them. The meaning of quantitatively beneficial here is a hypothetical "dose-response" relation in which more biodiversity leads to more or better functions which in turn lead to more human welfare and hence value.

Ecosystem function is a product of interactions at many levels of biological organisation, giving rise to ecosystem stability, resilience, and complexity. Ecologists are still trying to discover whether, and to what extent, and in what circumstances, higher diversity gives higher capacity to recover from perturbations, and therefore higher resilience and resistance (see, e.g., Srivastava and Vellend, 2005). There are many different hypothetical models describing the relationship between biodiversity and ecosystem functioning (Figure 1.1); but in general we are not able to identify a specific form.

To identify the form of the relationship we need to discern the processes underlying ecosystem functions (Loreau et al., 2001), as well as to quantify the relationship between biodiversity change and ecosystem functioning (Balvanera et al., 2006), and, for this we need at least to conduct experimental investigations (Naeem and Wright, 2003). Some progress has been made, but in most of the studies so far, only manipulations on the level of primary producers have shown evidence that change in biodiversity affects ecosystem processes and these vary with level i.e., weaker on ecosystem compared with community level (Balvanera et al., 2006). Taking account of food-webs may be the next step and this has been attempted by de Ruiter et al. (2005), O'Gorman et al. (2011), and others. Overall, it seems we are still at an early stage in quantifying biodiversity-function relationships, even though qualitatively it is clear that biodiversity affects ecosystem functioning in a number of ways (Hooper et al., 2005). The present lack of quantitative knowledge about these relations poses severe limitations on the biodiversity-ecosystem services approach to valuation that is currently so popular. This leaves open the possibility to attempt a direct valuation of biodiversity, which is the path explored in this thesis. In the next section I shall explore the steps needed to allow for a direct valuation of biodiversity in terms of its indirect-use value, rather than its philosophical existence value.

## 1.3 Biodiversity as information?

Biodiversity is not a thing – it is literally the amount of difference (meaning of diversity) in a biological system. It seems, therefore, that it cannot have a direct value. However, in the next chapter I will build on the idea that degree of difference is in fact a way of saying information content and thereby I will identify biodiversity as a concrete and real property – information. Given a concrete property of a biological system, which can be quantified, we have something with the potential to be valued, and this is, by definition, an intrinsic property of the system. Thus, the identification of biodiversity with the information contained within the biological system enables an objective value to be quantified. This way, intrinsic value is transformed from a subjective, philosophical notion, into an objective, scientific, and concrete notion, suitable for economic calculations.

Therefore, the research aim of this work is *to demonstrate, at least in principle, that biodiversity can be interpreted as the information content of the biological system enstantiated as the degree of difference among it components.* Then, this degree of difference, aggregated over all levels of the biological organisation, can be used to fully quantify biodiversity.

To think about biodiversity in this way, certain requirements need to be satisfied. To begin with, its definition must be precise and independent of context. This sets the first research question which I address in Chapter 2 – *Is it possible to form the basis of a scientific measure of biodiversity through identifying patterns in difference?* To answer this, I gain insight into the meaning of biodiversity as completely as possible using the principles of information theory. This involves setting a formal structure to the concept of biodiversity and establishing the logical connections between its components. Using a formal vocabulary my argument throughout Chapter 2 results in a precise definition of biodiversity. In so doing, I answer the first research question in principle, affirmatively because I show a logical derivation of 'potential to be valued' from aggregated difference.

This conclusion leads to the second research question of putting the principle into practice (operationalising as a method), which I address in Chapter 3: *Can patterns in difference be used in practice to form the basis of a scientific measure of biodiversity?* I answer this question through a meta-review of the empirical biodiversity literature which shows variety, frequency distributions, and relationships among different biodiversity metrics. The meta review is supported by a relational database framework, in which all the biodiversity concepts that are found in the literature, are organised into a formal structure. This structure then makes the prototype for organising biodiversity metrics in further studies. Unfortunately, I find the existing published data insufficiently standardised at present to perform the practical task post-hoc. I conclude that aggregating across studies to show meaningful patterns in biological difference is not yet a practical objective.

The failure of the literature-based approach motivates the next question: *Could biodiversity as difference be operationalised, at least in principle?* The corollary to this question being: what kind of data and which processes are needed to achieve operational methods? I address this by simulating a set of ecological communities in Chapter 4 for which I then attempt to find patterns among the components of biodiversity in Chapter 5. For this to be realistic, I use an algorithm that imitates the distributional properties of real benthic marine communities. Applying multivariate analysis I show that my overall research aim of interpreting biodiversity as information can be achieved, at least in principle. The analysis shows which data are essential and how they might be combined in a standard method for estimating biodiversity more comprehensively. The results demonstrate a substantial advantage in representing biodiversity as information, compared to the most commonly used metric – species richness.

Finally, I conclude this work in Chapter 6 by assessing these conclusions for their relevance to the policy-driven applied science of biodiversity research, using the following question: *Can understanding of biodiversity as information be put in a practical policy-relevant context?* Here I show how the more formal and fundamentally derived concept of biodiversity as information could assist through standardising and making more comprehensive the estimation of biodiversity for the international policy arena. Relating it to economic questions of prioritising, I highlight the advantages of the information approach in cost-benefit analysis. One of the main recommendations arising from this is to build a comprehensive, accessible, and integrated relational database to include species-indexed phylogenetic and functional data. My work therefore gives a formal support to current global efforts in biodiversity database management and highlights the shortcomings of species richness as a common currency for biodiversity.

## Summary

1. Practical action to conserve biodiversity usually requires estimates of its value and this is hampered by the conceptual diffusion so that formal, objective derivation of a comprehensive and quantifiable definition is now needed;

2. Using principles from information theory, meta-review of biodiversity literature and mathematical simulations with assessments of biodiversity metrics, I propose one such definition, demonstrating its operational potential.

# Chapter 2

# What (really) is biodiversity?

“O how they cling and wrangle, some who claim
For preacher and monk the honored name!
For, quarreling, each to his view they cling.
Such folk see only one side of a thing.”

Jainism and Buddhism. Udana 68-69:
Parable of the Blind Men and the Elephant

## 2.1 Setting out the problem

As we have seen, great concern over biodiversity loss has stimulated efforts to quantify its value, but these efforts have been impeded by difficulties over the definitions of both biodiversity and value, leading to a wide range of concepts, methods, and outcomes. Biodiversity is often considered as a crucial link between ecosystem functioning and human well-being (Naeem and Wright, 2003), justifying scientific interest. However, the fact that “biodiversity means different things to different people” (Noss, 1990) and “may itself have a diversity of meanings” (Begon et al., 2006), mitigates against scientific clarity and does not help conservation.

Since the first introduction of “biodiversity” as a portmanteau word blended from “biological” and “diversity”, biodiversity was used both on its own and in combination with other concepts appearing in a variety of contexts. It is more or less agreed that the meaning of biodiversity is context dependent, because standards of biodiversity knowledge, and its justification, vary with context. Despite the large body of literature devoted to biodiversity,

Based on Farnsworth, K. and Lyashevska, O. Functional complexity: the source of value in biodiversity? Submitted to Ecological Complexity, December 2010 there has been no agreed definition of biodiversity which is both precise and comprehensive (see, e.g., Norton, 1994; Purvis and Hector, 2000; Ricotta, 2005a; Mooers, 2007). The last two decades of research leave it still unclear what constitutes biodiversity and how it should be measured (DeLong, 1996; Feest et al., 2010). Although many individual scientists use a clear, self-consistent definition within their own sub-field, biodiversity *as a whole* remains a vague concept with its scientific meaning being questioned in a number of publications (e.g., Ghilarov, 1996; Sarkar and Margules, 2002; Ricotta, 2005a; Mayer, 2006). For example, ecologists use it primarily to denote species richness or evenness, but geneticists refer to genetic variability within species. Ambiguous understanding of biodiversity generates a proliferation of measures of biodiversity (King, 2009), further exacerbating the ambiguity by introducing different meanings in the measures. Biodiversity can be quantified in many different ways (May, 1994) which drive the meaning of biodiversity. More generally, the meaning of a concept is definable by its corresponding set of operations or measures. Each contextual definition of the meaning of biodiversity is predefined by inconsistent methods of quantification that lack standardisation (Mellin et al., 2011). These measures, since they pertain to contextual definitions, are contextual as well. Therefore, a diversity of meanings encompasses a diversity of measures, each of them intended to represent some facet of total biodiversity. Examples include genetic and phenotypic variance, species numbers, ecosystem structural properties, and patterns of functional heterogeneity.

In this way, discrepancies among contextual definitions of biodiversity has led to overlap and potential redundancy as well as possible incompatibility of metrics. This proliferation calls us to rationalisation and synthesis: to identify which features of biodiversity are mathematically independent and thereby to find the irreducible set of metrics which must be included to encompass, what one may call "total biodiversity". Implied in that goal is the identification of redundant metrics, those which are so mutually correlated that any one of them may be taken to approximate the others. This especially matters in the context of valuation, where biodiversity is weighed against economic goods in cost-benefit analysis. According to the economist Weitzman (1992), we need a "more or less consistent conceptual framework" for decision making to achieve sustainable outcomes, and this requires an understanding of the "real" meaning of biodiversity. To find such consistency through acceptance of a unifying definition among disparate interests, we must dig deep towards a common root of meaning. We will have to trace back to first principles, the meaning of biological diversity, using theoretical tools: ontology, information theory, and complexity theory.

Thus, I identify the following problem to be solved: that the current state of understanding of what constitutes biodiversity is very fragmented. Scientists from different disciplines see biodiversity in a number of different ways without any clear agreement, often without realising there is a disagreement. In the epigraph at the head of this chapter, taken from an old Indian parable, it is suggested that the current state of understanding of what constitutes biodiversity is fragmented: separate, perhaps incompatible views and that this arose from our approaching a complex and multidimensional concept from separate directions. The resulting uncertainty may be exploited by antagonists to biodiversity conservation.

To construct a precise and scientific definition of biodiversity that would allow its quantification, all ambiguities in the meaning of biodiversity should be eliminated. Therefore, this chapter presents a theoretical argument aiming to develop a definition of biodiversity which is comprehensive enough to encompass other scientific meanings, but specific enough to be unambiguous, objective, and quantifiable in a single currency. The overall method used here is that of reductionism, followed by synthesis, I shall first seek the primitive elements of biodiversity and then reconstruct the concept from them in a way which maintains the logical connection between primitive and higher concepts; carrying units of metrication with it.

## 2.2 Methods

Answering the question in the title of this chapter will be tackled in two different ways, each best suited to a different level of understanding biodiversity: the conceptual and the quantitative. Starting with a conceptual analysis of biodiversity it will be shown how biodiversity and information can be connected to derive the information-based meaning of biodiversity. Having defined the concept, then the focus will be shifted towards biodiversity operationalisation, understood here as a process of making the concept measurable. Here a more quantitative approach will be used, within the framework created by the new conceptual definitions arising from the first part. This will result in a formal structure for decomposition of biodiversity from which metric analysis can follow in the remaining chapters.

Conceptual analysis will consist first of syntactic decomposition, constrained by a set of defined axioms, but will take the form of an argument using sequences of inferences, rather than a formal language proof. Nevertheless, it will lead to elemental concepts to which I shall apply a philosophy of information developed by Luciano Floridi (Floridi, 2003, 2005). From it I will obtain an information-based definition of biodiversity. I shall then use concepts from information theory as it applies to life, to develop a theoretical argument that operationalises Gregory Bateson's (1972) famous statement that information is "a difference that makes a difference". I shall interpret making a difference in the biological context, as meaning "functional" and go on to isolate the functional part of the total information which biodiversity represents. Further information theory-based argument will be used to show how functional information is equivalent to complexity, so the argument will lead from biodiversity as difference to biodiversity as functional information and biocomplexity (see Figure 2.1).

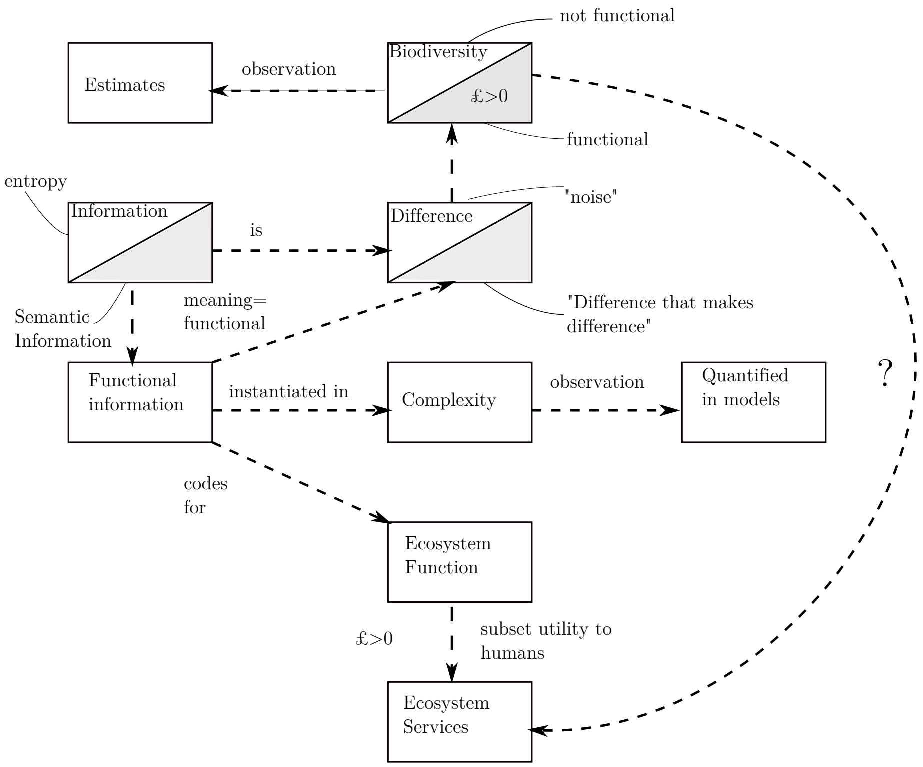


**Figure 2.1:** Conceptual analysis of biodiversity. Sequences of inferences (represented by arrows) applied to an information-based definition of biodiversity suggest a link between biodiversity and ecosystem services

Focussing on function is deliberate, since by definition only functional attributes can provide instrumental (e.g., indirect use) value, which is ultimately what I am searching for (following the arguments in Chapter 1). This relationship between function and potential to be valued will be justified and elaborated by the conceptual analysis of biodiversity as functional information.

For the second, quantitative part of the answer to my title question, I will use concepts from the preceding part to decompose biodiversity as a metric into elements, each being a member of a class, of which there are two: termed “descriptor” (symbol D) and “level” (symbol L). Descriptors perform the function of axes of variation, or difference, whereas levels specify the position in an hierarchy of biological organisation (from molecules to the global Gaia system) at which variation is measured. With this, a general measure of biodiversity will be defined as a couplet of descriptor and level elements, such that every possible measure is an element of an $\mathbf{D}|\mathbf{L}$ matrix. Then every possible index of biodiversity (an indeterminate set) can be expressed as a combination of D|L components. This decomposition and matrix representation allows for construction of arbitrary biodiversity indices and the rapid assessment of existing indices in terms of their orthogonality or redundancy. Thus, the quantitative tools will be available for the following chapters.

## 2.3 The nuts and bolts of biodiversity

I start my conceptual analysis by stating four (common language) axioms, which form the foundations of my argument.

A1: Biodiversity is not organisms or species, or forests, or beautiful environments; it is a concept that attempts to capture the richness of variety in living things, not the things themselves.

A2: Biodiversity is a *material* property of a biological system, which is defined generally to include anything capable of (or having previously exhibited) life; where life is defined in the most general terms as an autopoietic and cognitive molecular system (Maturana and Varela, 1980; Bitbol and Luisi, 2004). A system is a stable assembly of components which interact to produce consistent responses (outputs) to environmental influences (inputs). An autopoietic system is one which constructs and maintains itself, and a cognitive system is one which *selects* inputs from the environment.

A3: The word "biodiversity" literally means the diversity within a biological system, where diversity quantifies the total difference among the system's parts. These parts are the system components and also the interactions among them.

Note: Biodiversity as a concept carries a notion of difference, which suggests that it is not a number of elements that should be counted (as, e.g., in species richness) but how these elements differ from one another. According to Gregorius and Gillet (2008) "any concept of diversity invokes the notion of difference".

A4: Differences can be organised into independent categories (e.g., colour, shape, and size of objects), so that total difference is a multidimensional property (orthogonal axes of difference variation representing the independent categories).

Note: Biodiversity is frequently discussed in multidimensional terms (e.g., Purvis and Hector, 2000).

Having laid these foundations, it is clear that I am looking for a fundamental definition of diversity, or difference, relevant to the biological system as described. The answer to this comes from philosophical enquiry into the precise and fundamental meaning of information. One of the major recent break-throughs of that research was the development of the data-based definition of information using a General Definition of Information (GDI) (Floridi, 2005). This definition is formally expressed from basic principles and is used in computer science and related fields. It essentially says that information is data arranged in a way so as to give it meaning. This is not merely stating the obvious: it only seems obvious because in most cases it is the way one experiences information; GDI constructs information from basic components. Crucially, it shows that information must be made of data. In the simplest case information can consist of a single datum: this is the elemental form of information. Further, a datum is reducible to a lack of uniformity , so a general definition of a datum is: a putative fact regarding some difference or lack of uniformity within some context (Floridi, 2011).

Thus, the elemental difference which I need to form the basis of diversity in biodiversity coincides with the "diaphoric definition of data" (diaphora is the Greek word for "difference") introduced by Floridi (2003, 2005), in which a binary (Boolean) bit is the unit of information and a bit is a single difference.

A digital image (see Figure 2.2) provides a useful analogy to explain why this works for biodiversity, including its extension to a multi-dimensional concept (A4 above). First start with a blank page – there is no information here and it requires no data to describe it. In order to establish any difference there should be at at least two entities that differ. In this case, I divide the image into two parts: one white the other black, thus creating a single difference. This requires a single bit of data to describe it: the image codes one bit and instantiates one bit of information. If it is divided again, along a different line, then further difference will have been added and more difference means more data which means more information: there are more pixels in the image.

In a colour image, each pixel must be coded in three independent quantities: the three primary colours (Red, Green, and Blue). These can be represented in a three-dimensional space and this description is the most compact possible. Note that other systems such as RGB+brightness are used in practice, but can be mathematically reduced to 3 orthogonal components. This fact will serve as an analogy for the combination of biodiversity measures later on.

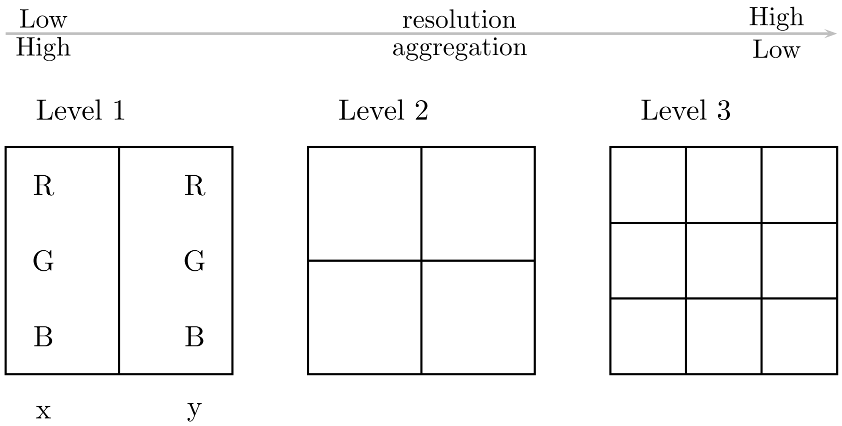


**Figure 2.2:** The notion of difference in the concept of biodiversity illustrated using the digital image analogy

Figure 2.2 shows this simplest case of a two-pixel colour image consisting of a pixel $x$ and a pixel $y$. Each pixel is fully described by the amount of RGB, putting it more formally, Red, Green, and Blue colours are necessary and sufficient to describe each pixel of the image. Where Red, Green, and Blue are primary and therefore elemental axes of variation in the image, I could describe it in terms of composite colours (e.g., yellow, magenta, and cyan), still in just three orthogonal dimensions, but not elemental. Further, I could describe it using a colour library (a pallet or swatch card), in which case I would have potentially many more than three axes of variation, but these axes would be neither elemental nor orthogonal: each colour in the library would be reducible to its RGB elements. This too will serve as a good analogy for the multitude of biodiversity indices.

The number of pixels in the image increases with the aggregation level: moving from the left side of the figure (high aggregation level) towards the right side (low aggregation level) the total amount of information that is carried by the image increases. This is because as aggregation decreases, more pixels means more pixel-pairs and so more potential differences among the pixels. This aggregation is analogous to that found in biodiversity, where the ecosystem level can be considered as the highest aggregation level and the molecular level of life (see A2 above) is the lowest aggregation level.

Using the digital image analogy, difference between any pair of pixels $i$ and $j$, is the vector sum over R, G, and B values, giving a distance: $\mathbf{D}(i,j) = ([R(i)-R(j)]^3 + [G(i)-G(j)]^3 + [B(i)-B(j)]^3)^{1/3}$. Aggregated over all unique pairs for an $n$ pixel image, the total scalar difference (and therefore information content) in the image is:

$$D = \sum_{j=2}^{n} (|\mathbf{D}(1,j)|) \qquad (2.1)$$

Note that all other possible pairs of the $n$ pixels, could be written in terms of those appearing in the sum, so do not represent additional information.

The principle illustrated by this digital image analogy applied equally well to biodiversity, showing that it too can be interpreted as total difference by forming a distance measure from the vector sum of pairwise differences among the components of the biological system. Such a measure built from biological diversity will serve as a measure of the information contained in the biological system.

This leads to the following statement:
Since (a) biodiversity is the measure of total difference in a biological system; and (b) difference is data (which is information), then biodiversity is a measure of information content. This statement is the information-based definition of biodiversity.

### 2.3.1 Biological systems as information

Some, but not all information is meaningful in the sense that it can cause a predictable change in a system: Bateson (1972) called (meaningful) information "a difference that makes a difference". This sense of meaning does not refer to language or human perception, it merely indicates that the information can interact with something (including other information) to create a predictable effect. Unpredictable effects may result from random information but their lack of predictability implies a different effect on each interaction, producing a sequence which would not make sense, it would be random noise, rather than signal; it would be meaningless. Meaningful information by definition has a context, which in general is a system component that responds in a predictable way upon detecting the information. This is made possible by selection which acts as a filter on the noise of random information: in other words, cognition (Maturana and Varela, 1980). For example, we could imagine a large collection of possible protein molecules in a solution: random shapes. A few of these fit into receptor molecules and when they do, something predictable happens. These few have a context, they are selected by the receptors from the random mix and interact with them. Having a predictable consequence, they carry meaning: they are signals (which are context dependent). Meaning is therefore not an intrinsic property of information, rather it describes predictable interaction. We should not think of meaning as a noun, but as a verb (Neuman, 2008), describing the phenomenon of cognition.

In contrast, information theory defines entropy as the measure of unpredictability, sometimes termed “surprise”. “Meaning” is not considered in Shannon’s (1948) definition of information, where maximising entropy maximises total information content by increasing unpredictability. Entropy is a measure of the unknown random information content of a system and physically, this determines the system’s capacity to do thermodynamic work. If you burn a plant, you obtain from it the entropic information which you can use as heat and this may have some (relatively small) value. In the process you will have destroyed the semiotic information, which if left intact would have specified how to reproduce the plant to get more, it would have told you its evolutionary history, and would carry the blueprint for thousands of potentially useful chemicals that the plant could synthesise. This information is potentially far more valuable than the fuel-energy you obtained. Measures of biodiversity that count the total information content of lists of organism categories or components do not make this distinction, so weigh entropy equally with semiotic information and are consequently misleading indicators for value. Isolating semiotic information will lead to an intuitive and justifiable link between quantified biodiversity and its value. The question I must now address is, how to identify and quantify semiotic information?

In order to include only “difference that makes a difference”, it is necessary to identify the predictable patterns accompanied by the noise of random information. With this in mind, I will classify the total information content of any system by two distinct components: $I_{tot} = I_S + I_E$, where $I_S$ is the, semiotic, “meaning” information and $I_E$ is the entropic, random information. Each of these terms can be quantified by the Algorithmic Information Content (Chaitin, 1990) if the terms can be isolated. $I_S$ could, in principle, be quantified by the Gell-Mann and Lloyd (1996, 2003) “Effective Complexity”, defined as the minimum description length of regularities, but only given prior information about the regularities (see McAllister, 2003, for an expansion of this criticism). I am searching for a way to identify $I_S$ without such prior information.

Bates (2005), quoting earlier works, defines information as: “the pattern of organization of matter and energy”. This definition explicitly addresses semiotic information. Patterns of organisation are not random since organisation is the alternative to randomness: patterns show either order (characterised by symmetry) or complexity (broken symmetry). Crystal lattices and DNA provide concrete examples of these two kinds of pattern. Schrödinger (1944) realised that symmetrical order was insufficient to account for the genetic information coding life, concluding that it must be in some aperiodic (non-symmetrical) molecule (well before the discovery of DNA). The required organized aperiodicity is commonly known as “complexity”; a defining characteristic of which is a high capacity for semiotic information. Adami et al. (2000) subsequently showed how all biological systems are complex systems in this scientific sense. Information is therefore not just stored in DNA and RNA: it is the whole biological system that embodies semiotic information, and, hence, biocomplexity as a whole is the storage of semiotic information in living nature. Valentine (2003) explained that biological complexity exists as a set of hierarchical levels, an example being that shown in Table 2.1. Spontaneous creation of semiotic information from complex order is a key property of such hierarchies: every level spontaneously *emerges* from the one below (Adami et al., 2000).

Realizing this has an immediate consequence for what one means by “biodiversity”. Biological complexity exists within a set of hierarchical levels (see Table 2.1) and is added to by interactions among them. This modular hierarchical structure means that biodiversity includes the diversity of: molecular structures; networks and pathways (responsible for processes such as metabolism and protein synthesis); cell types; tissues and organs as well as whole organisms and the way they interact in community networks, i.e., at multiple scales (Bar-Yam, 2004). (Note: Sarkar and Margules, 2002, argued that including all this amounts to biodiversity becoming all of life, including its behaviours.) One of the key properties of these hierarchies is self-organisation and emergent complexity – the spontaneous creation of semiotic information from complex order (Adami et al., 2000). As a result, even a complete description of genetic information fails to account for the full complement of semiotic information, which is why, for example, seed-banks are no substitute for community conservation, as noted intuitively by Lee (2004) (see also Cowling et al., 2004).

So, analytically, we are looking to distinguish complexity from its accompanying random information, within the algorithmic information content, embodied in a biological system. Using Bates’ (2005) definition, I count biocomplexity (complex pattern) as the storage of semiotic information in nature and set this as the target for biodiversity measurement. Measures of information based on message-length are intuitive and well known, notably as the Algorithmic Information Content (A(IC)) (Chaitin, 1990), which is an operationalisation of the Kolmogorov complexity, discussed by Gell-Mann and Lloyd (1996), who proposed the "effective complexity" as an alternative measure of semiotic information.

Accordingly, I now define biodiversity as a measure of the total complexity of a biological system (biocomplexity), including complexity at each of the nine levels shown in Table 2.1. This is equivalent to the total semiotic information of the system, a substantial amount of which may be found in the genome of its constituent organisms. Crozier (1997) concluded that phylogenetics should form the basis of biodiversity measures, understanding that the goal of biodiversity conservation was to preserve information, much of which is held in the genome. According to Table 2.1, I must add to this the supra-organism level complexity. Having identified biocomplexity as meaningful information, I now need to show how it may be quantified, which I do by returning to Bateson's (1972) definition of meaning as "making a difference". This is formalised by Functional Information.

**Table 2.1:** A nine-level hierarchy of biocomplexity. Left column names the level of organization and right column gives examples of the complex interactions and processes that take place at that level, contributing to biocomplexity. Complexity is also added by interactions among levels, both upwards and downwards, producing feedback circuits. Note: at the top level, one may place the planetary bio-geochemical regulatory system.

| Organization Level | Interactions |
|---|---|
| ecological communities | nutrient cycling, environmental regulation |
| populations, species | competition, predator-prey, symbiosis. |
| multi-cellular organisms | reproduction, social behaviour |
| tissues, organs, and organ systems | organ function, first messenger regulation |
| cells | specialization, first messenger communications |
| sub-cellular structures | cellular homeostasis |
| molecular networks | biochemical networks, second messengers |
| DNA sequences: codons to genes | evolution and expression control |
| molecular surfaces | lock and key information gates, e.g., enzymes |

### 2.3.2 Functional information – the potential for value

In this section I consider specifically functional component of biological information in isolation, discuss its quantification and the way it can be identified as the source of value in biological diversity, which is the ultimate goal of my thesis.

#### In biological organisation up to the species level

In an application of Boltzmann's entropy concept at the genetic level, Szostak (2003) defined "functional information", in terms of a gene string, as $-\log_2$ of the probability that a random sequence will encode a molecule with greater than any given degree of function – in other words a *design brief*.

In the case of genes, this "function" may be thought of as the biochemical activity (for example a digestive enzyme's cleaving rate) of whatever molecule is produced from reading the nucleotide sequence. For a practical degree of function at the DNA level, the probability of a random sequence producing greater function than the observed sequence is approximately zero. This implies that if the information content of the genome were compressed (removing repetition) one would be left with only the Functional Information Content (F(IC)), but the compressed genome is by definition the A(IC), hence for the genome F(IC) = A(IC). Despite historical references to "redundant" or "junk' DNA, substantial modern evidence points to the adaptive significance, hence function, of all extant genes – this being discussed in Barbieri (2007). It is therefore reasonable to assume that the total genetic complexity identified by gene-based biodiversity is entirely functional F(IC)=A(IC), implying that all unique elements of information at this level are potential sources of value. The phylogenetic measures of biodiversity called for by Crozier (1997), implied by Weitzman (1993) and reviewed by King (2009) are designed to quantify the number of these unique elements.

### Can sub-species level information have value?

For more than ten years, genetic information has been recognised as an important part of biodiversity (see Crozier's review, 1997). A few economists have adopted this idea to aggregate the genetic information content of an assembly of species through totaling the inter-species genetic-distance (Weitzman, 1992). This was elaborated into the "Noah's Ark Problem" (Weitzman, 1998), in which a hypothetical choice is made of which species to "save" in order to maximise the genetic information of the "Ark". The problem is expressed in economic terms as finding the optimal level of "biodiversity", given a budget constraint (or, as recently restated by Béné and Doyen (2008), find "how big Noah's Ark must be to host the optimal level of biodiversity"). Genetic differences are aggregated into a dissimilarity index and it is assumed that the greater the dissimilarity, the more desirable (hence, valuable) the biological system to which they belong, though Brock and Xepapadeas (2003) make the reasonable complaint that this assumption has not been justified. This is a problem for economists, since, unless information is to be valued in and of itself, it is not clear how maximising genetic diversity maximises welfare. In an alternative approach, Nehring and Puppe (2002) describe species in terms of attribute sets (in practice assemblies of biological traits), but their economic valuation entails a subjective choice of attributes which were selected for specific human welfare goals, rather than describing the species' ecological role. This approach returned valuation to species level and above as well as re-introducing the abstraction of organisms into a few non-living components of identifiable utility. To overcome these problems, Eppink and van den Bergh (2007) suggest using less stylised representations of ecological processes in economic models of biodiversity conservation, but the approach offered relied on hypothetical (subjective) valuation methods: a return to opinion-collection.

Despite these conceptual drawbacks, it is possible to recognise that the previously mentioned environmental economists have closely linked the idea of unique genetic information to that of function, taking it as axiomatic that genetic information may be valuable only because it codes for potentially valuable functions. These functions are normally thought of as those performed, not by genes, but by whole organisms. The conventional understanding of genetic information in biology is that it provides the "blueprint" for making the molecular components that are responsible for the complexity and functionality of all the levels between DNA and the whole organism, inclusively. Thus, differences among genes lead to (typically) functional differences in organism traits, which is to say: functional diversity. For this reason, the lower seven levels of Table 2.1 may be counted together in considering the functional uniqueness of organisms as a result of genetic-level complexity. Phylogenetic diversity may therefore be used to characterise the information content at and below species level, even though this is not directly measuring it (see Figure 2.1). The hypothesis is that this contributes to the instrumental value of biodiversity in so far as it constitutes the necessary information for the functioning of those species present in a community.

### Beyond the species level: Noah's Ecosphere

At the levels above species, Szostak's (2003) functional information approach requires a quantitative specification of the function of each system component (species), from which to find the proportion of "all possible components" which can fulfil the design brief; but what is the set of all possible components? To the extent that a biological system is composed of a set of inter-dependent components, each optimised by natural selection (for its natural environment), it is composed of approximately unique solutions (Smith, 2000). The alternative is that the biological design brief $h$ is specified sufficiently broadly that more than one available design may suffice. If that were true then the F(IC) of any observed design would, by definition, be less than or equal to its A(IC) – in all but the special optimal case $\mathrm{F(IC)} < \mathrm{A(IC)}$. In this case, designs (e.g., species) are substitutable, since there would be more than one way to achieve the design brief. Thus, one can think of the ratio A(IC)/F(IC) as a measure of substitutability. Clearly, if $h$ is specified in broad terms, such as – "this ecosystem must sequestrate $k$ tonnes of carbon per year", then there is opportunity for substitutability because F(IC) is likely to be less than A(IC) for any particular candidate ecosystem. In that case, it might be said that ecosystems are "over-specified", using engineering language. However, a human choice of $h$ is inevitably arbitrary, partial, and subjective. We are, in general, ignorant of the biological design criteria, only able to partially infer them in cases where the loss of system components (e.g., populations) has led to wider measurable ecological effects. In this sense, humanity is in the position of the first brain surgeons, learning which structures do what from studying trauma victims.

The ecologically most important development of the economic Noah's Ark idea recognises for the first time that the assembly of "saved" organisms must work together as a functioning system, not just a "zoo" (Perry, 2010), thus extending the idea to ecological levels. For Perry, Noah selects species for their functional diversity, explicitly recognising "ecological function" as the contribution to value. However, Perry (2010) defined function only qualitatively and the analysis is limited to a single function, in practice, again, leaving valuation as a subjective choice of function by the human valuer. He clearly identified the "functional uniqueness" of a population as the source of indirect use-value. For Perry (2010), substitutability defines "ecological importance", by counting the number of populations that perform an identified function in a community (the functional set **F**). His "ecological importance" measures "function" in terms of the number of populations affected (members of the affected set **A**). In practice, the network-nature of ecological communities ensures that through indirect effects, **A** contains all populations in the community. This accords with the established model of an ecological community as a system of differential equations of the form $\partial \mathbf{n}/\partial t = f(\mathbf{n})$, where $\mathbf{n}$ is the vector of *all* populations. Perry's (2010) measure further assumes that members of a functional group are quantitatively equivalent because the measure is qualitative – a population either contributes to the function or it does not. For this reason, functional populations appear substitutable, though quantitative empirical evidence contradicts that (e.g., O'Gorman and Emmerson, 2009, and references therein). Quantitatively, we would expect every population to perform some (not necessarily known) unique function, which is the direct effect of functional information, so again, F(IC) is approximately equivalent to A(IC); this time at levels beyond the species level in Table 2.1. Evidently, we need to look for functional information and therefore potentially valuable information in every level of the hierarchical organisation of life. To reflect this multi-level information, we would need a multi-level measure of biodiversity.

### 2.3.3 Summary of conceptual analysis

The conclusion from these considerations is that functional information, instantiated at multiple hierarchical levels, codes for ecological functions, which accord the biological system with the potential to be valuable (see Figure 2.1).

Biodiversity has been defined here as a measure of information, expressed as a degree of difference among constituent parts of the biological system. The total information is a mixture of functional and entropic, so biodiversity estimation for valuation must isolate and count only the functional information. This was identified as biological complexity, which can be manipulated and measured in models, but rarely directly observed in real systems. Thus, if the purpose of estimating biodiversity is to quantify indirect use value, then it must estimate the functional information content, recognising that this exists at multiple levels of biological organisation. The most direct means to estimate this is through measures of functional diversity, but below species level, functional diversity is easier to specify via the phylogenetic diversity which derives from the way it is instantiated. Above species level, it is likely that ecological structure can act as a surrogate for some difficult to identify functions. *Thus, it is proposed that for a comprehensive measure of biodiversity as functional information, in practice, all three – phylogenetic, structural, and functional diversity should be estimated and combined.* Each of these is a description of a different kind of diversity; though all ultimately reflect function, they are potentially independent axes of variation (like the primary colours in the digital image analogy). Whether or not they are truly orthogonal remains to be determined. Furthermore the combination of these measures should be organised so as to recognise the multiple hierarchical levels of biological organisation.

This now implies a way to categorise existing biodiversity measures according to what kind of biodiversity they estimate and by the organisational level to which they refer. I shall now refer to the kind of biodiversity as the descriptor and the level of biological organisation as level. With level (L) and descriptor (D), existing and hypothetical biodiversity measures can be classified in a ($\mathbf{D}|\mathbf{L}$) permutation matrix, each element of which is a different combination of the kind of biological diversity and the organisational level of its measurement. This constitutes a formalisation of the influential ideas presented by Noss (1990), based on primary attributes recognised by Franklin (1988), who incorporated the descriptor categories composition, structure, and function into a hierarchy of indices. This formal structure now provides a starting point for a more quantitative analysis of biodiversity as information, following the conceptual scaffolding I have just established.

## 2.4 Decomposition of biodiversity

To offer a formal structure for biodiversity decomposition, I shall now use the conceptual framework defined within the information-based meaning of biodiversity in the preceding sections. For this I will closely follow the logic of the digital image analogy (Figure 2.2) in applying a more quantitative approach to the elemental components of biodiversity: a descriptor (D) and a level (L). While these two concepts have been briefly introduced earlier, here they will be further elaborated. Once formalised, this formal structure for metrics decomposition will provide support for biodiversity quantification which accords with the notion of hierarchical organisation of biodiversity (see Table 2.1).

### Descriptors

Starting off with a formal definition of a descriptor I define it as following:

**Definition 2.1.** Descriptor D characterises properties of an entity that arise as a function from their variation on a certain aggregation level.

Since each individual descriptor only partially characterises the entity by referring to its certain properties, in order to obtain a comprehensive description, a set of descriptors is required. Literature review (detailed in the next chapter) shows a variety of descriptors suggesting several groups to which these elements can be ascribed. These typically refer to numbers, features, patterns, distances or functions of the entity. Grouping these elements is important, because in practice, there is an unlimited number of descriptors making it hard to disentangle them. Their (often) overlapping meanings introduce many redundancies, which can be effectively minimised by reducing the set of descriptors to its necessary and sufficient elements.

Typically, this notion of redundancy appears due to double counting of similar properties with variation between them being not strictly orthogonal. This leads to a correlation between certain groups of descriptors – for example, there is evidence that an increase in the number of species makes patterns between them more complex (Hooper et al., 2005).

This leads me to a new definition – *empirical descriptors*, denoted as $\mathbf{D}_E$:

**Definition 2.2.** Empirical descriptors $\mathbf{D}_E$ are descriptors derived empirically from the literature; these descriptors are neither elemental nor orthogonal.

From which it follows that:

**Definition 2.3.** Elemental descriptors $\mathbf{D}_O$ are descriptors that cannot be divided or reduced any further; these are necessarily orthogonal.

Using these definitions and following their implicit meaning, I group all empirical descriptors to form a tree of descriptors (Figure 2.3).

The list of descriptors shown on the Figure 2.3 is not exhaustive – it is based on a sample of the literature on biodiversity (described in the next chapter). Therefore, this structure is only intended to demonstrate empirical descriptors, with different types of biodiversity – phylogenetic, structural, and functional – implied within it. There is no clear-cut distinction between them, thus suggesting the interdependence of the empirical descriptors. The meaning of each descriptor typically used in the literature to denote biodiversity is generalised below:

1. *Numbers* represent a particular quantity and used to count the number of identical replicates. This descriptor has a wide range from species richness to number of alleles;

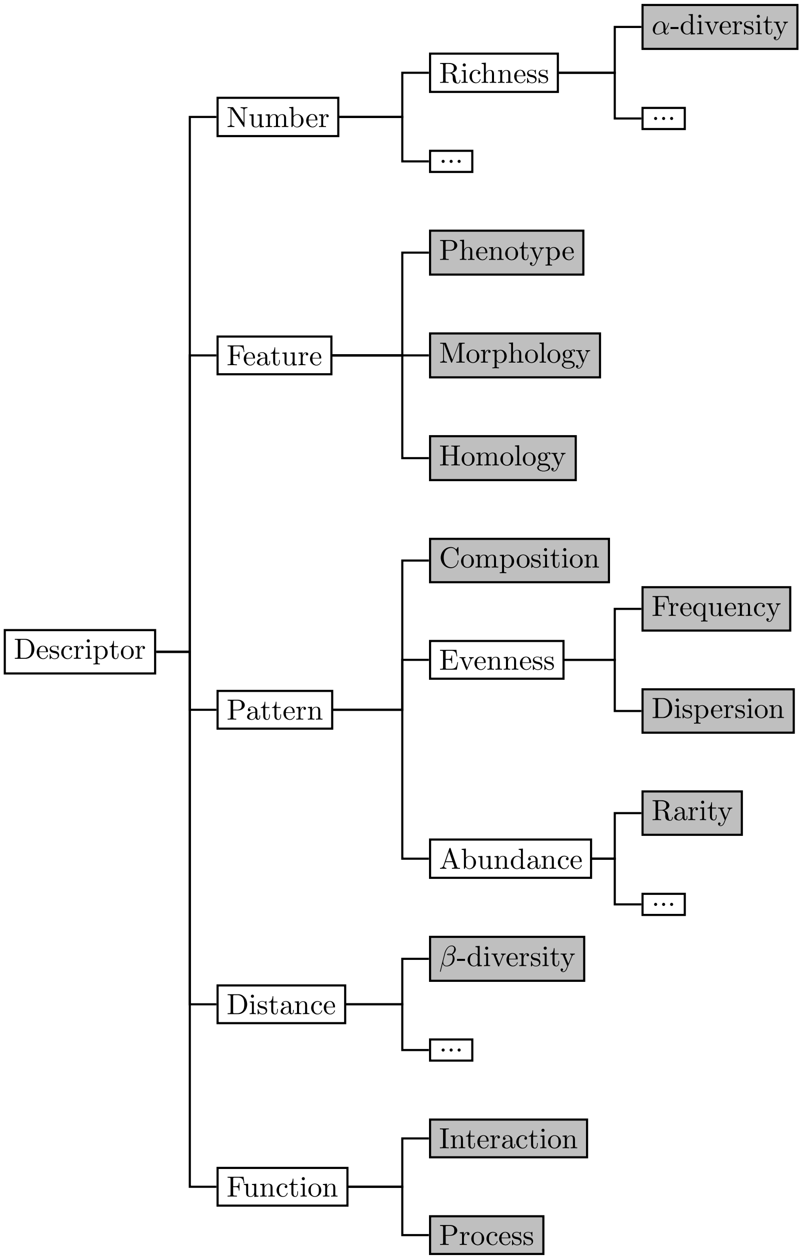


**Figure 2.3:** Decomposition of biodiversity as a metric into empirical descriptors. Empty boxes mean that potentially more descriptors can be considered

2. *Feature* is a distinctive attribute or aspect of something, also referred to as trait. Terminal nodes of this descriptor are phenotype, morphology, and homology. Phenotype is a set of observable characteristics of an entity resulting from the interaction of its genotype with environment. Morphology describes forms and relationships between structures. Homology (e.g., genetic) is a similarity in sequence of a protein or nucleic acid. Examples of these descriptors include genetic markers, phenotypic difference or morphological difference;

3. *Pattern* is a way something is organized, modelled or designed. There are four terminal nodes: composition, frequency, dispersion, and rarity that can be described by, e.g., species assemblage, landscape pattern or frequency of alleles;

4. *Distance* is a description of how far entities are from each other referring to physical length, time length or other criteria like distance on taxonomic tree. Examples are pairwise species distance and taxonomic distance; and

5. *Function* is a sequence of changes in properties of an entity and their relationships which result in a certain output. Terminal nodes are interactions and processes. One output can be achieved through various interactions and processes.

Note that these groups may not be strictly orthogonal, and used here only as an illustration of the variety of descriptors, while introducing an important distinction between empirical and elemental descriptors.

### Empirical versus Elemental descriptors

The descriptors on the Figure 2.3 are empirical descriptors, which in fact, describe the literature rather than the concept of biodiversity. Empirical descriptors convey some information about underlying biodiversity, and they can be seen as a subset of a necessary and sufficient set of descriptors for quantifying total biodiversity as total functional information. Is it possible to filter out orthogonal and elemental descriptors from empirical descriptors? It can be achieved by making the best use of existing empirical knowledge on measures of biodiversity and analysing any patterns among empirical descriptors. The relationship between $\mathbf{D}_E$ and $\mathbf{D}_O$ is shown on Figure 2.4:

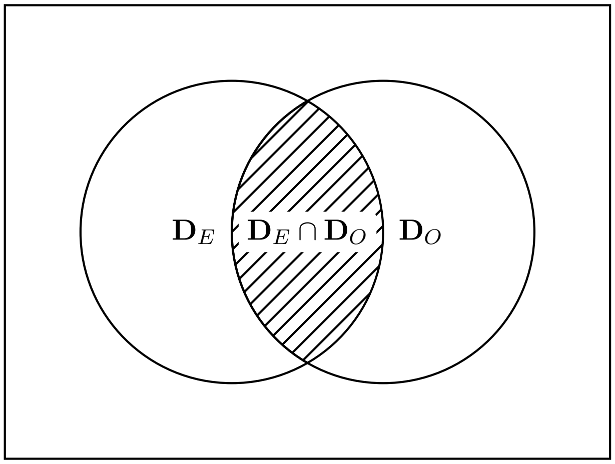


**Figure 2.4:** An intersection between empirical ($\mathbf{D}_E$) and orthogonal ($\mathbf{D}_O$) descriptors illustrates those empirical descriptors that are also orthogonal

### Levels

**Definition 2.4.** Level L is an aggregation level or position in an hierarchy of biological organisation.

Various aggregation levels on which biological entity might be considered are shown on Figure 2.5. They are generically grouped here into taxonomic and functional levels, including their sub levels. Note, that this representation of levels differs from that outlined in Table 2.1. This is an empirically-derived classification with its levels by definition being not strictly orthogonal. An advantage of using this classification for biodiversity decomposition is that it can readily be used for coding empirical measures of biodiversity.

Taxonomic level is based on a formal classification of all organisms from species on lower level to kingdoms on higher level. Additionally to these formal taxonomic ranks I consider genetic level, although, strictly speaking, it does not belong to taxonomic hierarchy. The reason for specifying genetic level under taxonomic level is that biological foundation of taxonomy is based on genetic variation of organisms. It is apparently the best effort to map this variation and as taxonomy becomes closer to an evolutionary tree it also becomes closer to genetic variation. There are some problems with this idea though as genetic level is not well enough understood yet. Often organisms that are very similar taxonomically are not necessarily similar genetically. For example, Taylor and McPhail (1999) studied morphologically similar species pairs of threespine stickleback (*Gasterosteus aculeatus*) that co-exist in several lakes in British Columbia. Based on examination of mitochondrial DNA, the authors concluded that species have evolved independently, thus confirming their genetic dissimilarity. Similar findings were demonstrated in Cano et al. (2008), Alexandrou et al. (2011).

Functional level can be subdivided into two distinct concepts which depend on (1) specific role that organisms play; and (2) aggregation of organisms that play specific roles. Among the roles that an individual organism can play I distinguish producers, consumers, and recyclers. Consumers can be further subdivided into trophic levels. Organisational levels can be in a form of communities, metacommunities, biome, and ecosystem. Community, best defined as an ecosystem without abiotic part, is not simply a collection of species – there must be meaningful interaction among them. Metacommunity is a collection of communities. As a result of interaction between community and biogeographical factors (e.g., climatic variables), biome will be used. Finally, on the last functional level I have ecosystem which is not only a group of organisms but also assemblies of functions that work together to coordinate it.

An alternative approach to classification of levels of biological organisation has been considered by Sarkar and Margules (2002), who suggested two hierarchical schemes to classify biological entities: (i) a spatial ecological hierarchy starting from biological molecules to

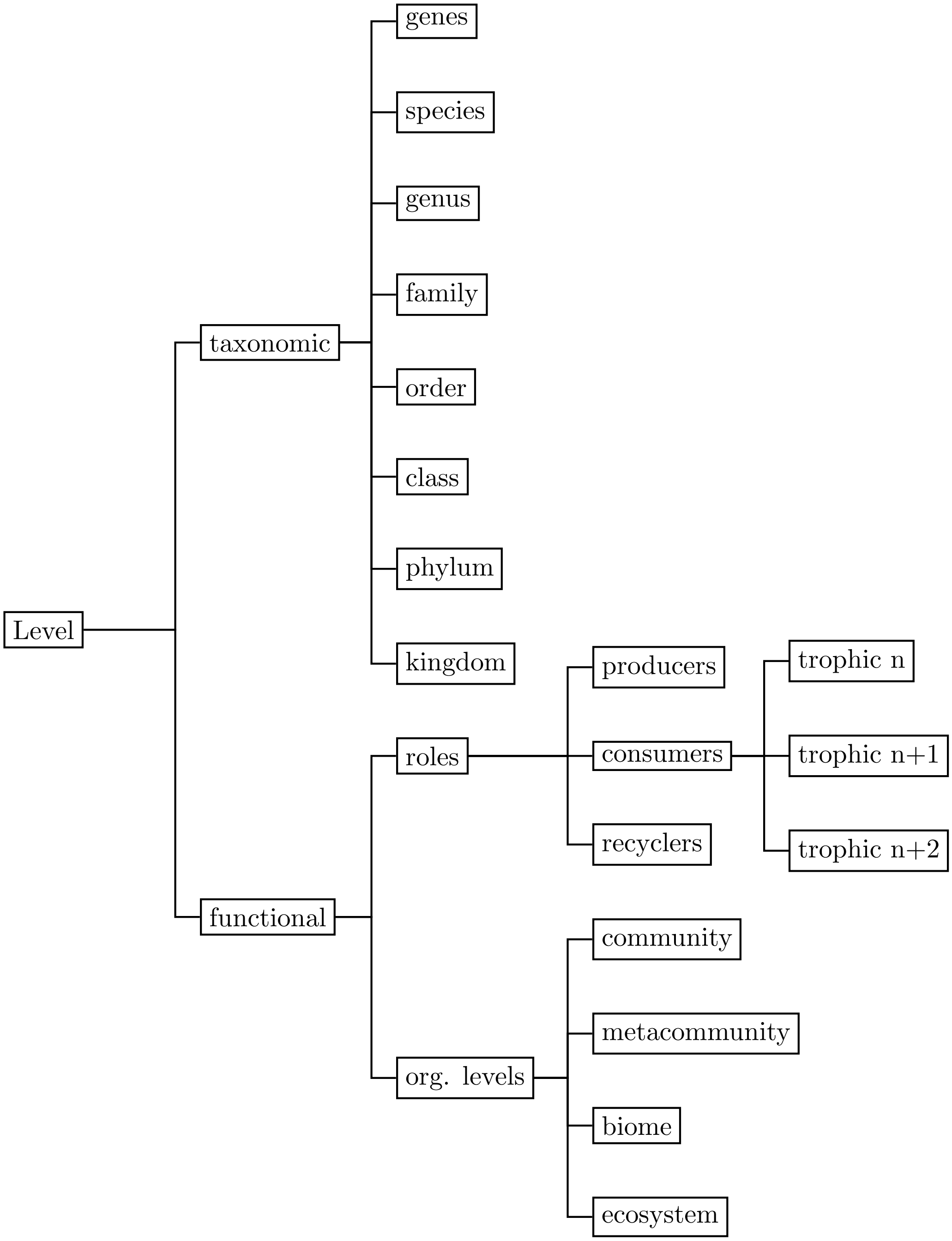


**Figure 2.5:** Decomposition of biodiversity as a metric into empirical levels. Here “trophic n” for consumers refer to a group with a specified trophic level

populations, meta-populations, and communities; and (ii) a taxonomical hierarchy spanning from genotypes to species, and kingdoms. The approach suggested here combines both hierarchical schemes as a tree of levels of biodiversity and accounts for the functional diversity.

## 2.5 A formal structure

Having introduced descriptors and levels as attributes of a more fundamental concept I shall further generalise this idea by proposing that *any* descriptors and *any* levels expressed as combination D|L generate a measure of biodiversity. Now, putting it more formally:

**Definition 2.5.** Measure $\mathrm{M}_{i,j} \equiv(\mathrm{D}_i|\mathrm{L}_j)$ is a scalar combination of one descriptor D at one level L specifying a component of biodiversity. The vector measure **M** is the set of all components consisting of all possible combinations $\mathrm{M}_{i,j}(\forall i, j)$ representing their projections on the coordinate axes.

Note, D|L may be null because not all combinations are presented in the literature, so that **M** may be sparse. The total set of measures is finite and yet unknown.

Let $\mathrm{D}_i$ denote a $i \times 1$ column matrix of all possible descriptors **D** and $\mathrm{L}_j$ is a 1× j row matrix of all possible levels **L**. A product of these two matrices is matrix **M**. Expressed as $\mathrm{D}_{i,1}\otimes \mathrm{L}_{1,j}$ with dimension $i \times j$ it contains all possible D|L combinations.

$$\mathbf{M}_{i,j} = \begin{pmatrix} M_{D_1L_1} & M_{D_1L_2} & \dots & M_{D_1L_j} \\ M_{D_2L_1} & M_{D_2L_2} & \dots & M_{D_2L_j} \\ \vdots & \vdots & \ddots & \vdots \\ M_{D_iL_1} & M_{D_iL_2} & \dots & M_{D_iL_j} \end{pmatrix} \tag{2.2}$$

Since some elements are zeros, this matrix is sparse, which implies that not every combination of level and descriptor exists or known. In general, matrix **M** is not commutative (i.e., $\mathbf{D} \otimes \mathbf{L} \neq \mathbf{D} \otimes \mathbf{L}$), but elements of matrix $\mathbf{M}_{D|L}$ are identical to $\mathbf{M}_{L|D}$, regardless the order of its elements – permutation equivalents.

Measures M can be combined to form an index of biodiversity, defined as follows:

**Definition 2.6.** Let $\mathbf{M}'$ be a subset of **M**, then $\mathcal{I}\equiv f\ (\mathbf{M}')$ is an "index".

Index is a scalar combining the member measures of $\mathbf{M}'$ to represent biodiversity as composite measures expressed in a mathematically meaningful way. Similarly to measures, the total set of indices is infinite and unknown. Now, armed with definitions for measures and indices I can make the following proposition:

**Proposition 2.1.** Any two indices of biodiversity that are composed using *the same* D|L elements can be mathematically derived from one another.

This will have an important implications for analysing different biodiversity metrics and redundancy among them in the following chapters of the thesis.

## 2.6 Diversity in space

Up to this point, the meaning of biodiversity has been the diversity within a particular biological system. In practice, many ecologists consider biodiversity as an attribute of a region of space (usually an area of land). Certainly, spatial variation of biological material has been one of the major interests of biodiversity research, in fact it even appears as one of the early definitions of ecology (as "the study of the distribution and abundance of organisms": Krebs (1972)), referred to by Magurran and Dornelas (2010). It is therefore necessary to show how the metrics which I have been discussing and which will be elaborated in subsequent chapters can be interpreted as spatial variables. To achieve that, I now briefly elaborate the key spatial biodiversity concepts from the axiom of (bio)diversity as (bio)information. Since I have now concluded that only functional information is responsible for ecosystem services, the meaning of information in what follows should be taken as strictly functional (meaningful), non-random information.

Let $\mathbf{a}$ be a set of unique units of functional information (f.i.): $\{a_i\}\,[i = 1 \cdots N]$, where $N$ is a finite positive integer. Let each member of $\mathbf{a}$ be distributed in space $\mathbf{z}$, with density $\rho_i(\mathbf{z})$, such that the probability density of $a_i$ is $\rho_i(\mathbf{z})$ (i.e., it is scale-free), then within a finite region of space $\Omega$, the prob. of $a_i$ is:

$$\mathrm{p}(a_i|\Omega) = \int_\Omega \rho_i(\mathbf{z})d\mathbf{z}. \tag{2.3}$$

Based on this, the total *expected* functional information within $\Omega$ is:

$$\mathrm{E}\left[I_\Omega\right] = \sum_i^N q_i \int_\Omega \rho_i(\mathbf{z})d\mathbf{z}, \tag{2.4}$$

where $q_i$ is the quantity of functional information contributed by the $i$th unit.

As $\Omega \rightarrow \mathrm{U}$ (the universal space), $\mathrm{p}(a_i|\Omega) \rightarrow 1, \forall i$, so $\mathrm{E}\left[I_\Omega\right] \rightarrow I$: the total f.i. in the universal space. $\mathrm{E}\left[I_\Omega\right]$ is the quantity estimated by the traditional $\alpha$-diversity.

Consider two regions of space $\Omega_1$ and $\Omega_2$, for which we can estimate the combined total functional information content. Using the probability "addition rule":

$$\mathrm{E}\left[I_{\Omega_1+\Omega_2}\right] = \mathrm{E}\left[I_{\Omega_1}\right] + \mathrm{E}\left[I_{\Omega_2}\right] - \mathrm{E}\left[I_{\Omega_1\wedge\Omega_2}\right], \tag{2.5}$$

(with standard notation: $\Omega_1 \wedge \Omega_2$ to represent information appearing in both regions). Well known rules for adding probabilities from $k$-trials, readily generalise this to multiple regions ($k > 2$) to obtain $\mathrm{E}\left[I_{\mathbf{\Omega}_k}\right]$, where $\mathbf{\Omega}_k$ is a set of $k$ regions (Note: McGill (2011) applies this concept to grid-based atlas data). This is the quantity which traditional $\gamma$-diversity attempts to estimate (from the original definition by Whittaker (1960), interpreted in Tuomisto (2010a) and also McGill (2011)). In set notation, for the two-region example, the quantity $\mathrm{E}\left[I_{\Omega_1+\Omega_2}\right]$ estimates $\mathrm{E}\left[I_{\Omega_1}\right] \setminus \mathrm{E}\left[I_{\Omega_2}\right] + \mathrm{E}\left[I_{\Omega_2}\right] \setminus \mathrm{E}\left[I_{\Omega_1}\right]$, which counts everything that is unique to each region.

If $\rho_i(\mathbf{z})$ is uniform in $\mathbf{z}$ $\forall i$, then aggregating regions over $\mathbf{\Omega}_k$ has the effect of increasing (effective) region size only: $\mathrm{p}(a_i|\mathbf{\Omega}_k)$ increases with $k$, following equation (2.3). Cases with non-uniform $\rho_i(\mathbf{z})$ are more interest to biodiversity research because biota are uneven in spatial distribution. As a gentle introduction, consider a 1-dimensional system $U$ in which only two units of f.i. exist: $\mathbf{a} = \{a, b\}$, distributed with probability density: $\rho_a(\mathbf{z}) = 1 - sz$ and $\rho_b(\mathbf{z}) = sz$, scaled s.t. $z$ is constrained to $[0, 1]$. In this example, take two regions: $\Omega_1 : z[0, 0.1]$ and $\Omega_2 : z[x, x + 0.1]$, with a "separation variable" $x[0.1, 0.9]$.

If $s = 1$,

$$\mathrm{p}(a|\Omega_1) = \int_0^{0.1} 1 - z \, dz = 0.095 \tag{2.6}$$

and

$$\mathrm{p}(b|\Omega_2(x)) = \int_x^{x+0.1} z \, dz = 0.005 + 0.1x, \tag{2.7}$$

similarly $\mathrm{p}(b|\Omega_1) = 0.005$ and $\mathrm{p}(a|\Omega_2) = 0.095 - 0.1x$, so, letting $q_a = q_b = 1$, for simplicity, using equation (2.4), $\mathrm{E}\left[I_{\Omega_1}\right] = 1$ and $\mathrm{E}\left[I_{\Omega_2}\right] = 1 \forall x$: because of the symmetry of this simple example, both regions have the same expected information content. To combine them using equation (2.5), we need the sum of joint probabilities:

$$\mathrm{E}\left[I_{\Omega_1 \wedge \Omega_2}\right] = \sum_i^N q_i \left[ \int_{\Omega_1} \rho_i(\mathbf{z}) d\mathbf{z} \int_{\Omega_2} \rho_i(\mathbf{z}) d\mathbf{z} \right], \tag{2.8}$$

in this particular example,

$$\mathrm{E}\left[I_{\Omega_1 \wedge \Omega_2}\right] = \mathrm{p}(a|\Omega_1)\mathrm{p}(a|\Omega_2) + \mathrm{p}(b|\Omega_1)\mathrm{p}(b|\Omega_2), \tag{2.9}$$

which evaluates to $0.00905 - 0.009x$ , so $\mathrm{E}\left[I_{\Omega_1+\Omega_2}\right]$ rises linearly, though only slightly, from $1.99185; (x = 0.1)$ to $1.99905; (x = 0.9)$ as the separation distance between regions $x$ increases. This is an example of "decay of similarity with distance" relation studied by, e.g., Nekola and White (1999).

In general, of course, the distribution of information units in space is not nearly so simple. The introductory example helps to explain the meaning of the terms in a general statement of spatial diversity:

$$\mathrm{E}\left[I_{\boldsymbol{\Omega}_k}\right] = \sum^{k}\left[\sum_{i}^{N} q_i \int_{\Omega} \rho_i(\mathbf{z})d\mathbf{z}\right] - \sum_{i}^{N}\left[\prod^{k} q_i \int_{\Omega} \rho_i(\mathbf{z})d\mathbf{z}\right], \tag{2.10}$$

which finds the total ($\gamma$-diversity) information over a set of regions in space, in which a set of discrete information units **a** is distributed.

So far, I have not mentioned the $\beta$-diversity, which traditionally represents the difference among regions as a whole, ideally independent of their individual $\alpha$-diversities. This notion immediately raises two questions relating to information content. Firstly, are differences among regions due simply to the assembly of functional information units captured within these regions, or do they include higher-level organisational information, treating regions as *systems*, as for example by McGill (2011)? Secondly, how can $\beta$-diversity differ from $\gamma$-diversity? The answer to the first question depends on the meaning of regions – if representations of distinct ecological communities, then maybe higher order information can be included, but if, as is often the case, they are simply separate samples from the environment, then they should be treated as assemblies (this uncertainty demonstrates the vagueness often complained about in discussions on $\beta$-diversity). Whether or not regions are treated as systems, with their own added organisational, community-level information, the answer to the second question is that, in set notation:

$$\mathrm{E}\left[I_{\boldsymbol{\Omega}_k}\right] = \sum_{j}^{k} \mathrm{E}\left[I_{\Omega_j}\right] \setminus \mathrm{E}\left[I_{\Omega_{\bar{j}}}\right], \tag{2.11}$$

where $\Omega_{\bar{j}} = \bigcup_{i \neq j}^{k}(\Omega_i)$, which is identical to the definition given above for $\gamma$-diversity (that is, the total expected information) distributed among the regions. To be clear, this way of thinking finds no difference between $\beta$- and $\gamma$-diversity. The reason the information-based perspective sees no distinction between $\beta$- and $\gamma$-diversity is twofold. Firstly, it formally specifies information as difference, so measures of difference (diversity) are identical to measures of information content (Tuomisto, 2010a) recognised that $\beta$-diversity indices are inconsistent in units and even concepts). Secondly, though traditional $\beta$-diversity indices are all calculated from the differences in taxon abundances among regions, the information approach calculates only from presence-absence data, focussing attention on the functional information found *within* the biological system. Certainly, heterogeneity in species abundance is information, but it is information *about* the system, not functional information coded within it.

This analysis leads to the conclusion that the large number of definitions for $\beta$-diversity, found for example in Tuomisto (2010a,b) and Anderson et al. (2011), which were designed to describe field data, do not directly describe variation in the functional information diversity of a biological system. On the other hand, $\gamma$-diversity (also primarily designed to summarise field data) is a well defined concept which is compatible with and so serve as an estimator for the information content coded within the biological system. If, however, $\beta$-diversity is defined following Whittaker (1960) with its original meaning of $\gamma/\alpha$-diversity, then it is useful as a measure of how well the information content of the whole space has been sampled. It is striking how the large literature concerning $\alpha$-, $\beta$-, and $\gamma$-diversities has concentrated on deriving scores for comparing communities based on the numerical distribution (common or rare) of species among them. Very little has been said about other, perhaps more meaningful, characters of community structure, such as foodweb connectance (Dunne et al., 2002). Genetic and functional variation in space, are even more neglected, other than through the surrogate of species identity. What follows in the next chapters will show that these forms of difference are important to the total of biodiversity. The question of spatial variation is one of defining system boundaries. Under the information interpretation of biodiversity, it is legitimate to consider information density as the probability of finding a given amount of information over a defined region of space, enabling diversity calculations via integrating over space.

## 2.7 Discussion and Conclusions

A connection between biodiversity and information is now well established in the form of various indices of biodiversity inspired by the communications theory of Shannon (1948). Pioneering authors regarded biodiversity as the measure of information contained in an assembly of organisms (see, e.g., Margalef, 1958), this information being the raw material (i.e., data) for ecological study. Biodiversity indices would quantify the information yield, but crucially, this meant information *about* the system, or a sample of it, not the information *embodied within* the system. The quantitative value obtained would depend on sample size, sampling effort, and the arbitrary choice of categorising level (e.g., species, genes, or higher order systems): it described the information perceived by the observer. There is very much more information present in a biological system than can be counted by simple observation, so its quantification via counting species or even genes amounts to gross bias by discarding. As well as missing a great deal of the information present, biodiversity indices based on such observations may be sensitive to information that has no functional significance. This is especially the case in recording abundances, because particular abundances in a sample are only “snapshots” of constantly changing variables, taken at an arbitrary time. System-level information is held in the relationships governing these variables, not a set of their values at any particular time. Again, the problem can be identified as one of registering information about the system, rather than within it. The information within the system is the form of relationships among its components, not a transitory count of these components. Recognising the distinction, Bateson (1972) called information “difference that makes a difference” and this concept can be used to distinguish between functionally significant and random information, with a view to isolating the former in biodiversity measures. The justification for that is that information which makes no difference, by definition contributes nothing to instrumental value.

Species richness and related metrics are often used as a practical surrogate for biodiversity (see, e.g., Aubert et al., 2003; Tian et al., 2007; Joshi et al., 2008; Moreno et al., 2008; Campos and Fernando Isaza, 2009). To manage the full spectrum of biodiversity (Redford and Richter, 1999) at every level of biological organisation and to go beyond subjective preference for "the cute and the cuddly" (Mace et al., 2003), and "furries and featheries" (May, 1994) a more clear understanding of the deeper meaning of biodiversity is required. Given this, it is clear that biodiversity is not species themselves, it is rather a concept that attempts to capture the meaningful information held in the system composed of those species, when information itself is very hard to quantify.

In ecological economics, information is seen by some as a primary source of the value in biodiversity, motivating phylogenetic information measurement (Faith, 1992), or counting species as information (Weikard, 2002). Closely related is the idea of biodiversity-information as an insurance against loss of ecosystem services (Baumgartner, 2006). Practical economic applications have so far been limited to highly specific contexts (Brock and Xepapadeas, 2003), probably because links between future welfare and biological information are typically obscure. If the question is limited to one of choosing (from a set) which ecological community is to be preserved, then Weikard's (2002) application of the Noah's Ark problem at the species level can objectively guide decision makers. The "ecosystem" distance measure he proposed is effectively the complementarity measure presented by Faith et al. (2004), but without the need for phylogenetics – an important advantage given our very incomplete knowledge. Weikard (2002) pragmatically replaces ecosystem information content with species counts, whilst acknowledging that the true information store lies at genetic, species, and system levels. It will be noted in the next chapter that taxonomy may provide a suitable surrogate for phylogenetic data, enabling below-species level functional information to be represented.

The recognition of descriptors as independent axes of variation, representing distinct types of biodiversity and levels, representing the hierarchical organisation of living systems led to a compact summary of biodiversity through the **D**|**L** permutation matrix. Whilst this is a novel formalisation, the idea of biodiversity decomposition is not new. According to DeLong (1996), for instance, the definition of biodiversity terms should consist of two parts: class and differentia. Class identifies the group that includes the term and the differentia distinguish object from all other members of that class. Noss (1990) distinguished composition, structure, and function. It has also been proposed to measure biodiversity in terms of different components (e.g., genetic, population/species, community/ecosystem) and attributes (e.g., composition, structure, function) (Redford and Richter, 1999). Components could be related to levels and attributes to descriptors, but the great advantage of the formal decomposition of biodiversity into a set of descriptor-level couplets is that this enables quantitative analysis of all biodiversity measures and indices within a common framework: This is the topic of the next chapter.

In the course of the chapter many contextual and implicit definitions were replaced by explicit ones. Recalling the research objective – I have looked into the real meaning of biodiversity and established logical connections between different concepts related to biodiversity – measure, index, level, and descriptors. This newly proposed definition can be appropriately used on any level of abstraction (hierarchical level) without a change in its meaning. The conclusion can be made that the information-based definition of biodiversity does not depend on the context of use – and it is, therefore, capable of being truly scientific and objective.

# Summary

1. The current state of understanding of what constitutes biodiversity is fragmented with no agreed definition. A diversity of meanings encompasses a diversity of measures of biodiversity;

2. Biodiversity is a multi-dimensional concept that can be decomposed into elements, each being a member of a class, of which there are two: termed “descriptor” (symbol D) and “levels” (symbol L);

3. Every possible index of biodiversity can be expressed as a combination of D|L components;

4. Biodiversity, at its most basic, is a measure of the degree of difference among constituent elements of the biological system, which is information;

5. The functional fraction of this – functional information – gives a system its potential use-value, so is a concrete, intrinsic, and system-independent currency for biodiversity;

6. Functional information can be measured from identifying and quantifying pattern (systematic difference) within biological systems.

# Chapter 3

# Knowledge engineering biodiversity

"Biodiversity is an abstract concept only facets of which can be made operational and measured"

*Gaston, K.*
*Biodiversity: a biology of numbers and difference*

## 3.1 Introduction

Since biodiversity loss and conservation have become scientifically important, empirical studies on measures of biodiversity have increased greatly (e.g., Tian et al., 2007; Zamora et al., 2007; Joshi et al., 2008; Anderson, 2008; Moreno et al., 2008; Sharma and Rawat, 2009). Published literature provides empirical evidence of biodiversity estimates varying in their scale, units, objects of the study (i.e., study systems), and their environments. Manifested at different organisational levels and described in a variety of ways, empirical biodiversity knowledge consists of multidimensional data represented by many (possibly) interrelated measures (Escarguel et al., 2011). So, given the variety of biodiversity appearances, when scientists measure it do they really have a common trait in mind? An answer to this question comes from the literature reviews of DeLong (1996) and Feest et al. (2010), which strongly suggest that it is not always the case.

The results of this chapter were presented at International Conference on Biodiversity Informatics, London, June 1-3, 2009

### Meta-analysis

A formal framework of statistical meta-analysis is required to synthesise in quantitative terms the empirical results from different biodiversity studies (Nijkamp et al., 2008). It is a starting point for integrating biodiversity knowledge across all known metrics and creating multidimensional diversity. As a tool, meta-analysis has been developed initially for quantitative generalisations in various disciplines (notably in medical sciences Dickersin et al., 1994; Villar et al., 1995; Egger et al., 1997; Graudal et al., 1998; Sutton et al., 2000; Turner et al., 2008) with its relatively recent extensions to biodiversity. Some of the applications include statistical generalisations of economic value of biodiversity (Brander et al., 2007; Nijkamp et al., 2008; Richardson and Loomis, 2009), species richness (Mittelbach et al., 2001; Cardinale et al., 2006; Vanderwel et al., 2007; Felton et al., 2010; Paillet et al., 2010; Prieto-Benitez and Mendez, 2011) or genetic diversity (Reed and Frankham, 2001, 2003). All these meta-studies, posing questions related to different aspects of biodiversity, demonstrate that the distribution of target variable – biodiversity estimate seems to be significantly heterogeneous.

This target variable is called an “effect size” (Gurevitch and Hedges, 1999), and in a typical meta-analysis it expresses some kind of effect across comparable studies (e.g., blood pressure level). In contrast, meta-analysis of biodiversity literature, given a variety of measures and, hence, their estimates, has a diversity of objects under analysis. Measures, arising at different levels (e.g., species, genes, ecosystems) and characterising different descriptors (e.g., composition, abundance), pose certain difficulties in generalising the overall effect size – total biodiversity. This requires any variation among biodiversity estimates and sources of this variation to be addressed explicitly. At least four sources of variation were envisioned by Osenberg et al. (1999); these include experimental, parametric, functional, and structural variation (Table 3.1). All these are attributed to a meta-analysis itself, and therefore can be considered as a some sort of measurement bias which ideally we want to control.

Meta-analytic patterns and robustness of findings can also be affected by data selection criteria. Using the results of stream predation experiments Englund et al. (1999) suggested minimising the use of selection criteria that are based on judgements of study quality. Another useful conceptual overview highlighting problems that need to be addressed can be found in Osenberg et al. (1999). In their paper, authors specify models that develop metrics of effect size and relate them to underlying ecological processes by examining their systematic variation across systems and conditions. Meta-analysis as a tool in quantitative reviews has been also addressed by Gates (2002), who concluded that methods to reduce bias and enhance the accuracy of the findings traditionally used in medical research, are still rare in ecological research. This variability can be minimised by adopting methodological developments (e.g., accounting for types of biodiversity studies and strength of biodiversity

**Table 3.1:** Types and sources of variation in biodiversity estimate. Adapted from Osenberg et al. (1999)

| Type of variation | Source |
|---|---|
| Experimental variation | caused by the way biodiversity measure were obtained, study system type, as well as conditions and manipulations |
| Parametric variation | due to variation among measures of biodiversity which depends on their types and indices calculated from them |
| Functional variation | occurs when shape of interaction between level and descriptors has different underlying function |
| Structural variation | arise when biodiversity estimates are derived from different study systems |

estimate they provide).

In studying patterns of variability in biodiversity estimates, my aim is to establish a statistical aggregation of multiple measures of biodiversity and their quantitative estimates over biodiversity *in general*. The presented analysis, therefore, is not restricted to any specific dimensions of biodiversity, nor any specific system: its goal can be described threefold:

1. to construct an aggregate quantitative measure of biodiversity through re-integration of the single dimensions into the multidimensional concept of biodiversity;
2. to examine the relationship between estimates of biodiversity and data-structural factors which may affect these results; and
3. to explore the patterns of variation in biodiversity estimates across biological levels and descriptors.

Answering all these questions presents a considerable research challenge, since current evidence shows that extending meta-analytic techniques to biodiversity might be complex (see, e.g., Nijkamp et al., 2008). The great variety of ways in which biodiversity may be characterised requires consistency in data extraction. For meta-analysis to succeed, a unifying framework providing both an accurate representation of the concept and sufficient flexibility to accommodate the variety of measures of biodiversity is needed. This calls for a research tool which should ideally allow a decomposition of the conceptual model of biodiversity, into a set of elemental constructs. The formal discipline of Relational Database (RDB) with its Entity Relationship Diagram (ERD) has been shown to fulfill these requirements (Teorey et al., 1986; Watson, 2006) and is therefore used as a practical and conceptual support for this part of my study. In the next section I will briefly justify this choice.

### The use of databases in biodiversity research

The use of RDB to integrate and structure the empirical biodiversity knowledge is a part of a broader concept of machine learning research – knowledge engineering. It involves integrating knowledge into computer systems that solves complex problems normally requiring a high level of human expertise (Fox, 2011). Even though, machine learning research attracted great interest for a long time (see, e.g., Elliot et al., 1995; Studer et al., 1998), the novelty of the present approach is its application to empirical biodiversity knowledge. This is achieved by modelling and implementation of relational infrastructure with the aim to realise the problem-solving capabilities of biodiversity quantification through its meta-analysis.

Recently, several initiatives have begun constructing digital biodiversity resources or integrating existing ones (e.g., WBD, 2010; GBIF, 2010) both on a local and international scale. In most cases databases are constructed with a view to cataloguing natural history and survey collections where biodiversity is typically considered at species level. As a result, such databases represent species banks covering a wide variety of organisms. My present application of RDB construction differs from these others in a number of ways. One of the most important distinctions is that this is the database of measures of biodiversity aiming to represent as many biodiversity levels as possible, without limitation to either species level or any particular study systems. If the intention is ultimately to have all biodiversity estimates collected together in one quantitative searchable database, then it needs to be robustly constructed.

While ensuring that the ecological meaning is retained, I explicitly organise biodiversity data into a set of interrelated entities (biodiversity concepts) with a finite set of attributes. This organisation precisely matches definition of the commonly used conceptual design tool – an ERD (Chen, 1976), which allows use of “entity” in RDB terminology to refer to an abstraction of biodiversity concepts. Based on Teorey et al.’s (1986) review, the merits of ERD in providing a unifying structure for biodiversity data across different study systems, environments, descriptors, and their levels become evident. Agrawal et al. (1993) outlined rules that underly this structure as a set of relationships or associations. Developing directly from this, Nijkamp et al. (2008) first explicitly applied RDB association rules for biodiversity science to examine variation in comparable biodiversity studies through meta-analysis of economic valuations of biodiversity.

Inspired by this recent application, I take a step further and consider a battery of measures of biodiversity and their estimates. Given the information definition of biodiversity, this implies accounting for multiple components of biodiversity which are estimated by different measures and indices. It is important to maintain their separate identities in an organised framework since treating each kind of estimate as interchangeable with the others, often results in pseudo-replication and consequent redundancy. As Jost (2006) reported, the concept of biodiversity may be confounded by the multiplicity of indices used to measure it, risking it becoming meaningless.

To minimise the effect of this conceptual pitfall I map each abstraction of biodiversity (descriptors, levels, measures) as a separate "entity" on the ERD. The formality of this treatment lies in precision and completeness in capturing the logical relationships while keeping the level of redundancy at minimum. The formal steps of RDB design (such as data normalisation, atomic values and links between different entities established via primary-foreign keys, see, e.g., Teorey et al., 1986; Watson, 2006) facilitating use of the D|L structure to distil biodiversity data from the literature.

To summarise, the formal procedure of RDB facilitates comparison and integration of biodiversity studies and their estimates, which is otherwise hindered by the variety of measures used to quantify biodiversity and the variety of ways in which these measures were applied (Koleff et al., 2003). The organisational discipline of RDB, combined with data mining techniques, ideally supports the meta-analysis of the biodiversity literature, an overall purpose of which is to collect as much empirical evidence on measures as possible in a structured and consistent manner.

## 3.2 Methods

The necessary biodiversity data was first harvested from the published literature by systematic search detailed in Section 3.2.1. A RDB of measures of biodiversity was constructed to accommodate the empirical biodiversity literature, as described in Section 3.2.2 and, finally, using data mining techniques, patterns in biodiversity estimates across studies were analysed as reported in Section 3.2.3.

### 3.2.1 Biodiversity literature search

Sources that were searched for biodiversity studies with quantitative estimates for measures and indices are shown in Table 3.2.

**Table 3.2:** Online databases being searched for biodiversity literature

| | |
|---|---|
| ScienceDirect | www.sciencedirect.com |
| Web of Science | www.isiwebofknowledge.com |
| JSTOR | www.jstor.org |
| SpringerLink | www.springerlink.com |
| Scirus | www.scirus.com |

To collect relevant biodiversity publications random and specific searching strategies were used. The former was achieved by conducting Boolean searches with different combinations of general search terms ("biodiversity" OR "biological diversity" OR "diversity" OR "biodiversity value") AND ("measure" OR "index" OR "estimate" OR "value"). Direct search was augmented by backward/forward "snowball" search. The latter strategy, related to a more specific or refined search, was based on the outcome of the random search and was intended to fill the gaps in the representativeness of the estimates. Thus, specific indices and measures of biodiversity, and their quantitative estimates were targeted.

Two different strategies were needed to gain a comprehensive sample: the first – random search – characterises the empirical distribution of measures of biodiversity across ecological communities, reflecting the frequency with which they appear in the literature. The second strategy – specific search – ensured that all common measures of biodiversity are represented, even though some are rare.

Englund et al. (1999) discussed the importance of data-selection criteria, related to the relevance, quality, and independence of data in ecology. Using meta-analysis, it was shown, that with the same research question, using different criteria to screen studies and select data within studies, different quantitative generalisations may be produced. Englund et al. (1999), argued that selection decision should be based on the relevance of the data, advocating the use of 'content relevance criteria'. Practically, this warns the researcher that if the question of interest and the chosen metric of effect size are unclear, the relevance of the selected data might be unclear too. Following this in my study, biodiversity publications were rejected if they did not report any quantitative estimate (i.e., effect size) and also the number of replicates. Studies with contrasting estimates addressing different L and D were especially valuable. Wherever it was possible, the study-system, duration of study, definitions of measures used, identification of level and descriptor in biodiversity metrics, additional indices and forms of relationships between them were all recorded as meta-data. It was anticipated that to produce an example dataset which is large enough to illustrate the use of RDB as a unifying framework for the meta-analysis of biodiversity literature the minimum number of studies collected by the literature review (having met all inclusion criteria) should be at least 50 (see Appendix 3.A for a list of studies).

### 3.2.2 A relational view of biodiversity data

This section is concerned with the application of relational database modelling tools to the concept of biodiversity, following the convention of Codd (1970) applied to biodiversity: the "relational view" is a set of concepts of biodiversity that are organised in some meaningful way. To conceptualise different measures of biodiversity in a consistent way, I apply D|L structure to data derived from the empirical biodiversity literature, which I then integrate as a relational database of measures of biodiversity (biodivDB). Further, following Codd

(1970), a relational data model implies that once conceptualised, multidimensional biodiversity data can be mapped onto a two-dimensional ERD. Therefore, RDB can be used as a physical implementation of the conceptual framework for understanding biodiversity, which was developed in Chapter 2. Traditionally, there are three major steps of the RDB construction which include (i) logical; (ii) physical database design; and (iii) specification of a set of mapping rules. I shall fully explain each of the steps below.

### (i) Logical database design

Using the empirical biodiversity literature as a guide, I used the basic building blocks of ERD (entities and relationships) to construct the relational biodiversity model (see Appendix 3.B). This model, being the first step of RDB design, leads to a set of interrelated entities with relationships among them specified. To map biodiversity data into the ERD, I proceeded as follows.

First, a top-down modelling approach was adopted to ensure that all entities had an atomic value. Codd (1970) defines atomic value as one that cannot be decomposed any further. In the context of biodiversity, this identifies elementary concepts of biodiversity, such as "level" L or "descriptor" D, defined by using the D|L structure.

Then, descriptors and levels, organised as hierarchical trees (see Figures 2.3 and 2.5) were re-designed using an adjacency list model (Celko, 2004). The adjacency list model, related to graph theory, is a special data structure for representing hierarchies, which allows the translation of the D and L hierarchy into a two-dimensional representation (flat file).

Finally, "robust design" was achieved by using lookup tables – an array-like data structures for D and L, which initially were populated with known values but provided the possibility for an extension if more descriptors or levels were revealed by the literature search.

It is important to consider referential integrity between different concepts of biodiversity, since this is an essential constraint of any referenced relationship (Ordonez and García-García, 2008). For instance, referential integrity enforced between "biodiversity study" and "biodiversity measure", implies that if there is no study, there is no measure. Referential integrity, once satisfied, ensures the consistency of the relationships and overall integrity of biodiversity data model. All these methods and properties of the logical database design, make ERD a particularly useful approach in arranging multivariate nature of biodiversity literature. This formal treatment establishes a way to distill the biodiversity literature into quantifiable elements.

### (ii) Physical database design

After the necessary refinements and verifications of the ERD, a biodiversity model was implemented physically (i.e., coded on a computer). This was achieved through a `MySQL` server `Ver 14.14 Distrib 5.1.49`, for `Debian-Linux-GNU (x86_64)`running remotely, using `InnoDB` as a storage engine (see Appendix 3.C for an SQL dump file to produce a database).

### (iii) Mapping Rules

To ensure biodiversity data integrity and precision as well as to minimise possible sources of subjectivity during transformation of the biodiversity data from publications into RDB, I use a protocol in which types of data to be extracted and methods of extraction are specified. This protocol is comprised of mapping rules which are explicit assumptions and decisions regarding certain elements of the concept of biodiversity. Clear mapping rules important not only for overall data integrity, but they also contribute towards reduction of the possible sources of variation during meta-analysis – this is outlined in Table 3.1. The mapping rules are listed as follows:

- According to the Definition 2.5, a measure of biodiversity is a product or scalar combination of one descriptor D at one level L. This is directly implemented in the ERD;
- Three types of relation in the D|L structure were included: cross-level (D-constant, L-variable), cross-descriptor relation (D-variable, L-constant) and both cross-level and cross-descriptor combined (D-variable, L-variable);
- Biodiversity estimates reported within a given publication can either refer to biodiversity measure (e.g., species richness or species evenness) or biodiversity index which is a combination of measures (e.g., Shannon diversity index);
- To identify the number of studies within a publication in a consistent way the following decision rule was established: biodiversity measures were attributed to a single study within a publication if (i) there was a clear intention to produce, accumulate or compare measures of biodiversity within one publication; (ii) measurements were repeated either in place or in time or both, with the intention of combining them, either as replicates or in comparison; (iii) a standard protocol for the study was reported. Alternatively, if there was no intention to produce, accumulate or compare various measures within one study, or no clear protocol for this, then the measures were treated as emanating from separate studies within one publication. Additionally, to be considered as a separate study, data had to be associated with either (i)

a different year of study; (ii) a different study system; (iii) an intentionally different study design; or (iv) a different study location. Studies that were conducted at different locations were distinguished from studies with different transects, which are just replicates of the same study. For studies at different locations to be treated as replicates, their study design were required be identical, (which is difficult to achieve in practice);

- When quantitative estimates were not shown in the primary literature, they were manually calculated from the reported data or visually derived from the graphs. Multiple estimates related to the same measure or index within one study were reduced to a single number by taking their average. When it was not possible to obtain any quantitative approximation, the following rule applied: “0” was assigned if there was a measurement and the value was known to be 0 or “NULL” if there was no measurement, and the value was not known or reported.

- To consider a sampling event as a replicate, it had to fulfil the criteria for independence. The number of replicates used to produce a biodiversity estimate gives a magnitude and variability associated with that estimate. Samples, taken across a period of time to draw conclusions about, e.g., seasonal changes, were not completely independent and, therefore, had to be considered as an extended single sample (pseudo-replication). For the studies that took place over several years, the initial and the final year of the study were recorded as meta-data.

### 3.2.3 Biodiversity data mining

Meta-analysis of biodiversity literature collected in RDB was used to interrogate multivariate patterns in biodiversity estimates spread across different publications, studies, measures, and indices. Biodiversity data-mining techniques applied in the form of a structured query language in conjunction with statistical modelling allowed to formally combine multivariate biodiversity estimates so that a more informed analysis of performance and sensitivity of biodiversity indicators was possible.

As a starting point, taking a query-based approach a formal set of queries (see Appendix 3.F) was implemented on the database to explore biodiversity data. These included description of the study systems and environments (Query 3.1); geographical spread of the studies (Query 3.2), measures, and indices that were calculated from them (Query 3.3); counting the occurrence of specific measures (Query 3.4) and the range of their values (Query 3.5); distribution of descriptors across specific levels (Query 3.6) and frequency distribution of different elements of D|L matrix in the empirical literature (Query 3.7). Using recursive queries it was also possible to retrieve the full tree of descriptors and levels (Query 3.8 and 3.9). Descriptive summary statistics including median, weighted mean, and a range within which estimates was reported based on the output of the queries. Considering methodological, ecological, and conceptual aspects of study design a degree of heterogeneity of biodiversity estimates arising at different levels was rated and compared using several grouping factors.

The variability patterns across biodiversity studies and their estimates revealed by the descriptive query-based mechanism were further explored using (General Linear Modelling (GLM)). More specifically, Ordinary-Least Squares (OLS) estimator was used to investigate whether it was possible to make a comparison among biodiversity measurements based on a single response variable (biodiversity estimate) and a combination of other explanatory variables. Linear regression, applied to the inherently linear equations, means that the relationship between the variables did not have to be exactly linear; only the relation between parameters in the model was linear (see Pindyck and Rubinfeld, 1998, for a concept of inherently linear model). The rationale of applying GLM to biodiversity data, is that if there were any consistent patterns in biodiversity estimates across biodiversity studies they should be revealed from a sample of empirical biodiversity literature represented in biodivDB.

## 3.3 A worked example: of the database

Here I develop an effective worked example to illustrate the potential of using RDB as a unifying framework for the meta-analysis of biodiversity literature. For this, I follow the RDB design steps introduced earlier and describe their implementation. I start by applying a predefined set of mapping rules to the empirical biodiversity literature. This results in the fully described and classified conceptual model of biodiversity data (ERD). Then, having established the relationship cardinalities between all entities of the ERD, I physically implement it and populate it the with the empirical biodiversity data.

### 3.3.1 Logical database design

#### Biodiversity data model

The mapping rules resulted in an ERD consisting of 13 entities that are linked in a conceptually meaningful way through their participation in one or more relations to define conceptual domain (Appendix 3.B). This conceptual domain can be thought of as a formal structure of the relational view of the biodiversity knowledge that contains a set of all possible concepts of biodiversity and the relationships between them. In what follows, I describe the way the most important concepts of biodiversity are captured in the ERD.

Recalling that the ERD can be seen as a physical implementation of the conceptual framework of understanding of biodiversity, the two most central entities would be those accommodating descriptors D and levels L. Both of them participate in a relationship with another entity "measure" M, thus explicitly referring to D|L structure (see Equation 2.2). The entity "measure" (`tbl_msr`), can be best mapped as a store of meta data about the relationship between "descriptor" (`tbl_dsc`) and "level" (`tbl_lvl`). Assuming here, that both entities have a many-to-many relationship, this implies that one descriptor (or one level) can participate in many relationships to form many measures of biodiversity. In the ecological context, for example, level "species" combined with descriptors "evenness" and "richness" may form at least two distinct measures – species richness and species evenness. The organisation of D and L is hierarchical following the Figures 2.3, 2.5 (see pages 27 and 30 respectively). This allows decomposition of each measure of biodiversity M into two components: D and L (applying Definition 2.5 on page 31).

The entity "index" (`tbl_indx`) represents another important concept. It stores the information about indices of biodiversity if they were calculated and reported. The definition 2.6 (page 31) suggests that certain measures $\mathbf{M}'$ from matrix $\mathbf{M}$ assembled in some meaningful way produce an index $\mathcal{I}$. From this, it follows that concepts "measure" and "index" should be related to one another within the ERD. To do so, `tbl_indx` was linked to `tbl_msr` via one-to-many relationship (with "many" on the "measure" end). Again, thinking about this relationship in the context of biodiversity, this conveys the meaning that "index" is nothing else but a combination of measures of biodiversity, and one measure can generate more than one index.

Thus, "measure", "descriptor", "level", and "index" are the essential entities; others being meta-data structures which play a supporting role. A list of all entities along with their types and description is shown in Table 3.3 (for a full description refer to data dictionary in Appendix 3.D).

Three types of entities that are possible according to the Table 3.3 include strong, weak, and associative entities. Traditionally defined (see, e.g., Watson, 2006; Garcia-Molina et al., 2008), strong entities can exist alone and contain a "one" end of the relationship, whereas weak entities usually depend on other entities to exist. For example, entity `index` is classified as a strong entity as it does not depend on other entities to exist. However, associative entities are of primary interest here, due to the fact that they are designed to store information about the relationship between other entities.

An example of associative entity can be given here using "measure", which links the two other entities: "level" (`tbl_lvl`) and "descriptor" (`tbl_dsc`) with a relationship between them classified as many-to-many. That is to say, that a single publication on measures of biodiversity may contain one or more descriptors which may appear on one or more biological levels. Similarly, a single descriptor may appear in one or more publications and can be

**Table 3.3:** Classification, type, and description of the ERD entities in biodivDB database

| Entity | Type | Description |
|---|---|---|
| tbl_pbl | strong | It is used to store meta-data about biodiversity publication. Its existence does not depend on the existence of any other entities |
| tbl_kwd | weak | Keywords that are used in biodiversity publication are stored here. Its existence depends on the existence of "publication" |
| tbl_auth | weak | It contains information about authors of the publication. Its existence depends on existence of "publication" |
| tbl_rel | weak | This entity links other entities, and it depends on existence of "publication" |
| tbl_std | weak | Study-specific data contains here. Its existence depends on the existence of "publication" |
| tbl_dsc | strong | This entity contains all possible descriptors of biodiversity. Its existence does not depend on existence of any other entities. |
| tbl_lvl | strong | This entity contains all levels of biodiversity organised in a hierarchic structure. Its existence does not depend on existence of any other entities. |
| tbl_msr | associative | Measure contains information about relationship between entity "descriptors", entity "levels", and "index" |
| tbl_indx | strong | Index of biodiversity. Its existence does not depend on existence of any other entities. |
| tbl_indx_alias | weak | Duplicates/Synonyms of the indices. Its existence depends on existence of "index" |
| tbl_relto_index | associative | It is used to store relationship between different indices |
| tbl_relto_std | associative | It is used to store relationship between different studies |
| tbl_relto_msr | associative | It is used to store relationship between different measures |

presented on one or more levels. Both assumptions are valid here. A conventional approach to resolve many-to-many relationship within the RDB is to promote this relationship to another associative entity, e.g., "measure". Now, this new entity stores information about the relationship and removes the cardinality many-to-many from the structure. A new cardinality is one-to-many between the outer entities ("publication", "levels", and 'descriptors). This newly created associative entity "measure" points with the many end towards the associative entity.

All entities should satisfy a referential integrity rule, which implies that for every attribute of one entity there exists a corresponding attribute in another entity. In terms of the "level-measure-descriptor" relationship, this would mean that only descriptors and levels

– elements of the finite (and defined) set can be used to form measure, and no single measure can exist without prior identification of its D and L. This rule can be enforced through assigning unique identifiers, which help, in turn, to link all entities in the ERD in a comprehensive manner by means of primary key-foreign key constraint. A complete overview of all entities in the ERD, their attributes and keys is given in data dictionary in Appendix 3.D.

### Relationship cardinalities and business rules

So far, I have referred a few times to different types of relationships that may exist in biodiversity data model, but have not defined them formally. Now, I will do so by introducing the ERD terminology, which suggest that this quantity relationship between different entities of the ERD can be denoted as cardinality. As a crucial aspect of the ERD, it defines (and more impotently quantifies) the way in which concepts of biodiversity are related. Relationship cardinalities are usually described through one-to-one, one-to-many or many-to-many relationships (see Table 3.4).

**Table 3.4:** Relationship cardinalities in biodiversity model

| Notation | Cardinality | Description | Example |
|---|---|---|---|
| `1:1` | one-to-one | each entity in the relationship will have exactly one related entity | |
| `1:m` | one-to-many | an entity on one end of the relationship can have many related entities, but an entity on the other end will have a maximum of one related entity | one biodiversity study may contain one or more measures of biodiversity; one publication may contain one or more biodiversity studies |
| `m:m` | many-to-many | entities on both ends of the relationship can have many related entities on the other end | many descriptors of biodiversity combined with many levels form biodiversity measure |

Most entities in the biodiversity model have one-to-many or many-to-many relationship cardinalities. This intricate structure of the concept is formed by many measures that are composed of many descriptors at many levels. Each relationship can be read in both directions. All `m:m` relationships were converted into `1:m` with the help of associative entities. While some of the relationships are intuitive others are not always readily inferred from the biodiversity data model. Business rules which are tightly related to cardinality by placing constrains on each entity, asserting the structure and behaviour are intended to avoid ambiguities. Text descriptions of the relational cardinalities between entities – concepts of biodiversity and their graphical representations are shown in Appendix 3.E.

To summarise, a set of mapping rules applied to the empirical biodiversity data resulted in a conceptual model of biodiversity knowledge. This was achieved by mapping a set of entities (concepts of biodiversity) to form the ERD which is a formal way of thinking of biodiversity. Each concept was used to store a specific type of biodiversity data with different properties. Concepts can be either weak or strong, depending on whether they depend on other concepts to exist. Concepts that are designed to store information about the relationship between other concepts are called associative. For example, entity "publication" is strong entity and "keywords" or "author" are regarded as weak entities. Putting it more practically, it means that if there is no publication there is no author. Another example is entity "measure", being a central concept of the ERD, it stores information about the relationship between "descriptors" and "levels". For this reason, it is classified as associative. Since, "measure" depends on "descriptor" to exist – i.e., one descriptor is used to create many measures of biodiversity – it is, hence, also weak.

The biodiversity data model accommodated within the ERD is evidently a robust way of describing biodiversity. While maintaining the rigid structure and relationships between the concepts of biodiversity, it also gives sufficient flexibility to allow for biodiversity data analysis. Additionally, the ERD reflects the cross-referenced structure of biodiversity knowledge in which one publication on measures of biodiversity can be linked to another publication based on the similarity of their study systems, environments or measures. A link between different measures or indices is also possible within or between publications provided that this type of the relationship has been established. Overall, this unifying framework, not only facilitates meta-analysis of the biodiversity literature, but also allows an interlinked network of biodiversity measures across multiple levels and descriptors to be built.

### 3.3.2 Physical database design, validation, and testing

A formal modelling of biodiversity concepts and their relationships as well as structuring of the data to construct a flexible model was achieved by applying Codd's principles of data normalisation. When dealing with complicated data structures, such as biodiversity data, one of the main principles of data normalisations lies in decomposing relations with non-atomic values. Normalisation was needed because biodiversity is inherently multidimensional concept. In contrast, if the concept was simple, then it could be represented in storage by a two-dimensional array. Through taking several consecutive steps in data decomposition (i.e., normalising data up to a high normal form), redundancies within the RDB are minimised.

As a test of the ERD for the robustness prior to its implementation, I used several publications to verify and validate the conceptual model of biodiversity. Several adjustments in database structure were made following this quality control step, including the introduction of the recursive relationship in "levels" and "descriptors" (i.e., the relationship at different taxonomic levels). Having verified the biodiversity data model, the RDB logical design was used to create its physical implementation (see Appendix 3.C).

## 3.4 Biodiversity data mining: Results

Here I show the results of the descriptive query-based approach used to provide insights into the empirical biodiversity literature (Section 3.4.1) and the inferential statistics approach to explore patterns in biodiversity estimates (Section 3.4.2).

### 3.4.1 Description of the biodiversity literature

A pilot study of 30 publications produced 189 quantitative estimates of measures of biodiversity. To describe biodiversity data and its estimates, I explore the content, statistics on central tendencies and variability. A detailed list of the descriptive queries that I used to extract numeric values is available in Appendix 3.F. The main results are summarised in Table 3.5.

The database showed that 73% of the measures involve a single taxonomic level with the majority of them referring to "species" (90%). This combined with the descriptor "richness" produces the most popular measure of biodiversity in D|L structure – "species richness". It is also suggested that in most of the studies the word "biodiversity" was not formally defined, implying that its equivalence with the "species richness" was assumed. Several geographically spread study systems for which biodiversity estimates (effect size) were calculated, have been identified.

#### The variability of the effect size

The degree of variability in effect size, the measures of central tendencies, and overall spread are shown in Table 3.6. For convenience, I shall refer to the effect size as $\mathcal{E}_{im}$, where $i$ is a set of $i = 1, \ldots, k$ studies on $m_i = 1, \ldots, m$ measures of biodiversity they contain. I calculate median, weighted mean, and range, denoted respectively as $\tilde{\mathcal{E}}_{im}$, $\bar{\mathcal{E}}_{im}$, and $R(\mathcal{E}_{im})$.

Median is a location parameter that gives a robust indication of central tendency. By arranging all (189) measures of biodiversity and separating the lower half of a sample from the higher half I obtain $\tilde{\mathcal{E}}_{im}$=62.

Weighted mean $\bar{\mathcal{E}}_{im}$, which is arithmetic average of the estimate $\mathcal{E}_{im}$ describes the central location of the data. It is shifted upwards by a small number of biodiversity estimates with very large values, so that the majority of biodiversity estimates are lower than the

**Table 3.5:** Summary of the database of measures of biodiversity (1995-2009)

| | Number |
|---|---|
| Publications | 30 |
| Studies | 53 |
| Measures | 189 |
| Indices | 63 |
| Levels L (distinct) | 26 |
| Taxonomic | 138 |
| Genetic | 15 |
| Others | 36 |
| Descriptors D (distinct) | 23 |
| Richness | 70 |
| Abundance | 38 |
| Composition | 15 |
| Others | 66 |
| Study systems: | |
| macrobenthos | 7 |
| insects | 12 |
| birds | 4 |
| microorganisms | 7 |
| plants | 22 |
| soil | 2 |
| Study type: | |
| Meta | 13 |
| Others | 40 |
| Country | 24 |

**Table 3.6:** Summary of biodiversity estimates across studies included in biodivDB

| | Median | Weighted mean | Range |
|---|---|---|---|
| Biodiversity measure ($\mathcal{E}_{im}$) | 62 | 283 | 0. . . 8196 |
| Biodiversity index ($\mathcal{I}_{im}$) | 4,62 | | -0,7. . . 84,17 |
| Number of estimates ($N(\mathcal{E}_{im})$) | 28 | 119 | 3. . . 297 |

mean. The difference between $\tilde{\mathcal{E}}_{im}$=62 and $\bar{\mathcal{E}}_{im}$=283 lies within one standard deviation, suggesting that the distribution of $\mathcal{E}_{im}$ is likely to be skewed.

The difference between the highest and lowest value of biodiversity estimate containing all the values that fall between the sample $min(\mathcal{E}_{im})$ and the sample $max(\mathcal{E}_{im})$ from the empirical literature demonstrates the degree of dispersion of biodiversity estimate. I obtain $R(\mathcal{E}_{mi})$=(0; 8196), which is least robust statistics, because it is maximally sensitive to outliers. The range for an index estimate $R(\mathcal{I}_{im})$=(-0.7; 84.17) indicates a milder degree of dispersion compared to measures, but it still implies certain degree of skewness in empirical data.

Replicates, defined here as a deliberate repetition within a study, are denoted as $N(\mathcal{E}_{im})$. While the minimum number of replicates that a single study contained was 3 and the maximum number of 297, $\tilde{N}(\mathcal{E}_{im}) = 28$ suggests that most of the studies have degree of precision greater than zero.

The variability in biodiversity estimates suggested by Table 3.6 needs to be further explored. Using "species" and "richness" as two grouping factors biodiversity estimates are plotted against their number of replicates on Figures 3.1 and 3.2. If the grouping factors are omitted, it appears that the range of biodiversity estimates decreases with the increase in the number of their replicates. Since biodiversity estimates are composed of different D|L elements which jointly contribute to the variance, we are likely to be dealing with multiplicative heteroscedasticity.

### Heteroscedasticity

To address this statistically, I perform the Breusch-Pagan test (Breusch and Pagan, 1979) against heteroscedasticity.

Under $H_0$: homoscedasticity, the test statistic of the Breusch-Pagan test $\xi = NR^2$ asymptotically follows a $\chi^2$ distribution. Fitting the simple model in which $\mathcal{E}_{im}$ is explained by $N(\mathcal{E}_{im})$, $H_0$ is rejected with $p$-value=0.024 ($\xi$= 5.116, $df$= 1) and conclude that the variance of biodiversity estimate changes with the number of replicates. Alternatively, fitting the inverse model in which $N(\mathcal{E}_{im})$ is explained by $\mathcal{E}_{im}$ no evidence of heteroscedasticity was found ($p$-value=0.894, $\xi$= 0.018, $df$= 1).

Overall, there is evidence of inter-comparability problems among biodiversity estimates, when they are generated by different studies using different measures. This is confirmed by the patterns on Figure 3.1, indicating that measures that belong to species level do not suggest any variability. A large spread of biodiversity estimates results in separate point estimates scattered across x-axis suggesting the lack of consensus in the empirical biodiversity literature.

Having established great variation among biodiversity estimates I used the database to find a graphically represented attribution of sources of variability. Factors considered were (1) methodological; (2) ecological; and (3) conceptual. While methodological differences such as different study design should be kept at minimum, ecological (e.g., different study systems) and conceptual (e.g., different level and descriptors) can be desirable, because they cover different facets of biodiversity and can give indicate why biodiversity estimates are so diverse. This makes the statistical inference of the sources of heterogeneity a very useful tool in explaining the differences in biodiversity metrics.

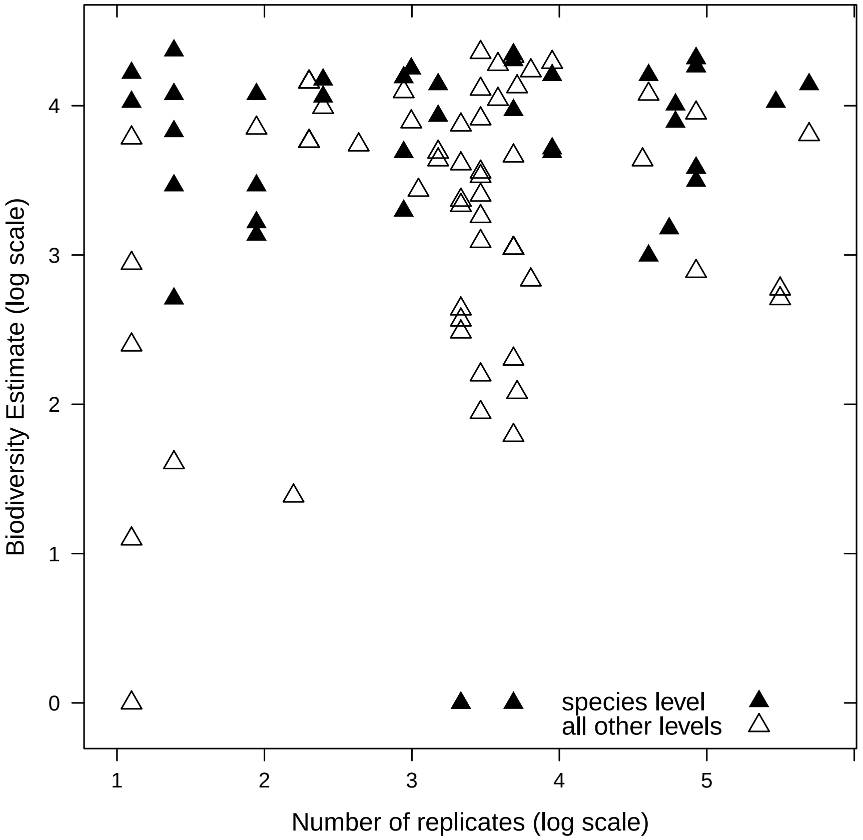


**Figure 3.1:** The number of replicates ($N(\mathcal{E}_{im})$) plotted against the biodiversity estimate ($\mathcal{E}_{im}$) on logarithmic scale and using species level as a grouping factor

### 3.4.2 Patterns in biodiversity estimates across studies

#### Specification of the models

To find an explanation for the variability in biodiversity estimates, I built a GLM statistical model, which is easier to interpret and computationally tractable, rather than using more complicated models that might fit the data better (e.g., GAMS in Gallardo et al., 2011), but at the cost of tractability. On the grounds of simplicity taking the general to specific modelling approach (see discussion by, e.g., Hoover and Perez, 1999) a full or General Unrestricted Model (GUM) specification is obtained (Equation 3.1):

$$y_i = x_i^{'}\boldsymbol{\beta} + \varepsilon_i, \tag{3.1}$$

where $y_i$ is response variable, $x_i$ is a vector of explanatory variables $(1 \quad x_{i2} \quad x_{i3} \ldots x_{iK})'$ and $\varepsilon$ is unobserved and referred to as an error term. The elements $\boldsymbol{\beta}$ are unknown model

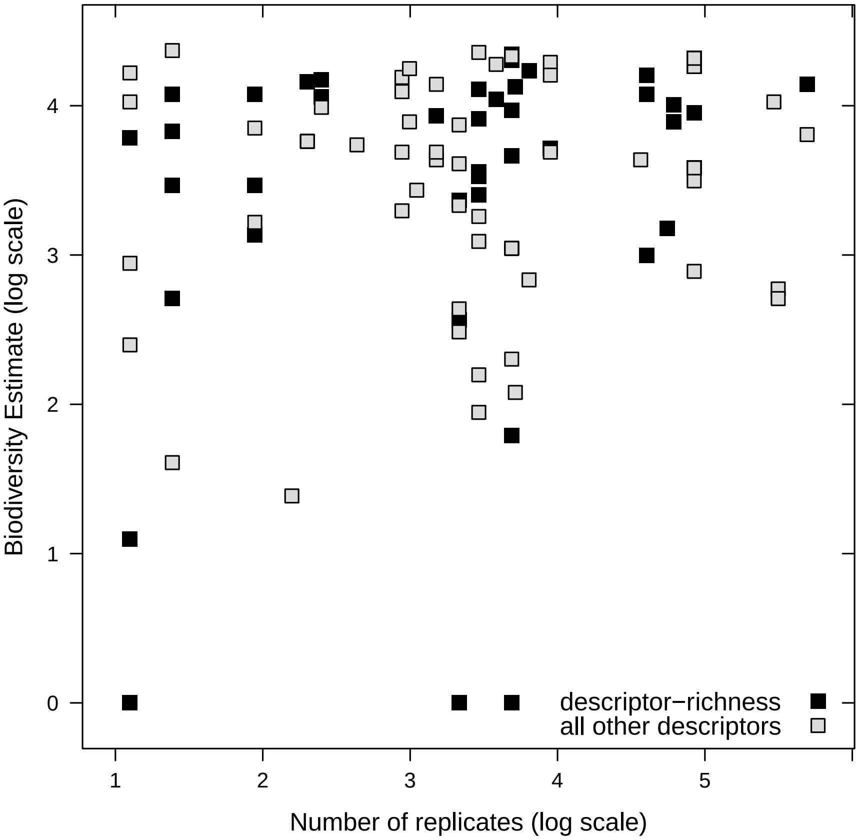


**Figure 3.2:** The number of replicates ($N(\mathcal{E}_{im})$) plotted against the biodiversity estimate ($\mathcal{E}_{im}$) on logarithmic scale and using descriptor richness as a grouping factor

parameters which can be collected in a $K$-dimensional vector $\boldsymbol{\beta} = (\beta_1 \ldots \beta_K)'$.

OLS estimator relies on Gauss-Markov assumptions (Verbeek, 2006):

$$E\{\varepsilon_i\} = 0, \qquad i = 1, \ldots N \tag{3.2}$$

$$\{\varepsilon_1, \ldots, \varepsilon_N\} \text{ and } \{x_1, \ldots, x_N\}, \qquad \text{are independent} \tag{3.3}$$

$$V\{\varepsilon_i\} = \sigma^2, \qquad i = 1, \ldots, Ni \tag{3.4}$$

$$cov\{\varepsilon_i, \varepsilon_j\} = 0, \qquad i, j = 1, \ldots, N, i \neq j \tag{3.5}$$

Assumption 3.2 states that the errors have an expectation value of zero. Assumption 3.3 imposes zero correlation between the error terms and the explanatory variables. Violation of this assumption as a result of any kind of measurement error or omitted variable bias typically can give rise to endogeneity. Homoscedasticity of the error terms is specified in 3.4 and any form of autocorrelation are excluded by assumption 3.5. These assumptions

**Table 3.7:** Summary of the variables used in GLM

| Variable | Code | Variable definition |
|---|---|---|
| $y$ | msr_val | Natural log of biodiversity estimate $\mathcal{E}_{im}$ |
| $x_1$ | 1 | constant |
| $x_2$ | pbl_year | Year of the publication |
| $x_3$ | lvl_id | Factor: levels according to Figure 2.5 |
| $x_4$ | dsc_id | Factor: levels according to Figure 2.3 |
| $x_5$ | species | Dummy: 1=species level, 0=otherwise |
| $x_6$ | richness | Dummy: 1=richness, 0=otherwise |
| $x_7$ | msr_rep | Natural log of number of replications |
| $x_8$ | std_id | Factor: levels 1...53 |
| $x_9$ | std_sys | Factor: levels – animalia, plantae, fungi |
| $x_{10}$ | std_meta | Dummy: 1=meta study, 2=original study |
| $x_{11}$ | std_dur | Duration of the study |
| $x_{12}$ | std_cntr | Factor: levels 1...24 |
| $x_{13}$ | std_rel | Dummy: 1=study is related to other studies, 0=otherwise |
| $x_{14}$ | indx_id | Factor: levels 1...63 |
| $x_{15}$ | indx_val | Natural log of numeric estimate |
| $x_{16}$ | pbl_id | Factor: levels 1...30 |

will be tested for as a part of diagnostic analysis in the following sections.

The variables that are used in GUM specification are shown in Table 3.7. I distinguish in total 16 variables related to different aspects of biodiversity literature, including meta-data and several dummy, and factor variables.

The GUM specification (Equation 3.1) can be interpreted as the conditional expectation of biodiversity estimate given the set of regressors $x_i$. This model includes all observable variables $x_i$ with $i = 1 \ldots 16$. The results of the model fit reveal inflated standard errors which can be either a sign of heteroscedasticity or model overspecification when more variables included than actually needed. This leads to loss of model efficiency. To mitigate heteroscedasticity a log transformation where appropriate $(y, x_7, x_{15})$ was taken. Inconsistencies and poor quality of the input data resulted in 89 observations with missing values. These were dropped from the model. As a result, the number of degrees of freedom decreased (*df*=32).

None of the explanatory variables were found to be individually significant in the GUM, and, hence their estimates are not reported here. There are a number of factors that can cause this, with the most relevant being: (1) possible multicollinearity between linear combinations of the (factor) explanatory variables; and (2) poor quality of the empirical biodiversity data (inconsistencies and missing data). Both factors are inherently linked as if variables are blended they are likely to be correlated.

While nothing can be done to improve the quality of the input data, the problem of the multicollinearity can be addressed in statistical terms. Dropping individual variables from the model would help to mitigate the collinearity. However, this would also result in inability to interpret the effect of any single coefficient in the model without knowing the other variables. Statistical theory suggests (Verbeek, 2006) that multicollinearity has typically little impact on the accuracy of model prediction, although it inflates the coefficients. However, the "total impact" of all explanatory variables is accurately identified by the GUM.

Even though at this stage, significant patterns in biodiversity estimates were not statistically attributable, the "overall model fit" can still be used. To infer the usefulness of the GUM, I shall look at $R^2$, which measures the proportion of sample variation in $y$ that is explained by $x$. Adjusted $R^2$= 0.374 multiple $R^2$ = 0.798 testing the difference between these two indicators using nested $F$-test ($F$-statistic= 1.884, $df$= 67 and 32) I find a significant increase ($p$-value= 0.026) in model's $R^2$.

To check that the Gauss-Markov assumptions (3.2-3.5) are not violated, a diagnostic analysis is performed. Figure 3.3 shows two diagnostic plots for the GUM specification: a plot of model residuals against fitted values and a normal Q-Q plot with theoretical quantiles shown against standardised residuals.

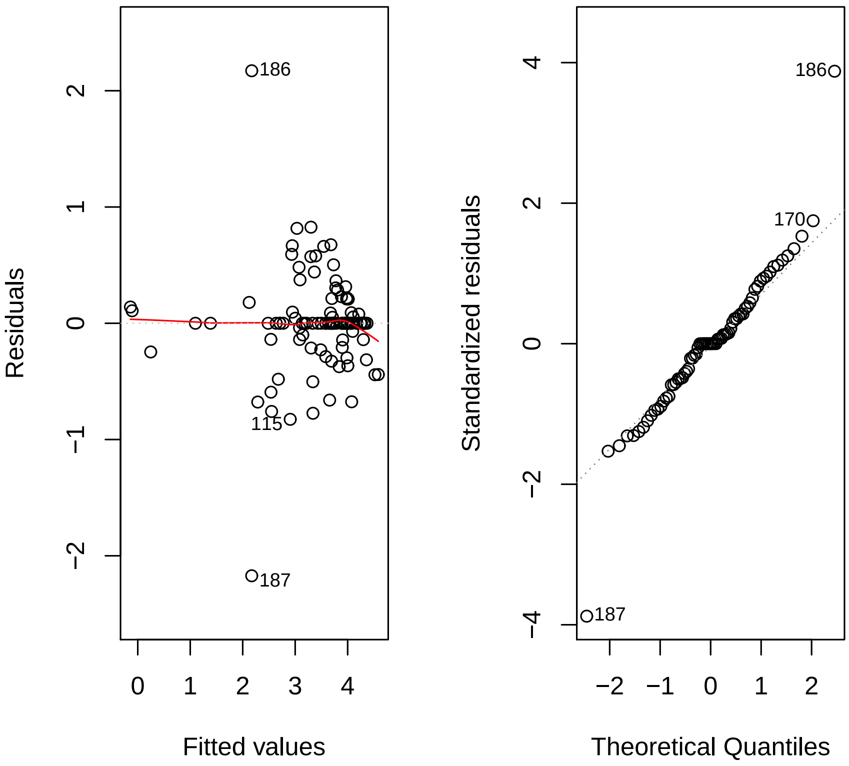


**Figure 3.3:** Diagnostic plots – General Unrestricted Model

A scatterplot of residuals against the fitted values indicates non-constant variance (especially around larger values) and potential outliers (e.g., points 186 and 187), hence an assumption on homoscedasticity of the error term (Equation 3.4) is likely to be violated. Comparison of theoretical quantiles against standardised residuals on the Q-Q plot gives visual assessment of the normality of the residuals. Since the data approximately follows a straight line, the assumption of normality is assumed valid.

A more parsimonious model was also generated, by dropping some of the variables from the GUM, a General Restricted Model (GRM) was specified as (Equation 3.6):

$$y = x_1 + x_3 + x_4 + \varepsilon \tag{3.6}$$

It is now assumed that if there are any patterns in biodiversity estimates as a result of their variation, it should be possible to determine those patterns based on descriptors and levels that generate particular biodiversity metrics. Therefore, only variables directly related to levels ($x_3$) and descriptors ($x_4$) are included and controlled for in the model. This reduced model is nested within the GUM and, as expected, reducing the number of variables resulted in some significant results.

According with the GUM, the results of the GRM fit do not show any significant estimates with the exception of those two individual levels of the factor variables $x_3$ and $x_4$ which correspond to biodiversity levels and descriptors. The first two significant levels are trophic and functional levels with the estimates 4.196 ($se$=1.108, $p$-value<0.001) and 3.843 ($se$=0.751, $p$-value<0.001) respectively.

The other two are descriptors related to function and composition of biodiversity metrics with the coefficients: -4.148 ($se$=1.097, $p$-value<0) and -3.807 ($se$=0.705, $p$-value<0.001) respectively. Further, comparing multiple $R^2$= 0.603 with adjusted $R^2$= 0.374 using nested $F$-test ($F$-statistic= 5.528, $df$= 22 and 80) a highly significant $p$-value<0.001 is obtained.

Diagnostic plots for the GRM are shown on Figure 3.4. Compared to the GUM model (Figure 3.3), non-constant variance at the right end of the curve (around large values) in the GRM has increased, which suggests considerable departure from the underlying assumptions (3.2-3.5). In the Q-Q plot, residuals with identical variance on the vertical axis plotted against theoretical quantiles on the horizontal axis, show the emergence of the non-linear patterns. Although the mean of the data is still approximately zero (no offset from the reference line), the left tail does not fit a normal model, suggesting non-zero skewness.

To summarise, the magnitude of variation (heteroscedasticity) among biodiversity estimates has been quantified, although not statistically attributed, most likely failing because of small sample size in this pilot study. The diagnostic methods based primarily on the model residuals indicate some departure from the model assumptions. By restricting the

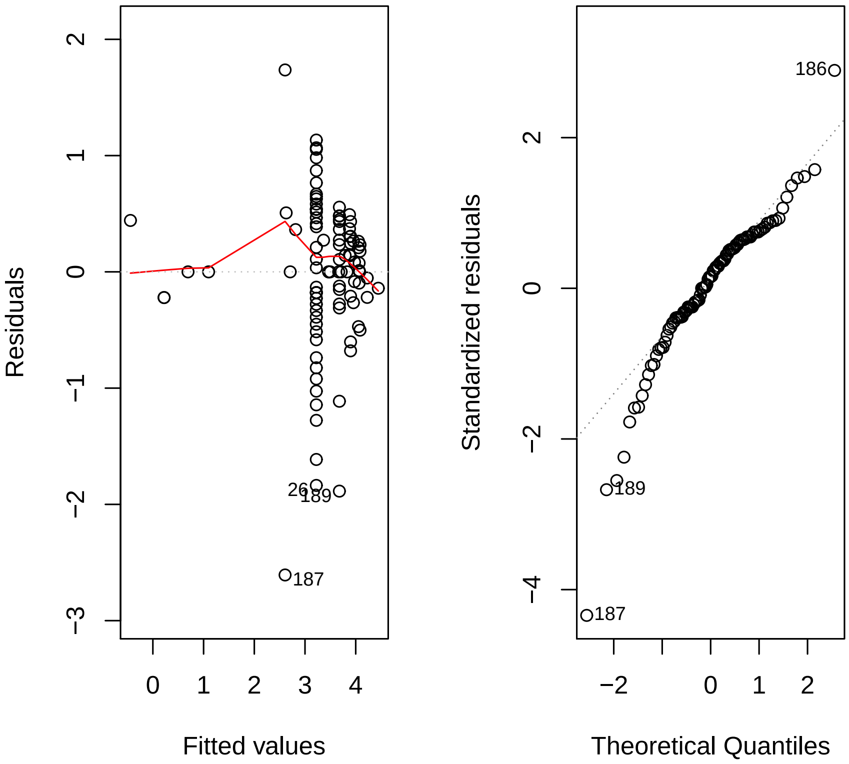


**Figure 3.4:** Diagnostic plots – General Restricted Model

number of explanatory variables to the minimum set representing D|L structure in the GRM model, some of the variables were found to be highly significant indicating a need for further investigation. Both model specifications leave a large unexplained variation, while indicating the presence of the clear patterns of variation in biodiversity estimates arising either from the diversity of biodiversity or methods used to measure it. Confirmed by the significant increase in coefficients of determination, the overall model fit in both the GUM and the GRM was statistically acceptable.

The pilot study has demonstrated how the low quality and paucity of the empirical data, combined with a proliferation of biodiversity metrics scattered across different study systems, imposes a strong limitation on cross-comparison of biodiversity estimates. This finding provides new insights into the different metrics of biodiversity and point to their limitations in defining the patterns of variation. Therefore, this meta-analysis not only reinforces the point on the multivariate nature of biodiversity measures, made earlier, but also suggests that despite the abundance of measures, they appear to be highly redundant on some levels while scarce on others. These problems presently preclude the use of empirically-based methods in the search for fundamental measures of biodiversity, suggesting the necessity for a modelling approach.

## 3.5 Discussion and Conclusions

As a response to biodiversity loss, the number of empirical studies attempting to quantify biodiversity has increased considerably (e.g., Zamora et al., 2007; Tian et al., 2007; Joshi et al., 2008; Anderson, 2008; Moreno et al., 2008; Sharma and Rawat, 2009). Impeded by difficulties over the precise definitions of biodiversity, these efforts have led to a proliferation of measures of biodiversity and their estimates. This trend has been intensified by inconsistencies of individual study designs among study systems, producing observer-dependent multivariate patterns of variability in biodiversity estimates.

Using field data to explore the patterns of variability in biodiversity estimates, multiple biodiversity measures were extracted and distilled from the literature. This followed two steps: (i) employing a combination of relational data modelling techniques as a unifying framework to organise biodiversity knowledge; and (ii) performing a statistical generalisation of multiple measures of biodiversity to test for the differences in their estimates via a formal framework of meta-analysis.

According to Arnqvist and Wooster (1995) meta-analysis is most useful under the following conditions: when there is empirical work available, results are variable across the studies with weak magnitude of the individual effect (biodiversity estimate) and limited sample size (replicates). All these requirements perfectly match the present state of biodiversity knowledge, demonstrating the appropriateness of the technique. In fact, the similarity across different studies is not even desirable, because it would imply that only one facet of biodiversity was being described. Here, in a departure from earlier mostly qualitative statements regarding the variability in biodiversity estimates, I have provided a quantitative explanation for this variability.

The meta-analysis used a dataset of 50 biodiversity studies obtained through the literature review as a pilot study. This dataset illustrates the usability of the RDB discipline in providing a suitable infrastructure for multivariate biodiversity meta-data. Once populated, the RDB has yielded in total 189 measures of biodiversity, 43 of which were distinct. Exploratory statistics have shown that the studies were conducted on different study systems – ecological communities, with approximately 40% attributed to plant communities, 23% insect, and 13% macrobenthos communities.

The presence of statistically significant variation in biodiversity estimates have confirmed the overall intuition of Englund et al. (1999) and Osenberg et al. (1999). The patterns of variation, explicitly addressed in Figure 3.1 and 3.2, suggested a decline in the range of biodiversity estimates with an increase in the sample size. The range of the values for biodiversity estimates was widely dispersed. Interestingly, the heteroscedasticity patterns revealed by the Breusch-Pagan test were lost at the species level on Figure 3.1. This implies, that when limited to a single level, the variability in estimates is likely to be caused by the diversity of measures used; thus reinforcing the point that multivariate biodiversity estimates give rise to variability. Other observations indicate the upward shift of the mean value of biodiversity estimates, caused by a small number of estimates with very large values – possibly outliers. The majority of estimates have been found below the mean, leading to the skewness.

An attempt to go beyond species in attributing various sources of the variability in biodiversity estimates was not successful. Based upon unrestricted GLM specification none of the individual variables were found to be significant, precluding any statistically justified attribution of sources of variability. However, the overall model-fit appeared satisfactory as it has been suggested by a significant increase in model's $R^2$ with $p$-value$= 0.026$ for a nested $F$-test. Study artifacts (e.g., sample size) may also contribute to the proportion of unexplained variability. However, a deeper problem may be poor quality of the test data, implied by the presence of the non-constant variance in the value for biodiversity estimates (Figure 3.3). This is shown more explicitly around large values, where the small sample size of this test study further increases the spread of biodiversity estimates. As a result, the quality of the data leads to a measurement error and non-zero correlation between the error term and explanatory variables, altogether producing endogeneity.

A more general limitation according to Osenberg et al. (1999) is that "even the most thorough and careful meta-analysis will contain bias". This typically includes the "file drawer problem" and "study selection bias" (Arnqvist and Wooster, 1995). While the file drawer problem, in which only significant results are published, is hard to control, study selection bias due to the dominance of some studies can be partially mitigated. This is achieved at the study-review stage when inclusion is determined by blind selection, following rules concerning study methods, rather than their results. Another possible bias can occur when extracting data to populate the relational database. This follows because it was not always possible to distinguish between different measures of biodiversity to match D|L structure. Despite all these shortcomings and the limited sample of all measures, some useful conclusions still can be made. The most popular D|L combination involved single taxonomic level with 90% on the species level which confirms a firmly established tradition in ecology of using species as a surrogate for biodiversity (Caro and O'Doherty, 1999).

For the first time, I have shown how knowledge engineering can be used to build a biodiversity infrastructure that adds understanding to the concept of biodiversity which is otherwise unstructured and disparate. Because biodiversity knowledge is unstructured, the process is not simply one of transfer into an appropriate computer representation, it takes the form of a model construction (Studer et al., 1998). This is a problem-driven process to structure and formalise biodiversity literature. The formal model has direct implications to biodiversity, since it allows one to minimise redundancy in measures of biodiversity through appropriately selected database attributes (e.g., constraints, multiplicity of the relationships, and referential integrity between entities-concepts of biodiversity).

While encapsulating the underlying structure of biodiversity, this approach has also quantified the problem of the intercomparability of biodiversity estimates that are generated by different studies using different measures and indices. It was not possible to attribute the observed variability in biodiversity estimates, partially because of inadequate sample size, but also due to the inconsistencies among published study study methods. The context in which these difficulties arise is that diversity, as a quantification of difference, is multidimensional, but of unknown dimensionality. Even species themselves differ in an uncounted diversity of ways, making for a very large and unknown dimensionality to total biodiversity. Existing indices provide transect projections and cross-sectional views to sample this multi-dimensional space. The most efficient sampling would be achieved by a set of orthogonal estimators, in particular those projecting along the major axes of variation: lines of greatest variance in diversity space. By deconstructing existing indices into their L and D components, In the previous Chapter I showed that it is possible to identify the available projections as "filled elements" in the permutation matrix of all possible level and descriptor pairs. Statistical ordination can then identify the desired orthogonal set of major axes, given a suitable data source.

Ideally, this data would be taken from empirical studies of real communities, but here I showed that the literature contains disappointingly little opportunity to make comparison among empirical biodiversity estimates, certainly not enough to perform an analysis with real data that is not affected by environmental co-variation. Lack of standardisation in methods and reporting of field studies account for some of this, but the wide range of study systems, their size, and location, as well as the variety of purposes for empirical studies seems to preclude success for the kind of meta-analysis, commonly found in medical research. This demonstrates the need for an analysis in a more controlled environment with more uniform studies, spanning over the range of indices for a large sample size. That is the subject of the next Chapter.

To summarise, relational database infrastructure has been used as a support for meta-analysis of the biodiversity literature allowing aggregation and comparison of biodiversity estimates from different publications. Current biodiversity research generates a multitude of measures, many overlapping as revealed by the resulting database. Overall, since these biodiversity estimates are point estimates scattered across different study systems it was difficult to compare them. It was not possible to elucidate different sources of variability in biodiversity estimates.

To reveal any statistically significant patterns a sufficiently large set of studies following a consistent protocol, with matched sampling effort, spanning a wide variety of communities is required. Empirical biodiversity data failed to meet most of these requirements with no explicit patterns found except for one: instances L and D were disproportionally represented in the empirical biodiversity literature. While a small group of permutations D|L found to be dominant, others were completely missing from the literature, precluding empirically based analysis. Evidently, using real ecological communities to attribute patterns of variation and to reduce the number of indices, is not yet possible and cannot contribute to the overall aim of my research. The lack of uniformity across biodiversity studies with different estimates and across different study systems strongly suggests the necessity to use artificial communities.

## Summary

1. Since biodiversity knowledge is unstructured and disparate, the organisational discipline of RDB, combined with data mining techniques was chosen for the meta-analysis of the biodiversity literature to collect as much empirical evidence on measures as possible in a structured and consistent manner;

2. Using field data to explore the patterns of variability in biodiversity estimates, multiple biodiversity measures were extracted and distilled from the literature. This followed two steps: (i) employing a combination of relational data modelling techniques as a unifying framework to organise biodiversity knowledge; and (ii) performing a statistical generalisation of multiple measures of biodiversity to test for the differences in their estimates via a formal framework of meta-analysis.

3. Once populated, the RDB contained 189 measures of biodiversity, 43 of which were distinct. Exploratory statistics have shown that the studies were conducted on different study systems – ecological communities, with approximately 40% attributed to plant communities, 23% insect, and 13% macrobenthos communities;

4. The majority of measures (73%) involves a single taxonomic level, 90% of which were referring to “species”. This combined with the descriptor “richness” produced the most popular measure of biodiversity in D|L structure – “species richness”;

5. The numerical values among biodiversity estimates were distributed with very large variance, but no statistical patterns were found at the study-level; and

6. The intercomparability of biodiversity estimates generated by different studies created a problem which may be reduced by coordination among empiricists to standardise methods and reporting techniques.

# Appendices

## Appendix 3.A Documentation of studies included in database of measures

**Index:** Species Diversity
**Reference:** Joshi et al. 2008
**Formula:** $H'(S) = \sum_{i=1} p_i \log p_i$
**Variables:** $p_i = n_i/N$; $n_i$ – number of individuals of species i; $N$ – size of the whole community; $S$ – total number of species
**Calculation:** avg(species abundance)=(Abundance Site 1 +Abundance Site 2 +Abundance Site 3)/3= (3980+3114+1780)/3=2958;
$H' = -2958/8875 \times \log(2958/8875) = 0.1586$
**Value:** $\approx 0,158$

**Index:** Index of evenness/equitability
**Reference:** Joshi et al. 2008
**Formula:** $H/H_{max}$
**Variables:** $H$ – realized value of diversity – Shannon, $H_{max} = \ln S$ – max possible value of diversity
**Calculation:** (Site 1+ Site 2 +Site 3)/3=(5,420+4,832+3,610)/3=4,62
**Value:** $\approx 4,62$

**Index:** Seasonal Diversity
**Reference:** Joshi et al. 2008
**Formula:** $H'(P) = \sum_{j=1} q_j log q_j$
**Variables:** $q_j = n_j/N$; $n_j$ – number of individuals present in season $j$; $N$ – size of whole community; $p$ – number of seasons
**Calculation:** avg($q_j$)=(106/3980+1423/3980+73/3114+1223/3114+40/1781+541/1781)/6=1,385
$H' = -1.385 \times \log(1,385) = -0,195$
**Value:** $\approx -0,195$

**Index:** Richness index (SR)
**Reference:** Tian et al. 2007
**Formula:** SR=S
**Variables:** S – total number of species
**Calculation:** $R = (R_{community}1$+$R_{community}2$+$R_{community}3$+...+$R_{community}n)$/6= (10+13+10+25+ ...+3)/6=90/9=10
**Value:** $\approx 10$

**Index:** Shannon-Wiener diversity index
**Reference:** Tian et al. 2007
**Formula:** $H' = \sum_{i=1}^{s} p_i \ln p_i$
**Variables:** $p_i = n_i/N$ – relative coverage of the i-th species; $n_i$ – number of individuals of species i; $N$ – size of the whole community; $S$ – the total number of species in each community
**Calculation:** $(H_1 + H_2 + H_3 + H_4 + \dots H_9)/9 = (0,782 + 1,144 + 0,803 + 1,018 + 1,697 + 1,354 + 0,049 + 0,479 + 0,567)/9 = 0,877$
**Value:** $\approx 0,877$

**Index:** Pielou evenness index
**Reference:** Tian et al. 2007
**Formula:** $J = (-\sum_{i=1}^{s} P_i ln P_i)/\ln s$

**Variables:**
**Calculation:** $avg(J) = (J_1 + J_2 + J_3 + J_4 + \dots J_9)/9 = (0,34 + 0,447 + 0,349 + 0,316 + 0,537 + 0,564 + 0 + 0,435 + 0,515)/9 = 0,3892$
**Value:** $\approx 0,3892$

**Index:** Jaccard dissimilarity index
**Reference:** Anderson 2008
**Formula:** Magurran 2004; $C_j = j/(a + b - j)$
**Variables:** j – number of species common to both sites; a – number of species in site a; b – number of species in site b
**Calculation:** $J' = 0,31$
**Value:** 0,31

**Index:** Shannon-Wiener
**Reference:** Sharma and Rawat 2009
**Formula:** $\bar{H} = \sum_{i=1}^{s}(n_i/N)\log_2(n_i/N)$
**Variables:** $n_i$ – the total number of individuals of a species; N – total number of all species
**Calculation:** $(\bar{H}_{july} + \bar{H}_{aug} + \bar{H}_{sep} + \bar{H}_{oct} + \bar{H}_{nov} + \dots + \bar{H}_{june})/12 = 4,26$
**Value:** $\approx 4,26$

**Index:** Coefficient Similarity (Jaccard)
**Reference:** Sharma and Rawat 2009
**Formula:** S=C/(A+B-C)
**Variables:** C – number of common species; A – total number of species in community A; B – total number of species in community B
**Calculation:** $(S_1S_2 + S_1S_3 + S_1S_4 + S_2S_3 + S_2S_4 + S_3S_4)/6 = 84,16$
**Value:** $\approx 84,16$

**Index:** Jaccard similarity
**Reference:** Moreno et al. 2008
**Formula:** Magurran 2004; $C_j = j/(a + b - j)$
**Variables:** j – number of species common to both sites; a – number of species in site a; b – number of species in site b
**Calculation:**
**Value:** 0,21

**Index:** Sorensen similarity
**Reference:** Moreno et al. 2008
**Formula:** Magurran 2004; $C_N = 2jN(aN + bN)$
**Variables:** aN – the number of individuals in site A; bN – the number of individuals in site B; jN – the sum of the lower of the two abundances of species which occur in the two sites.
**Calculation:**
**Value:** 0,348

**Index:** Chao-Jaccard
**Reference:** Moreno et al. 2008
**Formula:** Chao et al. 2005
$J_{abd} = UV/(U + V - UV)$
**Variables:** U – the total relative abundance of individuals belonging to the shared species in assembly 1; $U = p_1 + p_2 + \dots + p_n$;
V – the total relative abundance of individuals belonging to the shared species in assembly 2; $V = \pi_1 + \pi_2 + \dots + \pi_n$;
**Calculation:**
**Value:** 0,430

**Index:** Chao-Sorensen
**Reference:** Moreno et al. 2008

**Formula:** Chao et al. 2005; $L_{abd} = 2UV/(U+V)$
**Variables:** U – the total relative abundance of individuals belonging to the shared species in assembly 1; $U = p_1 + p_2 + \ldots + p_n$
V – the total relative abundance of individuals belonging to the shared species in assembly 2; $V = \pi_1 + \pi_2 + \ldots + \pi_n$
**Calculation:**
**Value:** 0,601

**Index:** Shannon (H)
**Reference:** Moreno et al. 2008
**Formula:** Magurran 2004; $H^{'} = -\sum p_i \ln p_i$
**Variables:**
**Calculation:** (Diversity overall managed direct search 1 + Diversity overall managed direct search 2)/2=(1,4+0,6)/2=1
**Value:** $\approx 1$

**Index:** Pielou Evenness (J)
temporal beta diversity
**Reference:** Moreno et al. 2008
**Formula:** Magurran 2004; $J = H/\log_2(S)$
**Variables:**
**Calculation:** (0,43+0,22)/2=0,325
**Value:** $\approx 0,325$

**Index:** Temporal species turnover
**Reference:** Zamora et al. 2007
**Formula:** Proportion of exclusive species to total species richness between consecutive sampling periods. Applied the complementarity index of Colwell and Coddington 1994 for each sampling site and calculated the average.
$C_{jk} = U_{jk}/S_{jk}$; $S_{jk} = S_j + S_k - V_{jk}$; $U_{jk} = S_j + S_k - 2V_{jk}$
or for computation form presence-abscence matrix: $C_{jk} = \sum_{i=1}^{S_{jk}}(X_{ij} - X_{ik})/\sum_{i=1}^{S_{jk}} max(X_{ij}, X_{ik})$
**Variables:** $S_j$ – richness in site j; $S_k$ – richness in site k; $V_{jk}$ – number of species in common between j and k
$X_{ij}, X_{ik}$ – the presence-absence (1,0) for species in list j and list k
**Calculation:** avg(beta diversity) = (avg (grassland)+ avg (shrubland)+avg(woodland))/6= (80,96+76,1+72,3+84,96+76,07+69,27)/6=76,54
**Value:** $\approx 76,54$

**Index:** Species richness (alpha diversity)
**Reference:** Aubert et al. 2003
**Formula:** SR=number of species
**Variables:**
**Calculation:** (9+16+14+17+18)/5=15
**Value:** $\approx 15$

**Index:** Shannon diversity index ($H'$)
**Reference:** Aubert et al. 2003
**Formula:**$H' = \sum_{i=1}^{s} p_i \log_2 p_i$
**Variables:** $p_i$ – relative frequency of species
**Calculation:** (1,25+2,5+2,25+3+3,5)/5=2,5
**Value:** $\approx 2,5$

**Index:** Evenness index ($J'$)
**Reference:** Aubert et al. 2003
**Formula:** $J' = H'/H'_{max}$
**Variables:**
**Calculation:** avg(0,4+0,6+0,55+0,7+0.825)/5=0,615

**Value:** $\approx 0,615$

**Index:** Factorial diversity (FD)
**Reference:** Aubert et al. 2003
**Formula:** $FD = \sum_{j=1}^{t} P_{j/i}[C_k(j) - L_k^{(c)}(i)]^2$
**Variables:** $p_{j/i}$ – the conditional relative frequency of sample i for species j; $L_k^{(c)}$ – the ordination of samples on gradient by averaging $C_k(j)$ – the ordination of species on gradient be weighted averaging.
**Calculation:** (max+min)/2=(0,05+0,46)/2=0,265
**Value:** $\approx 0,265$

**Index:** Jaccard index (pairwise similarity)
**Reference:** Aubert et al. 2003
**Formula:** C/(A+B-C)
**Variables:** C – the number of species shared; A,B – the total number of species occurring in record A and B
**Calculation:** (0,37+0,43+0,41+0,41+0,37)/5=0,39
**Value:** $\approx 0,39$

**Index:** Within record heterogeneity (WRH)
**Reference:** Aubert et al. 2003
**Formula:** Pairwise similarity among the four records of a plot. The mean similarities were determined using Jaccard values
**Variables:**
**Calculation:** (0,43+0,44+0,51+0,5+0,5)/5=0,47
**Value:** $\approx 0,47$

**Index:** Simpson index (D)
**Reference:** Campos and Fernando Isaza 2009
**Formula:** $r^2 = \sum_{n=1}^{s}(p_n)^2$; $D = r^2$
**Variables:** p – probability distribution representing the relative abundance; $p_n = N_n/N$; $N_n$ – number of organisms in each species; N – total number of organisms $N = \sum_{n=1}^{s} N_n$
**Calculation:** for larch middle-aged forest
**Value:** $\approx 0,46$

**Index:** Shannon index
**Reference:** Campos and Fernando Isaza 2009
**Formula:** $H(p) = H(p_1, p_2, \ldots, p_s) = -\sum_{n=1}^{s} p_n \ln p_n$
**Variables:**
**Calculation:**
**Value:** 1,6

**Index:** Geometrical index $B(S_m r_m)$
**Reference:** Campos and Fernando Isaza 2009
**Formula:** $B_k(s,r) = V_s(r)/V_s + k(r) = \alpha_k(s)/r^k = \alpha_k(s)\beta_k(r)$; $r \neq 0$
**Variables:** k – a non-negative integer; $\beta(r) = 1/r^k = 1/D^{k/2}$;
$\alpha_k(s) = \Gamma((s+k+r)/2)/\pi^{k/2}\Gamma((s+r)/r)$;
s – species richness (based on Simpson)
**Calculation:**
**Value:** 1,94

**Index:** Simpson diversity
**Reference:** Wilsey et al. 2005
**Formula:**$D = 1/\sum p_i^2$
**Variables:** $p_i$ – the proportional biomass of species i
**Calculation:** avg across 6 sites= 47,9+51,3+63,3+49,0+51,4+69,1)/6=55,3

**Value:** $\approx 55,3$

**Index:** Species richness (SR)
**Reference:** Wilsey et al. 2005
**Formula:** SR=S
**Variables:** S – total number of species
**Calculation:**(36,8+37,1+22,4+37,4+38,8+31,2)/6=33,95
**Value:** $\approx 33,95$

**Index:** Simpson Evenness (E)
**Reference:** Wilsey et al. 2005
**Formula:** D/S=Simpson diversity/Species richness
**Variables:**
**Calculation:** (41,7+42,1+63+41,4+41,3+63,2)/6=48,78
**Value:** $\approx 48,78$

**Index:** Berger-Parker dominance (BP)
**Reference:** Wilsey et al. 2005
**Formula:** Berger and Parker 1970; $D = N_{max}/N$
**Variables:** $N_{max}$ – number of individuals in the most abundant species; N – total number of individuals in sample
**Calculation:** (48,2+51,2+61,2+50,9+49,2+71,0)/6=55,28
**Value:** $\approx 55,28$

**Index:** Rarity (proportion of rare species)
**Reference:** Wilsey et al. 2005
**Formula:** Camargo 1992
Proportion of species whose relative biomass was less than 1/S
**Variables:**
**Calculation:**
**Value:** $\approx 54,06$

**Index:** Nei's diversity
measure of the average gene diversity per locus ($H_s$)
**Reference:** Lambertini et al. 2008
**Formula:** Nei and Li 1979; Kosman 2003
$H_S = 1/k \sum_{s=1}^{k} H_{S_s} = 1/k \sum_{s=1}^{k} [1 - q_s^2 - (1 - q_s)^2]$
**Variables:** $k$ – the total number of loci; $H_{S_s} = 1 - q_s^2 - (1 - q_s)^2$; $q_s$ – the frequency of one of the two alleles of the sth diallelic locus
**Calculation:**
**Value:** 0,175

**Index:** Shannon information index (I)
**Reference:** Lambertini et al. 2008
**Formula:**$H' = \sum_{i=1}^{s} p_i \log_2 p_i$
**Variables:** $p_i$ – relative frequency of species
**Calculation:** Among populations (po plain and Razim)
**Value:** $\approx 0,283$

**Index:** Number of polymorphic fragments
**Reference:** Lambertini et al. 2008
**Formula:**
**Variables:**
**Calculation:**
**Value:** $\approx 99$

**Index:** Species Dominance (Berger-Parker Index)
**Reference:** Oxbrough et al. 2007
**Formula:** $d = N_{max}/N$
**Variables:** $N_{max}$ – the number of individuals in the most abundant species; N – the total number of individuals
**Calculation:** Avg standard across grasslands = (0,22+0,34+0,23)/3=0,636
**Value:** $\approx 0,636$

**Index:** Species Richness (SR)
**Reference:** Oxbrough et al. 2007
**Formula:** SR=S
**Variables:** S – total number of species
**Calculation:** (16,9+23+26,8)/3=48,83
**Value:** $\approx 48,83$

**Index:** Shannon-Wiener (for $\alpha$-diversity
**Reference:** Jiang et al. 2007
**Formula:** $H = -\sum_{i=1}^{n}(p_i/np_i)$
**Variables:** $p_i$ – the relative importance of each species; n – number of species in the particular quadrat
**Calculation:** (min+max)/2
**Value:** $\approx 1,66$

**Index:** Sorensen index for $\beta$-diversity
**Reference:** Jiang et al. 2007
**Formula:** Sorensen 1948
$R = 2C/A + B$
**Variables:** C – the number of species shared in both belts; A – species occurring only in belt A; B – species occur only in belt B
**Calculation:** mid values (0,779+0,533+0,697+0,749)/4=0,69
**Value:** $\approx 0,69$

**Index:** Simpson Diversity (clonal diversity)
**Reference:** Chen et al. 2007
**Formula:** $D = 1 - \sum[n_i(n_i - 1)/N(N - 1)]$
**Variables:** $n_i$ – the number of ramets of the ith multilocus genotype; N – the number of samples collected for that population
**Calculation:**
**Value:** 0,95

**Index:** Species Richness (SR)
**Reference:** Ohsawa 2004
**Formula:** SR=S
**Variables:** S – total number of species
**Calculation:** avg (9,7+10,8+16,7+15,8+11,7)/5=12,94
**Value:** $\approx 12,94$

**Index:** Shannon-Wiener H'
**Reference:** Ohsawa 2004
**Formula:** $H' = \sum_{i=1}^{s} p_i \ln p_i$
**Variables:** $p_i = n_i/N$ – relative coverage of the i-th species; $n_i$ – number of individuals of species i; $N$ – size of the whole community; $S$ – the total number of species in each community
**Calculation:** avg over forest (3,0+3,1+3,4+3,5+3,1)/5=3,22
**Value:** $\approx 3,22$

**Index:** Species Richness (SR)
**Reference:** Ferrero et al. 2008

**Formula:** SR=S
**Variables:** S – total number of species
**Calculation:** (min+max)/2=(12+29)/2=20,5
**Value:** $\approx 20,5$

**Index:** Species Richness ES(51) = Rarefaction
**Reference:** Ferrero et al. 2008
**Formula:** Sanders 1968;
$ES = \sum_{i=1}^{S}[1-(\binom{N-N_i}{n})/\binom{N}{n}]$
**Variables:** $N_i$ – number of individuals in species i; $N$ – total number of individuals, $S$ – total number of species; $n$- number of individuals chosen for standardisation $n \leq N$, $\binom{N}{n}$ – number of combinations of n individuals that can be chosen from a set of N individuals = $N!/n!(N-n)!$
**Calculation:** (5+18)/2=12,5
**Value:** $\approx 12,5$

**Index:** Shannon-Wiener H'
**Reference:** Ferrero et al. 2008
**Formula:** Shannon and Weaver 1949; $H' = \sum_{i=1}^{s} p_i \ln p_i$
**Variables:** $p_i = n_i/N$ – relative coverage of the i-th species; $n_i$ – number of individuals of species i; $N$ – size of the whole community; $S$ – the total number of species in each community
**Calculation:** (1,1+2,8)/2=1,95
**Value:** $\approx 1,95$

**Index:** Evenness $J^{'}$
**Reference:** Ferrero et al. 2008
**Formula:** Pielou 1969; $J = (-\sum_{i=1}^{s} P_i lnP_i)/\ln s$
**Variables:**
**Calculation:** (0,3+0,9)/2=0,6
**Value:** $\approx 0,6$

**Index:** Equitability $V^{'}$
**Reference:** Ferrero et al. 2008
**Formula:** Platt and Lamberhead 1985; $V^{'} = H/H_{max}$
**Variables:** $H_{max} = \ln S$; $H$ – Shannon diversity index
**Calculation:** (-4+2,6)/2=-0,7
**Value:** $\approx -0,7$

**Index:** Shannon-Wiener
**Reference:** Gamito and Furtado 2009
**Formula:** Shannon and Weaver 1949
$H'_{FD} = -\sum_{i=1}^{n}(p_i \log_2 p_i)$
**Variables:** $p_i = fg_i/\sum_i^n fg_i$ – the relative abundance of the ith feeding group
n – the total number of feeding groups
**Calculation:** estimate of feeding diversity
(min +max)/2=(0,6+4,8)/2=2,7
**Value:** $\approx 2,7$

**Index:** Evenness index
**Reference:** Gamito and Furtado 2009
**Formula:** Pielou 1969
$j_{FD} = H'_{FD}/H'_{FD}max$
**Variables:**
**Calculation:** estimated visually from graph
**Value:** $\approx 0,7$

**Index:** Margalef index

**Reference:** Henry and Roberts 2007
**Formula:**$d = (s-1)/\log N$
**Variables:** N – the total number of individuals in the sample
**Calculation:** estimated visually from graph
**Value:** $\approx 7,75$

**Index:** Shannon diversity
**Reference:** Henry and Roberts 2007
**Formula:**$H' = \sum_i \log_{10} p_i$
**Variables:** $p_i$ – the relative abundance of the ith species in a sample
**Calculation:** estimated visually from graph
**Value:** $\approx 1,575$

**Index:** Richness (SR)
**Reference:** Henry and Roberts 2007
**Formula:** SR=S
**Variables:** S – total number of species
**Calculation:** estimated visually from graph
**Value:** $\approx 57,5$

**Index:** Pielou evenness (J')
**Reference:** Henry and Roberts 2007
**Formula:**$J' = H'_{obs}/H'_{max}$
**Variables:** $H'_{obs} = H'$
$H'_{max}$ – the highest possible H'
**Calculation:** estimated visually from graph
**Value:** $\approx 0,83$

**Index:** Shannon-Wiener
**Reference:** Gambi et al. 2003
**Formula:** $H' = \sum_{i=1}^{s} p_i \ln p_i$
**Variables:** $p_i = n_i/N$ – relative coverage of the i-th species; $n_i$ – number of individuals of species i; $N$ – size of the whole community; $S$ – the total number of species in each community
**Calculation:** avg across different stations
(3,1+3,2+3,2+2,7)/4=3
**Value:** $\approx 3$

**Index:** Evenness J'
**Reference:** Gambi et al. 2003
**Formula:** Pielou 1975
**Variables:**
**Calculation:** (0,87+0,83+0,89+0,86)/4=0,85
**Value:** $\approx 0,85$

**Index:** Species richness (SR)
**Reference:** Gambi et al. 2003
**Formula:**estimated from Margalef formula
$SR = (S-1)/\ln N$
**Variables:** S – number of species; N – number of individuals in a sample
**Calculation:** (7,9+13,0+14,5+14,1)/4=12,37
**Value:** $\approx 12,37$

**Index:** The index of trophic diversity (ITD)
**Reference:** Gambi et al. 2003
**Formula:** $ITD = \sum \theta^2$
**Variables:** $\theta$ – the contribution of density of each trophic group

**Calculation:** (0,275+0,31+0,28+0,34)/4=0,3
**Value:** $\approx 0,3$

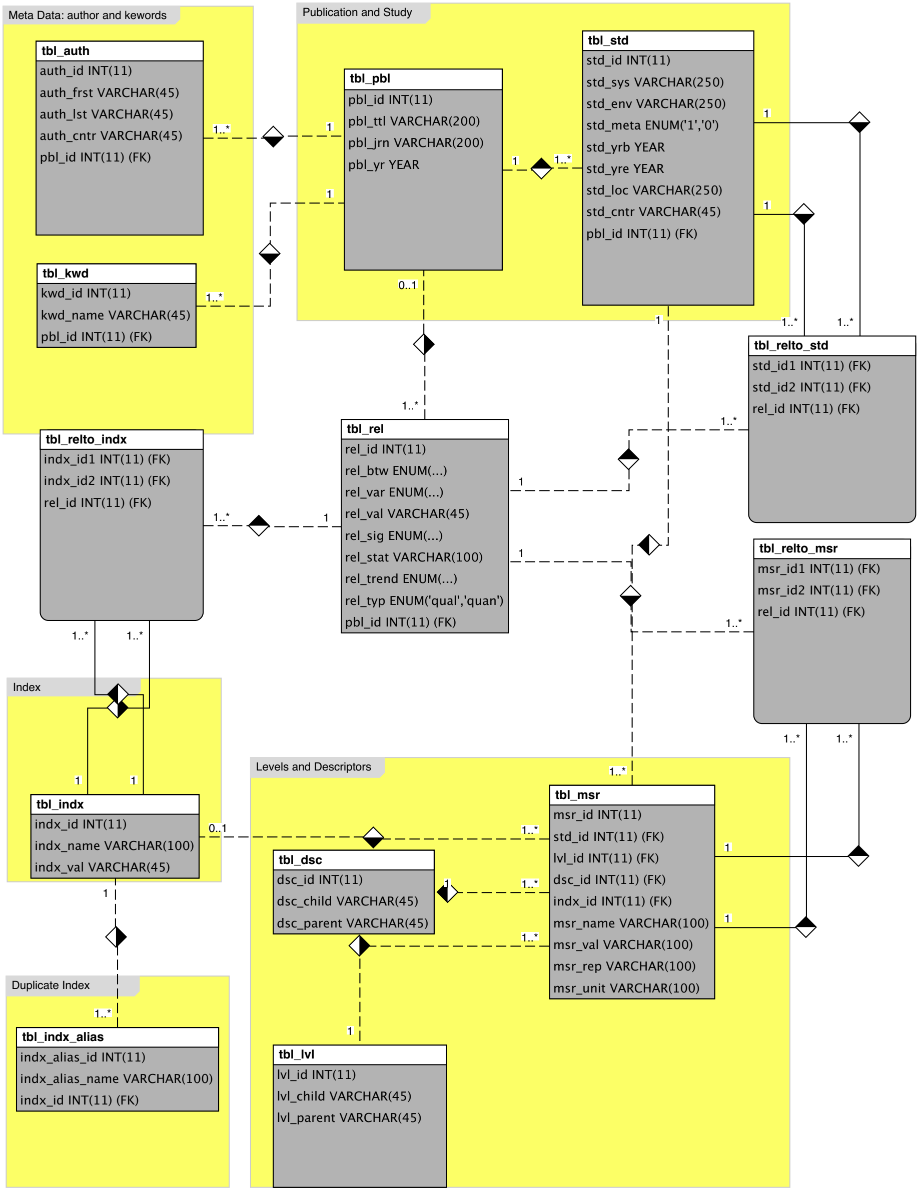
**Index:** The maturity index (MI)
**Reference:** Gambi et al. 2003
**Formula:** Bongers et al. 1991 $MI = \sum v(i)f(i)$
**Variables:** v – the c-p (colonisers-persisters) values of genus i;
f(i) – the frequency of that genus
**Calculation:** avg from various stations; (2,7+2,6+2,5+2,4)/4=2,55
**Value:** $\approx 2,55$

**Index:** Taxonomic diversity $\Delta$ (Taxonomic relatedness)
**Reference:** Gambi et al. 2003
**Formula:** Clarke and Warwick 1998; $\Delta = [\sum \sum_{i<j} \omega_{ij} x_i x_j]/[n(n-1)/2]$
for $\omega_{ij} = 1$ $\Delta = (1 - \sum_i p_i^2)/(1 - n^-1)$
**Variables:**$x_i$ – the abundance of ith species, n – the total number of individuals in the sample, $\omega_{ij}$ – the “distinctness weight” given to the path length linking species i and j
**Calculation:** (95,6+96,5+95,3+90,5)/4=94,47
**Value:** $\approx 94,47$

**Index:** Taxonomic distinctness $\Delta^*$
**Reference:** Gambi et al. 2003
**Formula:** Clarke and Warwick 1998; $\Delta^* = [\sum \sum_{i<j} \omega_{ij} x_i x_j]/[\sum \sum_{i<j} x_i x_j]$
**Variables:** $x_i$ – the abundance of ith species; $\omega_{ij}$ – the “distinctness weight” given to the path length linking species i and j
**Calculation:** (98,1+98,4+97,6+95,2)/4=97,32
**Value:** $\approx 97,32$

# Appendix 3.B Biodiversity data model

This appendix specifies ERD of relational database biodivDB which consists of 13 tables (entities). Key entities are in yellow boxes and include meta data, publication and study it contains, index $\mathcal{I}$, duplicate index, levels L and descriptors D that form a measure.



# Appendix 3.C Database biodivDB (SQL dump)

This is a dump file which can be used to reconstruct the database.

```
-- phpMyAdmin SQL Dump
-- version 2.11.3deb1ubuntu1.1
-- http://www.phpmyadmin.net
--
-- Host: localhost
-- Generation Time: May 07, 2009 at 12:21 PM
-- Server version: 5.0.51
-- PHP Version: 5.2.4-2ubuntu5.6
SET SQL_MODE="NO_AUTO_VALUE_ON_ZERO";
--
-- Database: `biodivDB`
--
-- --------------------------------------------------------
--
-- Table structure for table `tbl_auth`
--
CREATE TABLE IF NOT EXISTS `tbl_auth` (
  `auth_id` int(11) NOT NULL auto_increment,
  `auth_frst` varchar(45) default NULL,
  `auth_lst` varchar(45) default NULL,
  `auth_cntr` varchar(45) default NULL,
  `pbl_id` int(11) NOT NULL,
  PRIMARY KEY (`auth_id`),
  KEY `pbl_id` (`pbl_id`))
ENGINE=InnoDB DEFAULT CHARSET=utf8 AUTO_INCREMENT=88 ;
-- --------------------------------------------------------
--
-- Table structure for table `tbl_dsc`
--
CREATE TABLE IF NOT EXISTS `tbl_dsc` (
  `dsc_id` int(11) NOT NULL auto_increment,
  `dsc_child` varchar(45) default NULL,
  `dsc_parent` varchar(45) default NULL,
  PRIMARY KEY (`dsc_id`))
ENGINE=InnoDB DEFAULT CHARSET=utf8 COMMENT='descriptors tree' AUTO_INCREMENT=24 ;
-- --------------------------------------------------------
--
-- Table structure for table `tbl_indx`
--
CREATE TABLE IF NOT EXISTS `tbl_indx` (
  `indx_id` int(11) NOT NULL auto_increment,
  `indx_name` varchar(100) default NULL,
  `indx_val` varchar(45) default NULL COMMENT 'value index',
  PRIMARY KEY (`indx_id`))
ENGINE=InnoDB DEFAULT CHARSET=utf8 AUTO_INCREMENT=64 ;
-- --------------------------------------------------------
--
-- Table structure for table `tbl_indx_alias`
--
CREATE TABLE IF NOT EXISTS `tbl_indx_alias` (
  `indx_alias_id` int(11) NOT NULL,
  `indx_alias_name` varchar(100) default NULL COMMENT 'alias name',
  `indx_id` int(11) NOT NULL,
  PRIMARY KEY (`indx_alias_id`),
  KEY `indx_id` (`indx_id`)
ENGINE=InnoDB DEFAULT CHARSET=utf8;
-- --------------------------------------------------------
--
-- Table structure for table `tbl_kwd`
--
CREATE TABLE IF NOT EXISTS `tbl_kwd` (
  `kwd_id` int(11) NOT NULL auto_increment,
  `kwd_name` varchar(45) character set latin1 default NULL,
  `pbl_id` int(11) NOT NULL,
  PRIMARY KEY (`kwd_id`),
  KEY `pbl_id` (`pbl_id`))
ENGINE=InnoDB DEFAULT CHARSET=utf8 COMMENT='keywords for publication' AUTO_INCREMENT=165 ;
-- --------------------------------------------------------
--
-- Table structure for table `tbl_lvl`
--
CREATE TABLE IF NOT EXISTS `tbl_lvl` (
  `lvl_id` int(11) NOT NULL auto_increment,
  `lvl_child` varchar(45) default NULL,
  `lvl_parent` varchar(45) default NULL,
  PRIMARY KEY (`lvl_id`))
ENGINE=InnoDB DEFAULT CHARSET=utf8 COMMENT='levels tree' AUTO_INCREMENT=27 ;
-- --------------------------------------------------------
--
-- Table structure for table `tbl_msr`
--
CREATE TABLE IF NOT EXISTS `tbl_msr` (
  `msr_id` int(11) NOT NULL,
  `std_id` int(11) NOT NULL,
  `lvl_id` int(11) NOT NULL,
  `dsc_id` int(11) NOT NULL,
  `indx_id` int(11) default NULL,
  `msr_name` varchar(100) default NULL,
  `msr_val` varchar(100) default NULL,
  `msr_rep` varchar(100) default NULL,
  `msr_unit` varchar(100) default NULL,
  PRIMARY KEY (`msr_id`),
  KEY `std_id` (`std_id`),
  KEY `lvl_id` (`lvl_id`),
  KEY `dsc_id` (`dsc_id`),
  KEY `indx_id` (`indx_id`))
ENGINE=InnoDB DEFAULT CHARSET=utf8;
-- --------------------------------------------------------
--
-- Table structure for table `tbl_pbl`
--
CREATE TABLE IF NOT EXISTS `tbl_pbl` (
  `pbl_id` int(11) NOT NULL auto_increment,
  `pbl_ttl` varchar(200) character set latin1 default NULL COMMENT 'title of the
      publication',
  `pbl_jrn` varchar(200) character set latin1 default NULL,
  `pbl_yr` year(4) default NULL,
  PRIMARY KEY (`pbl_id`))
ENGINE=InnoDB DEFAULT CHARSET=utf8 AUTO_INCREMENT=31 ;
-- --------------------------------------------------------
--
-- Table structure for table `tbl_rel`
--
CREATE TABLE IF NOT EXISTS `tbl_rel` (
  `rel_id` int(11) NOT NULL auto_increment COMMENT 'primary key',
  `rel_btw` enum('std','msr','msrindx','indx','none') default 'none' COMMENT 'between
      measure or index',
  `rel_var` enum('lvl','dsc','lvldsc','none') default 'none' COMMENT 'variable level,
      descriptor, both or nothing',
  `rel_val` varchar(45) default NULL COMMENT 'value',
  `rel_sig` enum('yes','no','none') default 'none' COMMENT 'significance yes or no',
  `rel_stat` varchar(100) default NULL COMMENT 'statistics used',
  `rel_trend` enum('pos','neg','none') default 'none' COMMENT 'trend positive, negative or
       none',
  `rel_typ` enum('qual','quan') default NULL COMMENT 'type of the result: quantitative,
      qualitative',
  `pbl_id` int(11) default NULL COMMENT 'foreign key',
  PRIMARY KEY (`rel_id`),
  KEY `pbl_id` (`pbl_id`))
ENGINE=InnoDB DEFAULT CHARSET=utf8 COMMENT='relationship metadata' AUTO_INCREMENT=63 ;
-- --------------------------------------------------------
--
-- Table structure for table `tbl_relto_indx`
--
CREATE TABLE IF NOT EXISTS `tbl_relto_indx` (
  `indx_id1` int(11) NOT NULL,
  `indx_id2` int(11) NOT NULL,
  `rel_id` int(11) NOT NULL,
  PRIMARY KEY (`indx_id1`,`indx_id2`),
  KEY `rel_id` (`rel_id`),
  KEY `indx_id2` (`indx_id2`),
  KEY `indx_id1` (`indx_id1`))
ENGINE=InnoDB DEFAULT CHARSET=utf8;
-- --------------------------------------------------------
--
-- Table structure for table `tbl_relto_msr`
--
CREATE TABLE IF NOT EXISTS `tbl_relto_msr` (
  `msr_id1` int(11) NOT NULL,
  `msr_id2` int(11) NOT NULL,
  `rel_id` int(11) NOT NULL,
  PRIMARY KEY (`msr_id1`,`msr_id2`),
  KEY `rel_id` (`rel_id`),
  KEY `msr_id1` (`msr_id1`),
  KEY `msr_id2` (`msr_id2`))
ENGINE=InnoDB DEFAULT CHARSET=utf8;
-- --------------------------------------------------------
--
-- Table structure for table `tbl_relto_std`
--
CREATE TABLE IF NOT EXISTS `tbl_relto_std` (
  `std_id1` int(11) NOT NULL,
  `std_id2` int(11) NOT NULL,
  `rel_id` int(11) NOT NULL,
  PRIMARY KEY (`std_id1`,`std_id2`),
  KEY `rel_id` (`rel_id`),
  KEY `std_id2` (`std_id2`))
ENGINE=InnoDB DEFAULT CHARSET=utf8;
-- --------------------------------------------------------
--
-- Table structure for table `tbl_std`
--
CREATE TABLE IF NOT EXISTS `tbl_std` (
  `std_id` int(11) NOT NULL auto_increment,
  `std_sys` varchar(250) default NULL,
  `std_env` varchar(250) NOT NULL,
  `std_meta` enum('1','0') default NULL,
  `std_yrb` year(4) default NULL,
  `std_yre` year(4) default NULL,
  `std_loc` varchar(250) default NULL,
  `std_cntr` varchar(45) default NULL COMMENT 'country',
  `pbl_id` int(11) NOT NULL,
  PRIMARY KEY (`std_id`),
  KEY `pbl_id` (`pbl_id`))
ENGINE=InnoDB DEFAULT CHARSET=utf8 COMMENT='study' AUTO_INCREMENT=54 ;
--
-- Constraints for dumped tables
--
--
-- Constraints for table `tbl_auth`
--
```

```
ALTER TABLE `tbl_auth`
  ADD CONSTRAINT `tbl_auth_ibfk_1` FOREIGN KEY (`pbl_id`) REFERENCES `tbl_pbl` (`pbl_id`)
      ON DELETE CASCADE ON UPDATE CASCADE;
--
-- Constraints for table `tbl_indx_alias`
--
ALTER TABLE `tbl_indx_alias`
  ADD CONSTRAINT `tbl_indx_alias_ibfk_1` FOREIGN KEY (`indx_id`) REFERENCES `tbl_indx` (`
      indx_id`) ON DELETE CASCADE ON UPDATE CASCADE;
--
-- Constraints for table `tbl_kwd`
--
ALTER TABLE `tbl_kwd`
  ADD CONSTRAINT `pbl_id` FOREIGN KEY (`pbl_id`) REFERENCES `tbl_pbl` (`pbl_id`) ON DELETE
       CASCADE ON UPDATE CASCADE;
--
-- Constraints for table `tbl_msr`
--
ALTER TABLE `tbl_msr`
  ADD CONSTRAINT `tbl_msr_ibfk_1` FOREIGN KEY (`std_id`) REFERENCES `tbl_std` (`std_id`),
  ADD CONSTRAINT `tbl_msr_ibfk_2` FOREIGN KEY (`lvl_id`) REFERENCES `tbl_lvl` (`lvl_id`),
  ADD CONSTRAINT `tbl_msr_ibfk_3` FOREIGN KEY (`dsc_id`) REFERENCES `tbl_dsc` (`dsc_id`),
  ADD CONSTRAINT `tbl_msr_ibfk_4` FOREIGN KEY (`indx_id`) REFERENCES `tbl_indx` (`indx_id
      `);
--
-- Constraints for table `tbl_rel`
--
ALTER TABLE `tbl_rel`
  ADD CONSTRAINT `tbl_rel_ibfk_1` FOREIGN KEY (`pbl_id`) REFERENCES `tbl_pbl` (`pbl_id`)
      ON DELETE CASCADE ON UPDATE CASCADE;
--
-- Constraints for table `tbl_relto_indx`
--
ALTER TABLE `tbl_relto_indx`
  ADD CONSTRAINT `tbl_relto_indx_ibfk_1` FOREIGN KEY (`indx_id1`) REFERENCES `tbl_indx` (`
      indx_id`) ON DELETE CASCADE ON UPDATE CASCADE,
  ADD CONSTRAINT `tbl_relto_indx_ibfk_2` FOREIGN KEY (`indx_id2`) REFERENCES `tbl_indx` (`
      indx_id`) ON DELETE CASCADE ON UPDATE CASCADE,
  ADD CONSTRAINT `tbl_relto_indx_ibfk_3` FOREIGN KEY (`rel_id`) REFERENCES `tbl_rel` (`
      rel_id`) ON DELETE CASCADE ON UPDATE CASCADE;
--
-- Constraints for table `tbl_relto_msr`
--
ALTER TABLE `tbl_relto_msr`
  ADD CONSTRAINT `tbl_relto_msr_ibfk_3` FOREIGN KEY (`msr_id2`) REFERENCES `tbl_msr` (`
      msr_id`),
  ADD CONSTRAINT `tbl_relto_msr_ibfk_1` FOREIGN KEY (`rel_id`) REFERENCES `tbl_rel` (`
      rel_id`) ON DELETE CASCADE ON UPDATE CASCADE,
  ADD CONSTRAINT `tbl_relto_msr_ibfk_2` FOREIGN KEY (`msr_id1`) REFERENCES `tbl_msr` (`
      msr_id`);
--
-- Constraints for table `tbl_relto_std`
--
ALTER TABLE `tbl_relto_std`
  ADD CONSTRAINT `tbl_relto_std_ibfk_1` FOREIGN KEY (`std_id1`) REFERENCES `tbl_std` (`
      std_id`) ON DELETE CASCADE ON UPDATE CASCADE,
  ADD CONSTRAINT `tbl_relto_std_ibfk_2` FOREIGN KEY (`std_id2`) REFERENCES `tbl_std` (`
      std_id`) ON DELETE CASCADE ON UPDATE CASCADE,
  ADD CONSTRAINT `tbl_relto_std_ibfk_3` FOREIGN KEY (`rel_id`) REFERENCES `tbl_rel` (`
      rel_id`) ON DELETE CASCADE ON UPDATE CASCADE;
--
-- Constraints for table `tbl_std`
--
ALTER TABLE `tbl_std`
```

```
ADD CONSTRAINT `tbl_std_ibfk_1` FOREIGN KEY (`pbl_id`) REFERENCES `tbl_pbl` (`pbl_id`)
    ON DELETE CASCADE ON UPDATE CASCADE;
```

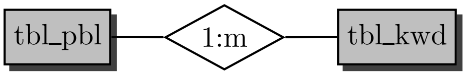
## Appendix 3.D Data dictionary

This appendix contains a data dictionary which gives an overview of all tables, their keys, attributes, data types and descriptions. Each atomised entity has alias and attributes (fields) which describe its properties including unique identifiers.

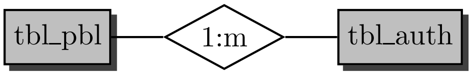
| Key | Attribute | Data type | Description |
|---|---|---|---|
| pk | pbl_id | int() | unique identifier |
| | pbl_ttl | varchar (200) | publication title |
| | pbl_jrn | varchar (200) | journal |
| | pbl_yr | year(4) | year of publication |
| pk | kwd_id | int() | unique identifier |
| | kwd_name | varchar(45) | keywords |
| fk | pbl_id | int() | |
| pk | auth_id | int() | unique identifier |
| | auth_frst | varchar(30) | first name |
| | auth_lst | varchar(30) | last name |
| | auth_cntr | varchar(30) | country |
| fk | pbl_id | int() | |
| pk | std_id | int() | unique identifier |
| | std_sys | varchar(250) | biological system |
| | std_yrb | year(4) | begin year |
| | std_yre | year(4) | end year |
| | std_loc | varchar(250) | location |
| fk | pbl_id | int() | |
| pk | rel_id | int(11) | unique identifier |
| | rel_btw | enum(“msr”,“indx”,“none”) | between measures or indices |
| | rel_var | enum(“lvl”, “dsc”, “lvldsc”, “none”) | variable: level, descriptor, both or none |
| | rel_sig | enum(“yes”, “no”, “none”) | significance: yes, no, none |
| | rel_stat | varchar(100) | statistics used |
| | rel_trend | enum(“pos”, “neg”, “none”) | trend: positive, negative, none |
| fk | pbl_id | int(11) | |
| pk | dsc_id | int() | unique identifier |
| | dsc_name | enum($D_1, D_2, \ldots, D_n$) | predefined set of descriptors |
| | dsc_val | int () | estimated value for descriptor |
| | dsc_unit | varchar (20) | unit |
| | dsc_oth | varchar(30) | other information |
| pk | lvl_id | int() | unique identifier |
| | lvl_name | enum($L_1, L_2, \ldots, L_n$) | predefined set of levels |
| pk | index_id | int() | unique identifier |
| | indx_name | varchar(20) | name of the index |
| | indx_rel | enum(“yes”, “no”) | relation |
| fk | lvl_id | int() | |
| fk | dsc_id | int() | |

# Appendix 3.E Relational cardinalities

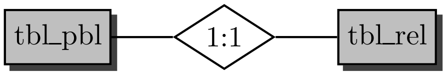
This appendix comprises of text descriptions for the relational cardinalities. The first relational cardinality to consider is “publication” `tbl_pbl` and “keywords” `tbl_kwd`. Entity `tbl_pbl` is an independent entity and it is used as a starting point of biodiversity data model.



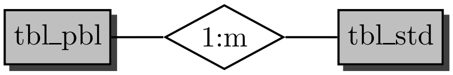
Each “publication” `tbl_pbl` that is found in biodiversity literature may have one or more “keywords” `tbl_kwd`; each “keyword” `tbl_kwd` may belong to only one “publication” `tbl_pbl`, hence one-to-many relationship. It does not imply, however, that each keyword must be unique, but a key assigned to it must be unique. For instance, the keyword “biodiversity” will certainly appear in most of the publications, but unique key assigned to each keyword on “many” end of the relationship (`tbl_kwd`) and reference to unique key on “one” end of the relationship (`tbl_pbl`) will ensure that the whole relationship can be uniquely identified. So, simple queries identifying all publications in the database which contain “biodiversity” as a keyword’ are possible.



Each “publication” `tbl_pbl` must have one or more “authors” `tbl_auth`; each “author” `tbl_auth` may appear only in one “publication”. Relationship between “publication” and “author” is, therefore, `1:m`. Although in reality one author may certainly publish more than one article, within a database each author will be assigned a unique identifier (primary key) which will allow us to distinguish between various papers published by the same author.



Each “publication” `tbl_pbl` may contain some information on “relationships” `tbl_rel` between different studies, measures or indices of biodiversity. This information can be some sort of statistics or comparison between different measures or studies and it belongs to at least one publication. Relationship between `tbl_pbl` and `tbl_rel` is `1:1`. Mapping a `1:1` relationship follows standard rules. Foreign key in `tbl_rel` may take `NULL` meaning that the information on relationship is not mandatory: publication may contain some meta-data on relationship but not necessarily. For instance, various measures of biodiversity may be reviewed and quantified in one publication but no relationship between them is considered.

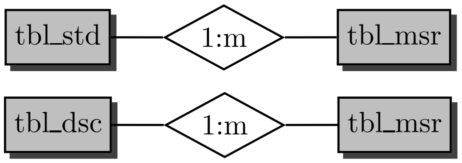


Each “publication” `tbl_pbl` must contain at least one “study” `tbl_std`; each “study” `tbl_std` can be found within a “publication” `tbl_pbl`. It gives us `1:m` relationship between `tbl_pbl` and `tbl_std` and implies that in each publication on measures of biodiversity at least one study can be found; equally, each study on measures of biodiversity must belong to at least one publication.

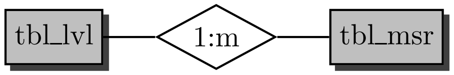
To accurately describe the semantics of an association among entities “study” `tbl_std`, “level” `tbl_lvl`, “descriptor” `tbl_dsc` and “index” `tbl_indx` an associative entity `tbl_msr` is used. As it has been discussed earlier, associative entity is a by-product of `m:m` relationship. Four entities that are simultaneously involved give a quaternary relationship. Each pair in this relationship shall be considered individually.

Each “study” `tbl_std` must contain at least one “measure” `tbl_msr`, each “measure” `tbl_msr` will appear in one “study” `tbl_std`. Relationship between this two entities is `1:m`, which reflects the fact that most of the studies have more than one measures. For instance, Balmford et al. (2000) have considered 4 measures in their study: species, genus, family, and order richness.

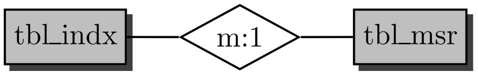
Each “descriptor” `tbl_dsc` can appear in one or more “measures” `tbl_msr`; each “measure” `tbl_msr` can belong to only one “descriptor” `tbl_dsc`, which gives `1:m` relationship. For instance, descriptor



“richness” can be used at species level to form species richness or at genetic level to form genetic richness.



Each “level” can be used to create different “measures”, but each “measure” refers to only one “level”. For instance, in species richness, species evenness, species rarity, and so on, taxonomic level species is used to generate various measures.



Each “measure” of biodiversity `tbl_msr` can be used to create only one “index” `tbl_indx` at a time, but each “index” `tbl_indx` can be made of one or more “measures.” That is compliant with the definition 2.5. For example, Simpson diversity index is calculated using two measures of biodiversity, i.e., species richness and species abundance. In turn, each of this measures can be used to calculate yet another index, e.g., Shannon index. Indices that are composed of identical measures can be mathematically derived from one another.

Tables `tbl_lvl`,`tbl_dsc`,`tbl_indx` are modelled as lookup tables which are a fixed list of data. Moreover, `tbl_lvl`,`tbl_dsc` contain hierarchical data (see Figures 2.3, 2.5). Only terminal node were recorded in `tbl_msr`, but there is a possibility to retrieve the whole tree.

# Appendix 3.F Biodiversity data mining

This appendix illustrates some of the queries performed on database biodivDB to explore the content, statistic on central tendencies and variability of biodiversity data.

**Query 3.1:** Measures and study environments for a given study system (e.g., grasslands)

```
select msr_name, std_env from tbl_msr, tbl_std
where tbl_std.std_sys like "%grass%"
and tbl_std.std_id=tbl_msr.std_id;
```

**Query 3.2:** Geographical spread of biodiversity studies

```
mysql> select std_cntr, count(*) from tbl_std  group by std_cntr;
+----------------------+----------+
| std_cntr             | count(*) |
+----------------------+----------+
| NULL                 |        1 |
| Africa               |        1 |
| Antarctica           |        1 |
| Atlantic             |        2 |
| Australia            |        3 |
| China                |        5 |
| France               |        1 |
| Iberian Peninsula    |        1 |
| India                |        3 |
| Ireland              |        3 |
| Italy                |        4 |
| Japan                |        8 |
| Jordon              |        1 |
| Mexico              |        1 |
| Morocco             |        1 |
| Portugal            |        1 |
| Romania             |        1 |
| South Africa        |        1 |
| South Pacific Ocean |        1 |
| Spain               |        3 |
| Sweden              |        1 |
| Tunisia             |        1 |
| UK                  |        2 |
| USA                 |        6 |
+---------------------+----------+
24 rows in set (0.00 sec)
```

**Query 3.3:** A list of biodiversity measures for which indices were calculated

```
select tbl_msr.msr_name, tbl_indx.indx_name
from tbl_msr, tbl_indx
where tbl_msr.indx_id=tbl_indx.indx_id
and tbl_msr.indx_id is not null;
```

**Query 3.4:** All instances of measure which contain “richness”

```
select msr_name, count(*)
from tbl_msr where msr_name like "%richness%"
group by msr_name;
+----------------------------------+----------+
| msr_name                         | count(*) |
+----------------------------------+----------+
| alpha diversity/species richness |        1 |
| families richness                |        3 |
| genera richness                  |        5 |
| orders richness                  |        2 |
| ribotype richness                |        3 |
| species richness                 |       46 |
| taxa richness                    |        6 |
| taxonomic richness               |        2 |
+----------------------------------+----------+
8 rows in set (0.00 sec)
```

**Query 3.5:** Range of the values reported for each measure

```
select msr_name, msr_val from tbl_msr
where msr_val is not NULL order by msr_name;
+----------------------------------+---------+
| msr_name                         | msr_val |
+----------------------------------+---------+
| alpha diversity                  | 15...139|
| beta diversity                   | 37      |
| community composition            | 23...94 |
| families richness                | 26...64 |
| functional groups                | 28      |
| genera richness                  | 6...168 |
| genetic distance                 | 16      |
| habitat assemblage               | 13      |
| mobility guilds                  | 2       |
| number alleles                   | 20      |
| number polymorphic fragments     | 41...62 |
| number species                   | 24      |
| orders richness                  | 8...26  |
| percentage polymorphic fragments | 32...48 |
| phylogenetic diversity           | 1582    |
| ribotype richness                | 23...94 |
| species abundance                | 18...8196|
| species assemblage               | 5       |
| species density                    | 2536    |
| species evenness                   | 0,92    |
| species richness                   | 10...315|
| taxa richness                      | 84      |
| taxonomic distinctness             | 96,8    |
| taxonomic diversity                | 93,5    |
| taxonomic relatedness              |         |
| taxonomic richness                 | 43...977|
| trophic diversity                  | 4...6   |
| trophic guilds                     | 5       |
+------------------------------------+---------+
```

**Query 3.6:** Distribution of D across L – “species”

```
select tbl_dsc.dsc_child as descriptor,
tbl_msr.msr_name, lpad('*',count(*),'*') as count
from tbl_msr, tbl_lvl, tbl_dsc
where (tbl_dsc.dsc_id=tbl_msr.dsc_id and tbl_lvl.lvl_id=tbl_msr.lvl_id)
and tbl_lvl.lvl_child='species'
group by tbl_dsc.dsc_child;
```

**Query 3.7:** Occurrence of different D|L and measures M

```
select tbl_dsc.dsc_child as descriptor, tbl_lvl.lvl_child as level,tbl_msr.msr_name, count
    (*)
from tbl_msr, tbl_lvl, tbl_dsc
where (tbl_dsc.dsc_id=tbl_msr.dsc_id and tbl_lvl.lvl_id=tbl_msr.lvl_id)
and tbl_dsc.dsc_child='richness'
group by tbl_lvl.lvl_child
union select tbl_lvl.lvl_child as level,
tbl_msr.msr_name, lpad('*',count(*),'*') as count
from tbl_msr, tbl_lvl, tbl_dsc
where (tbl_dsc.dsc_id=tbl_msr.dsc_id and tbl_lvl.lvl_id=tbl_msr.lvl_id)
and tbl_dsc.dsc_child='richness'
group by tbl_lvl.lvl_child;
```

```
+--------------+---------------------+---------------------------------------------------+
| D/L          | M                   | Occurrence (times)                                |
+--------------+---------------------+---------------------------------------------------+
| abundance    | species abundance   | **************************************            |
| assemblage   | species assemblage  | **                                                |
| bdiversity   | beta diversity      | **                                                |
| composition  | species composition | ******                                            |
| density      | species density     | *                                                 |
| dominance    | Berger-Parker       | ********                                          |
| evenness     | species evenness    | *********                                         |
| number       | number species      | *                                                 |
| rarity       | species rarity      | *********                                         |
| richness     | species richness    | ************************************************* |
| family       | families richness   | **                                                |
| genes        | ribotype richness   | ***                                               |
| genus        | genera richness     | ******                                            |
| order        | orders richness     | **                                                |
| species      | species richness    | ************************************************* |
| taxonomic    | taxa richness       | ********                                          |
+--------------+---------------------+---------------------------------------------------+
```

**Query 3.8:** Retrieving a full tree of levels

```
select t1.lvl_child as lev1, t2.lvl_child as lev2, t3.lvl_child as lev3,
t4.lvl_child as lev4, t5.lvl_child as lev5
from tbl_lvl as t1
left join tbl_lvl as t2 on t2.lvl_parent=t1.lvl_id
left join tbl_lvl as t3 on t3.lvl_parent=t2.lvl_id
left join tbl_lvl as t4 on t4.lvl_parent=t3.lvl_id
left join tbl_lvl as t5 on t5.lvl_parent=t4.lvl_id
```

```
where t1.lvl_child='level'
+-------+------------+------------+---------------+---------+
| lev1  | lev2       | lev3       | lev4          | lev5    |
+-------+------------+------------+---------------+---------+
| level | taxonomic  | genes      | NULL          | NULL    |
| level | taxonomic  | subspecies | NULL          | NULL    |
| level | taxonomic  | genus      | NULL          | NULL    |
| level | taxonomic  | family     | NULL          | NULL    |
| level | taxonomic  | order      | NULL          | NULL    |
| level | taxonomic  | class      | NULL          | NULL    |
| level | taxonomic  | phylum     | NULL          | NULL    |
| level | taxonomic  | kingdom    | NULL          | NULL    |
| level | taxonomic  | species    | NULL          | NULL    |
| level | functional | roles      | producers     | NULL    |
| level | functional | roles      | consumers     | trophic |
| level | functional | roles      | recyclers     | NULL    |
| level | functional | orglevel   | community     | NULL    |
| level | functional | orglevel   | metacommunity | NULL    |
| level | functional | orglevel   | biome         | NULL    |
| level | functional | orglevel   | ecosystem     | NULL    |
+-------+------------+------------+---------------+---------+
16 rows in set (0.00 sec)
```

**Query 3.9:** Retrieving a full tree of descriptors

```
select t1.dsc_child as lev1, t2.dsc_child as lev2, t3.dsc_child as lev3,
t4.dsc_child as lev4
from tbl_dsc as t1
left join tbl_dsc as t2 on t2.dsc_parent=t1.dsc_id
left join tbl_dsc as t3 on t3.dsc_parent=t2.dsc_id
left join tbl_dsc as t4 on t4.dsc_parent=t3.dsc_id
where t1.dsc_child='descriptor'
+------------+----------+-------------+------------+
| lev1       | lev2     | lev3        | lev4       |
+------------+----------+-------------+------------+
| descriptor | number   | richness    | adiversity |
| descriptor | feature  | phenotype   | NULL       |
| descriptor | feature  | morphology  | NULL       |
| descriptor | feature  | genhomology | NULL       |
| descriptor | pattern  | composition | NULL       |
| descriptor | pattern  | evenness    | frequency  |
| descriptor | pattern  | evenness    | dispersion |
| descriptor | pattern  | abundance   | rarity     |
| descriptor | distance | bdiversity  | NULL       |
| descriptor | function | interaction | NULL       |
| descriptor | function | process     | NULL       |
+------------+----------+-------------+------------+
11 rows in set (0.01 sec)
```

# Chapter 4

# A simulation study: Biodiversity from bootstrap resampled communities

## 4.1 Why simulation?

The evident lack of consistency among study designs strongly limits any attempt to describe relationships among real measurements of biodiversity. For this reason, comparative analyses of biodiversity metrics typically use simulated data. Several simulation studies have been aimed at revealing relationships among multiple levels of community organisation. These include taxonomic diversity (Clarke and Warwick, 1998, 2001), functional diversity (Villéger et al., 2008), species and genetic diversity (Vellend, 2005), and phylogenetic diversity (Nipperess et al., 2010) indicators.

Representing different aspects of biodiversity, these studies contain one of the primary attributes – composition, structure, and function. These attributes were first recognised by Franklin (1988) and, subsequently, elaborated into a nested hierarchy by Noss (1990). According to Noss's characterisation, a four-level organisation of biodiversity is needed (i.e., regional landscape, community-ecosystem, population species, and genetic). Since the total biodiversity is determined by these three attributes, individual categories of indicators that contain these attributes need to be united.

While comparing different indicators, Clarke and Warwick (1998) concluded that the overall comparability of biodiversity estimates is compromised by different sources of variability. Using the statistical sampling properties of indices only certain aspects of diversity may be validly compared. Clarke and Warwick argued that if this variability has a random

Adapted from Lyashevska, O. and K. Farnsworth, 2011. How many dimensions of biodiversity do we need? Submitted to Ecological Indicators, October 2011 character overall results should remain unchanged. Now, thinking in terms of variability in biodiversity estimates revealed by the empirical approach, it persisted even after the inclusion of the complete set of variables. Complementing their previous work, Clarke and Warwick (2001) explored taxonomic relatedness patterns by constructing a simulation distribution from random subsets of species lists of free-living marine nematodes, thus allowing to elicit unevenness in biodiversity structure at different levels of a taxonomic or phylogenetic tree.

The independence of several functional diversity indices from each other and species richness was demonstrated by Villéger et al. (2008) through simulation of artificial datasets. Based on their simulations, Villéger et al. (2008) suggested decomposition of these indices into their primary components to provide a meaningful framework for biodiversity quantification.

Vellend (2005) used a set of spatially explicit simulations to investigate patterns of species-genetic diversity that are influenced by different sources of heterogeneity (e.g., environmental heterogeneity). He assessed multiple emergent properties simultaneously and discovered moderate to strong positive correlation between species richness and genotypic richness. It has been discovered, that depending on the characteristics of the species for which genetic diversity was measured (e.g., rare versus common), the strength of the correlation between the two types of diversity varies.

Other studies that use artificial data adopting a variety of simulation approaches include Hubalek (2000), McGill (2003), Pla (2004), Chao et al. (2005), Mendes et al. (2008), and Bevilacqua et al. (2009). Mendes et al. (2008), for example, used an extensive fish dataset to simulate artificial communities through either log-normal distributions of species richness or rarefaction from real communities.

Now, looking at these examples of the simulation studies in biodiversity research, it is obvious that their application has been largely confined to individual levels of biodiversity that were typically attributed to a narrow study system. It still remains to expand these simulation studies by uniting the individual categories of biodiversity into one analysis. Therefore, here I extend simulation studies by uniting the individual attributes – composition, structure, and function (using Franklin's terminology) of biodiversity into one analysis.

By generating biodiversity complexity from bootstrap resampled species lists, I examine the correlation amongst and sensitivity to a mix of taxonomic, structural, and functional diversity indices – each corresponding to one attribute of biodiversity. Due to a relatively small species list, an assumption on asymptotic distribution might be weak bootstrap procedure seems to be more appropriate (Dixon, 2002). In theory, this includes drawing species randomly with replacement from a global species list which allows estimation of an unknown sampling distribution. However, in practice bootstrapping is more complicated due to the presence of hierarchy introduced by taxonomic levels. This requires estimation of empirical probability distribution functions for taxonomic distribution across levels – such as orders in classes, family in orders, genera in families, and species in genera. Knowledge of the empirical distributional properties of taxa among the categories of diversity allows for realistic relations to be built into the test data, which has not been done in previous simulation studies.

Previous studies examined indicators against unstructured, random assemblies of "species", rather than realistic simulations of communities. For instance, Villéger et al. (2008) simulated artificial data sets, by generating species and their abundances using uniform statistical distributions without reference to empirical distributions. This severely limited the possibility of finding relations that cut across the descriptor categories of Franklin (1988) or those operating at different levels of biological organisation. To enable possible relationships among very different aspects of biodiversity to emerge, I built artificial communities with taxonomic structures and distributions of species traits that statistically matched an example of near-shore temperate marine ecosystems.

Looking for a way to reduce the number of biodiversity metrics into a single, but comprehensive description of biodiversity I will simulate a wide range of model ecological communities with controlled variation. This allows comparison of biodiversity estimates across communities in a consistent, and comprehensive way providing quantitative evidence on the degree of their relatedness. Before this can be done, I need to be clear what constitutes a community. Community as a concept is not easily defined: an ongoing debate in community ecology indicates certain conceptual problems (see, e.g., Looijen and van Andel, 1999). Here, let a community be defined as a set of co-occurring organisms of different taxa that can co-exist in a given space and time. Similarly, artificial (i.e., simulated or hypothetical) communities are defined as a set of co-occurring elements of different taxa which are generated by a single model run. To ensure that they are not a random collection of organisms (Cohen et al., 2003) but a collection of organisms functionally interlinked, their taxa content and distribution of traits need to be plausible.

Caron-Lormier et al. (2009) highlighted the importance of an ecologically feasible representation of community for facilitating predictive modelling. Using an arable food-web as an example and applying a functional trait approach to simulate ecosystem, they demonstrated that "trophic-functional typing can be used to explore the structure, diversity, and dynamics" in an ecosystem. In the same spirit, here, I construct an entire hypothetical taxonomic tree through stratified taxonomic sampling, following Clarke and Warwick (1998) by using the global species list as a staring point for simulation. Different taxonomic sampling schemes which include different degree of randomness are outlined by Hillis (1998) (Table 4.1).

Since, the choice of taxonomic sampling can have important consequences for the realism

**Table 4.1:** "Hillis" taxonomic sampling schemes. Adapted from Hillis (1998)

| Sampling scheme | Description |
|---|---|
| Scheme 1 | Random sample from the tree of life |
| Scheme 2 | Random sample from the group of interest |
| Scheme 3 | Purposefully select representative taxa within the group of interest |
| Scheme 4 | Select taxa within the group that are expected to subdivide long branches in the initial tree |
| Scheme 5 | Add (and delete) taxa until feasible results are achieved |

of biodiversity estimates, using computer simulation (see, e.g., Graybeal, 1998; Kim, 1996; Yang and Goldman, 1997) the available strategies were explored. From these studies it follows that completely random taxonomic sampling from the tree of life or from the groups of interest (scheme 1 and 2 respectively) leads to the dominance of certain groups of taxa, which is an undesirable effect (Hillis, 1998). While Kim's study is mostly concerned with sampling strategy 1 and Graybeal's with strategy 4, Hillis expects most studies to select strategy 3.

I selected a type of stratified random sampling, which is a mixture of strategies outlined above. Through a computer simulation, each consecutive run generates an artificial ecological community, consisting of a set of elements (e.g., species) – a random model construct. Obviously, in bootstrapping hierarchical data while accounting for the distribution of species within higher taxonomic levels, resultant collections of species are not completely random. Effectively, this is achieved through applying so-called balanced bootstrap which forces each species to occur a specified number of times in the collection of bootstrap simulated communities. However, this does not force each community to contain all species: one species may occur many times in one community and not at all in others (Dixon, 2002).

To control the taxonomic structure and species distribution in the process of stratified taxonomic sampling realistic taxa-distributional properties (i.e., distribution of taxa within higher taxonomic levels) are needed. These properties, reflected in species body size, allow for prediction of species density distributions. General trends in community size structure suggest that if species are drawn at random, empirically we are likely to encounter more small-sized organisms (see, e.g., Peters, 1983). To satisfy this requirement simulation algorithm relies on taxa-distributional data at different taxonomic levels which is ensured by hierarchical bootstrapping. Therefore, empirical distributional properties of the species lists I use for simulations, need to determined. This includes the species richness distribution at each taxonomic level.

To summarise, artificial ecological communities with controlled properties and matching real ecological communities enable me to analyse patterns in relationships between biodiversity estimates that emerge during simulation. While building communities I want to move away from using random subsets of species lists (as in, e.g., Clarke and Warwick, 1998) to explicitly account for taxonomic structure and distribution of taxa as they appear in the primary source dataset – near-shore temperate marine ecosystems.

## 4.2 Empirical distributional properties

At every stage of community generation, from establishing and instantiating a taxonomic topology to assigning abundances, it is essential to have an understanding of the distributable properties of the species lists that are used. In this context, generally referred to as taxa-distributional properties, this process includes exploration of the variation of the form of diversity patterns with changes in taxonomic resolution. Controlling for these properties during bootstrapping ensures a close match between real and simulated communities. To allow for this, in this section I will explore in more depth, the empirical distributional patterns of the taxa across different taxonomic resolution using the following empirical biodiversity source data:

- ITIS – the Taxonomically Structured Species Database (Bisby et al., 2009);
- BioMar – Irish benthic marine database – (Picton et al., 1992); and
- Benthic Biological Traits Information Catalogue – (BIOTIC, 2010).

ITIS is a large-scale dataset held at the Global Biodiversity Information Facility website (GBIF) hosted in Copenhagen and it is a freely available SQL-dump of taxonomically structured data covering $4 \times 10^5$ species spread over nearly two hundred classes[1]. BioMar is a dataset containing information on approximately $2 \times 10^3$ species and it is held at the Environmental Science Unit, Trinity College Dublin. In order to generate artificial ecological communities, records from both datasets were merged in one database. BIOTIC was obtained from the University of Plymouth (UK) and was used to derive matrices of functional traits (maximum number of traits available is 40) per taxonomic group in order to calculate functional diversity indices. An overview of the traits is given in Table 4.2.

These traits were coupled with taxonomic units of the simulated communities.

At different taxonomic levels, changes in the shape of distribution are typically manifested either through (a) taxon delimitation variation; or (b) unit variation (Storch and Sizling, 2008). Considering option (a), taxon delimitation variation is responsible for patterns revealed for a set of species by narrowing or broadening taxonomic resolution. The question of interest is whether distribution within a given taxon also applies for every taxon delimitation (i.e., distribution of species at genera compared to classes). Then, it is said, that

[1] accessed on 01-30-2009

**Table 4.2:** An example of functional traits and their possible values obtained to calculate functional diversity indices

| Traits | Selected values |
|---|---|
| Food type | Zooplankton, Phytoplankton, Detritus, Suspended particles |
| Size | 1-50 cm |
| Habitat | Free living, Attached, Erect |
| Regeneration | Yes/No |
| Life span | 1-100 years |
| Reproduction frequency | Annual/Biannual, Protracted/Episodic |
| Fertilisation type | External/Internal |
| Biogeographic range | Cold/Temperate |
| Depth range | 0-1765 m |
| Biozone | Littoral/Pelagic |
| Environmental position | Epifaunal, Epifloral, Demersal, Pelagic |
| Feeding method | Herbivore, Predator, Scavenger, Suspension feeder |
| Growth form | Radial, Stellate, Turf |
| Mobility | Crawler, Drifter, Swimmer |
| Reproduction type | Vegetative, Budding, Self fertilisation |

if pattern is taxon invariant, it follows a strong principle of taxon invariance. This implies that patterns in species distributions do not change with taxonomic resolution. Option (b) – unit variation involves variation in the patterns that occurs by changing fundamental taxonomic units (changing scope from species to genera, etc.).

Here, following Storch and Sizling (2008) I did not assume taxonomic unit invariance, so that the distribution of number of daughter taxons within a parent differed among levels in the taxonomic hierarchy. I, therefore, need to calculate the set of empirical probability distributions: Orders in Classes (OiC), Families in Orders (FiO), Genera in Families (GiF), Species in Genera (SiG) from which the number of orders in a class, number of families in an order, number of genera in a family and number of species in a genus, respectively, can be sampled. (Note: the use of these distributions across all taxonomic units of equivalent level, tacitly assumes weak taxonomic invariance (Storch and Sizling, 2008) within taxonomic units).

### 4.2.1 Large-scale properties

To investigate taxa-distributional properties that inform empirical probability distributions on a large scale I will use a taxonomically structured species database – ITIS, which is a data source represented in a form of taxonomic tree. This data source is ideal for establishing a taxonomic tree topology through taxonomic sampling algorithm. Having correct classification system in place, it allows the identification of any individual taxonomic elements at any level. It has been recreated from an SQL-dump with entity `Taxonomic_units` being

**Table 4.3:** The unit variation at different levels in taxonomically structured species database (ITIS)

| | Bacteria | Plantae | Animalia | Fungi | Protozoa |
|---|---|---|---|---|---|
| Species | 1167 | 61822 | 291211 | 2952 | 2010 |
| Genus | 136 | 6309 | 41988 | 955 | 552 |
| Families | 45 | 1000 | 6036 | 460 | 257 |
| Orders | 18 | 285 | 636 | 110 | 64 |
| Classes | 2 | 47 | 101 | 25 | 18 |
| Phyla | 3 | 34 | 47 | 7 | 5 |
| Kingdoms | 2 | 1 | 1 | 1 | 1 |

of primary interest. As it contains a recursive (`1:m`) taxonomic structure, an adjacency list model proved to be the most appropriate.

Here, I will start by exploring the unit variation and the form of the diversity patterns that emerge from this variation. The following questions will be addressed: How many phyla are in one kingdom? How many classes are in each phyla? How many orders are in a typical class? ITIS taxonomic data spans over kingdoms of Bacteria, Monera, Plantae, Animalia, Fungi, Protozoa, Chromista, and, therefore, represents a 7-tier relationship.

The taxa-distributional properties across taxonomic levels for each of the kingdoms are shown in Table 4.3.

A brief analysis of numbers of subtaxa within each taxa (Table 4.3) shows that the most abundant kingdom is Animalia, closely followed by Plantae. These distributional properties of the unit variation are further visualised on the Figure 4.1.

The taxa-distributional properties emerging at each taxonomic resolution (Figure 4.1) are clearly distinct from the species richness-distributional properties, which proves Storch and Sizling theory. While distributable property "taxa" can include anything from species to kingdoms, "species richness" is conceived as a number of species or species count at each taxonomic level. Thus, this first variation is the unit variation and the second is the taxon delimitation variation.

Now, considering the taxon delimitation variation which comes in the form of a species count within each kingdom and genus, I construct Figure 4.2. This graph shows that all distributions are right-skewed while the left tail is extremely short. The median values for kingdoms Protozoa, Plantae, and Animalia are sufficiently similar because they have similar number of species per genus. The presence of potential outliers in Animalia implies that there are a several taxonomic groups with a particularly high number of species.

For a taxonomic tree topology to be realistic, it should resemble these statistical properties. The density plots of species richness distributions across several taxonomic levels are superimposed in one graph (Figure 4.3). This graph reveals patterns in the taxon

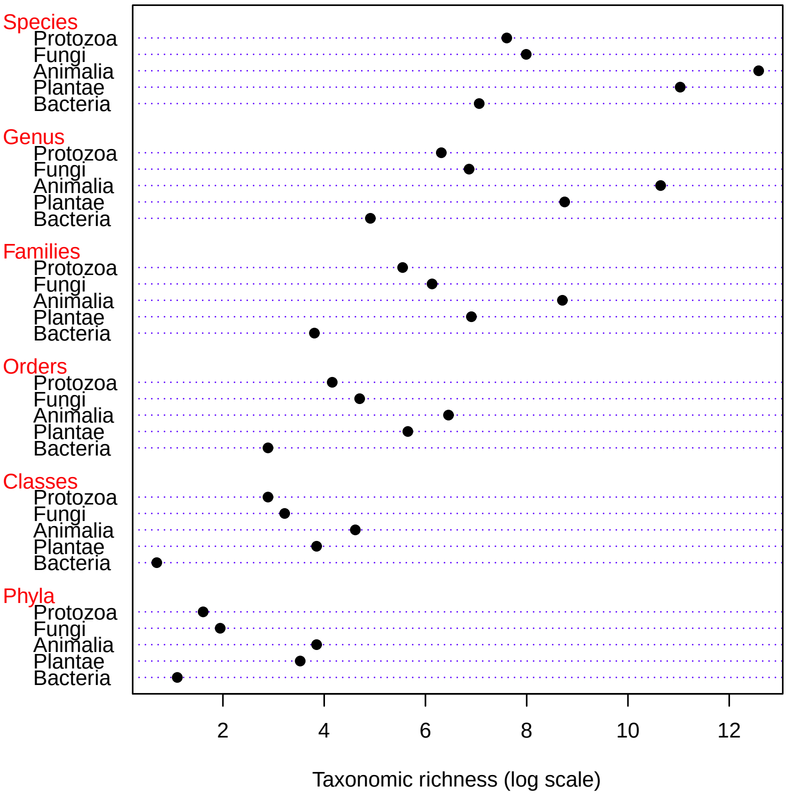


**Figure 4.1:** The unit variation in distributional properties of taxonomic richness at each taxonomic resolution

delimitation variation, which determine OiC, FiO, GiF, SiG.

The probability density function not only reveals the distributional properties and the shape of the data, but also allows one to determine whether distributions of taxa within higher taxa have similar distribution. This is of direct relevance for the assumptions made with respect to the taxonomic unit invariance. Thus, from the Figure 4.3 it is evident that the modality of the species-richness distribution varies with each taxonomic level: It is multimodal at low taxonomic level and unimodal at high taxonomic level. Using the non-parametric maximum likelihood estimate of cumulative distribution function (CDF) an empirical cumulative distribution function (ECDF) of species richness is formed (Figure 4.4).

From Figure 4.4 it follows that individual ECDF's converge towards the both ends. While density estimates emphasize local features such as modality and shape of the species richness-distributional properties, quantile-quantile (Q-Q) plots are more suitable for investigating global features. To assess whether data is well-described by a chosen probability distribution I shall use Q-Q plots (Figure 4.5).

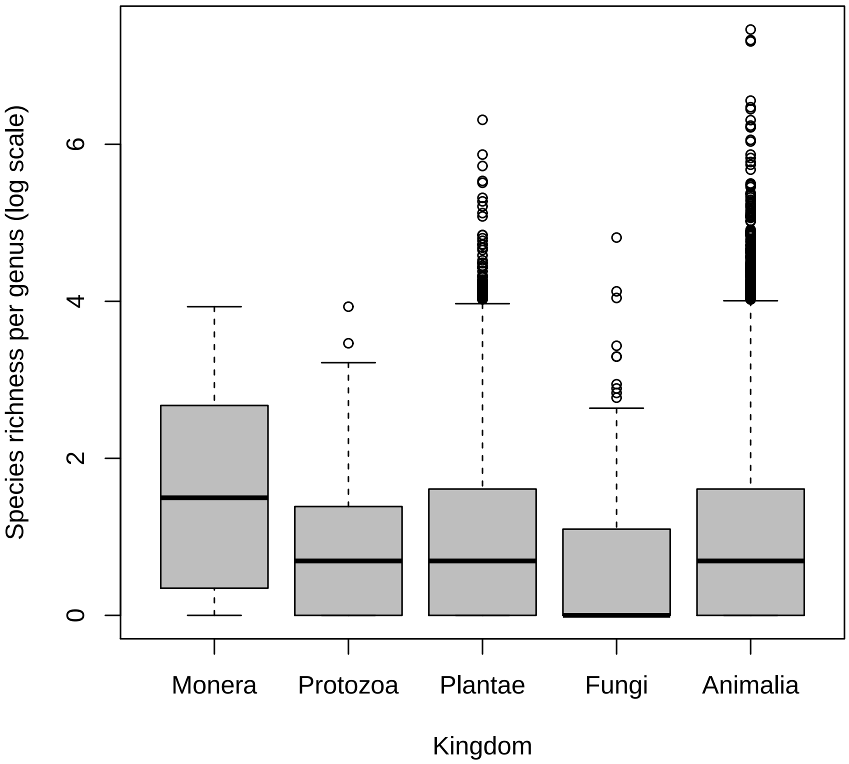


**Figure 4.2:** The taxon delimitation variation: species richness per genera on a logarithmic scale across different kingdoms. Distributions in Protozoa, Plantae, and Animalia, as it is shown by quantiles, have similar medians. All distributions are skewed to the right to some degree, with Plantae and Animalia having the heaviest tails. Points which are plotted at the end of the whiskers are potential outliers

Here, the quantiles of the observed data are plotted against equivalent quantiles of a normal probability distribution – the hypothetical match. The resulting approximately linear Q-Q plot suggests a good fit of species richness distribution at different taxonomic levels to normal distribution, despite deviations in the left tail for each of the taxonomic levels. These deviations are the extremes that do not match the quantiles well. Further interpreting Figure 4.5, no visible systematic convexity (which is consistent with the right-skewed distribution) is detected. As expected, high taxonomic level have steeper slope suggesting the systematic increase in variance in species richness distribution with taxonomic level. Overall, Figure  4.5 suggests that species richness-distributional properties are approximately equivalent under a linear transformation (i.e., they are a linear transformations of each other).

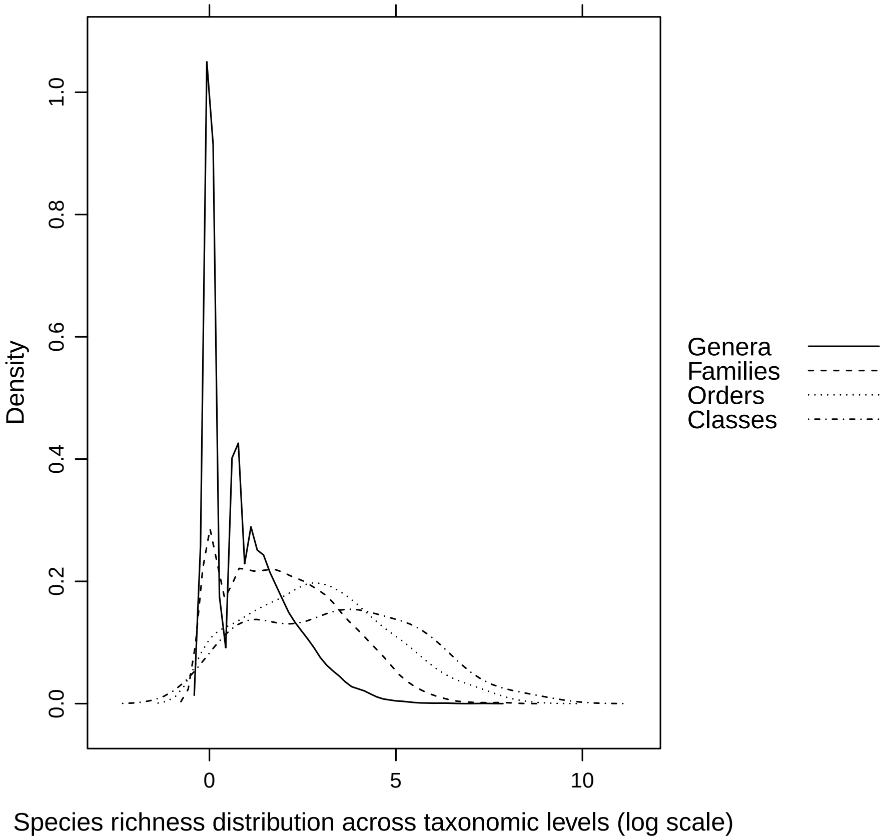


**Figure 4.3:** The grouped kernel density plot of distribution of species richness on a logarithmic scale across different taxonomic levels superposed in a single graph. The Gaussian kernel is used to compute the estimated density. Species richness distribution at all taxonomic levels in ITIS dataset has positive skew. This indicates that the bulk of the values lies to the left

### 4.2.2 Small-scale properties

Taking an approach similar to the large-scale data presented above, the taxon delimitation variation and its distributional properties for the log of species richness in BioMar Database is explored in Figure 4.6.

BioMar is a survey of littoral and sublittoral biotopes, and species found around Ireland conducted from September 1992 to June 1997. The results of this survey are organized in a form of relational database which I will use as guidance in identifying a plausible taxa-content for simulating ecological communities. Species surveys have been conducted across 16 counties in Ireland with county Down and county Leitrim having the highest and the smallest number of surveys (716 and 2 respectively); mean number of surveys per county was 155. As per BioMar specification, species which were encountered represent 27 phyla and 318 orders. The minimum number of orders identified within a phylum was 1,

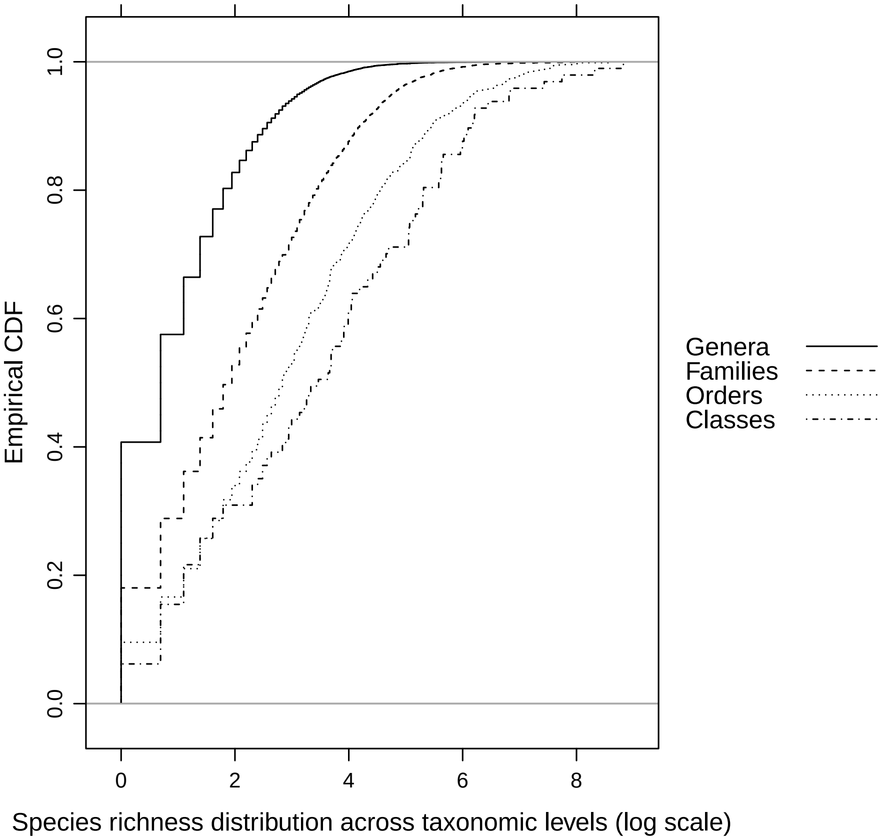


**Figure 4.4:** The ECDF for the log of species richness for each of the taxonomic level in ITIS dataset

the maximum was 39 and the mean value was 13. Number of families in phyla are in the range from 1 to 112 with the mean value of 23 families.

Overall, the taxon delimitation variation expressed as species richness distribution at all taxonomic resolutions in BioMar is positively skewed. This suggests that many taxonomic groups have extremely low number of species. The non-parametric maximum-likelihood estimate for the CDF is shown on Figure 4.7.

A shift towards negative skew observed at higher taxonomic resolution is revealed on the normal Q-Q plot for the empirical dataset (Figure 4.8). Although this was not visible from density plots (Figure 4.6), it accords with the expectations: species richness accumulated at high taxonomic levels are likely to produce taxonomic groups with large values.

As with the Q-Q plot in ITIS, observation made from the density plots can be confirmed – higher taxonomic levels have steeper slopes, suggesting an increase in the variance in species richness distribution with taxonomic level. Additionally, the presence of systematic curvature suggests a shift towards left-skewed distribution which is clearly visible at high

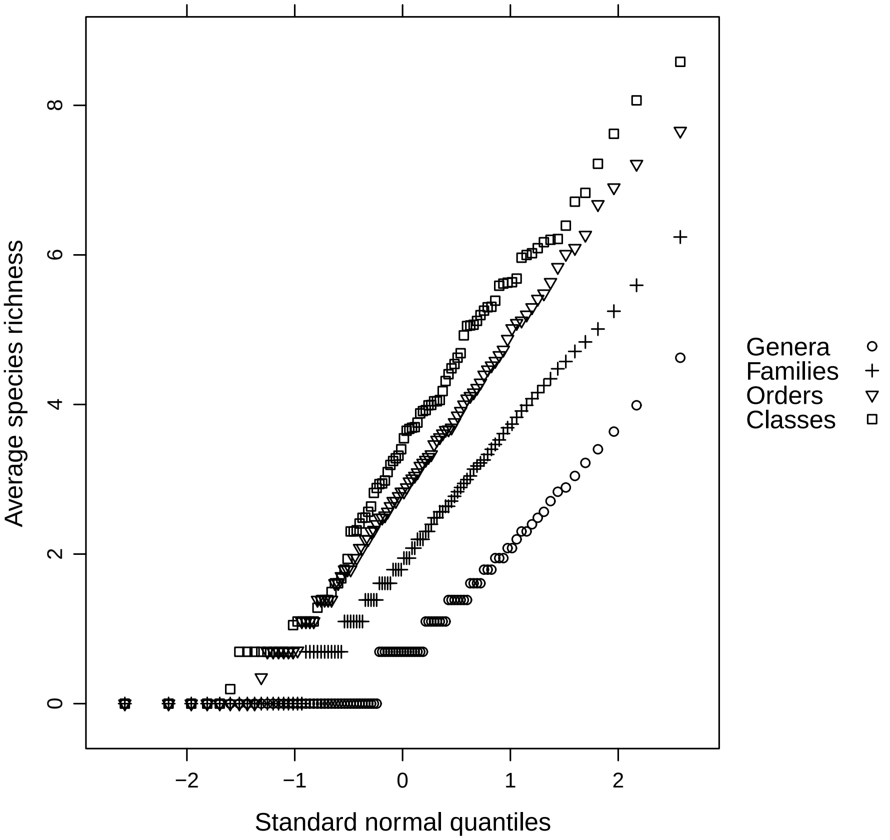


**Figure 4.5:** Normal Q-Q plot of species richness for different taxonomic levels in ITIS dataset for the first 100 quantiles

taxonomic levels (e.g class). When comparing Figures 4.5 and 4.8 a shift from positive (low species richness) to negative (high species richness) skewness at high taxonomic levels is more visible in BioMar dataset, which suggests that the scale of the data available for bootstrapping affects assumptions made on taxonomic unit invariance.

### 4.2.3 Cross-scale properties

Now, having identified the variation (taxon delimitation and unit) in biodiversity source data, it still remains to demonstrate that both properties are transferable across different scales represented by databases. Therefore, the question is to what extent taxon delimitation and unit variation properties of ITIS dataset match those of BioMar. To answer this question I will construct the  curves for both datasets and match their quantiles in two sample Q-Q plots (Figure 4.9). This figure is similar to Figures 4.5 and 4.8, with the only difference that quantiles from one dataset are plotted against quantiles from another dataset.

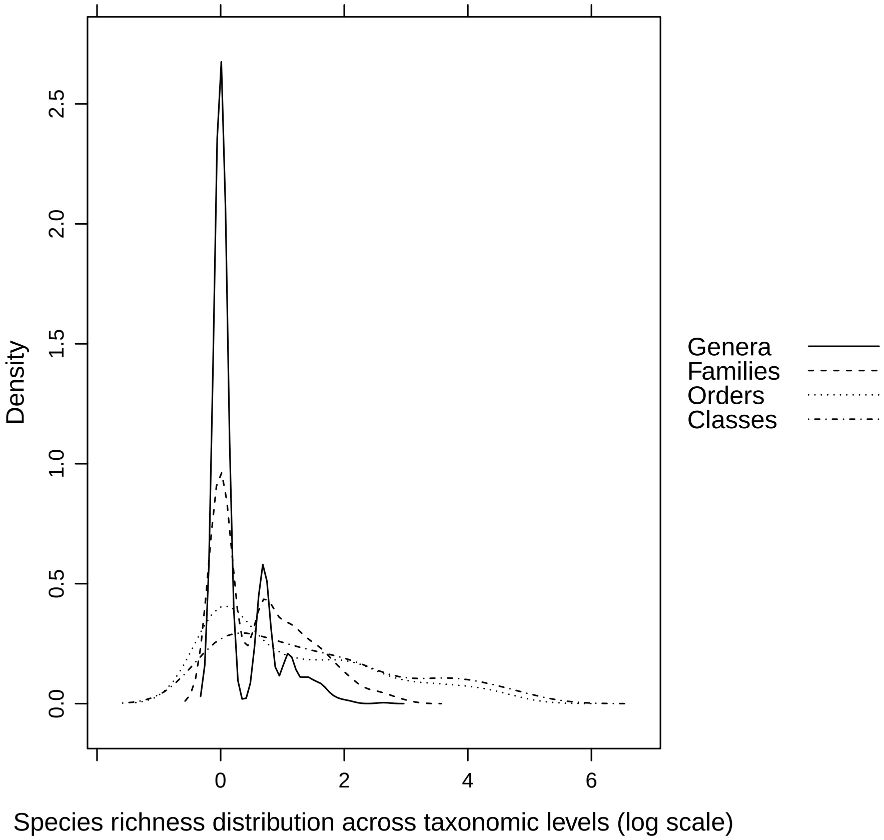


**Figure 4.6:** The grouped kernel density plot of distribution of species richness on a logarithmic scale across different taxonomic levels superposed in a single graph. The Gaussian kernel is used to compute the estimated density

A parametric curve (line) indicates that both SRD are positively skewed, which is expressed by a long right tail consisting of large values. Q-Q plots are independent of the location and the scale of the data, which make them a particularly useful tool to compare the distribution coming from different data sources, and therefore having different scale parameters.

Both SRD curves approximately lie on the same line, thus confirming that despite apparent scale differences in biodiversity data, their taxon delimitation variation seems to coincide. This gives sound grounds for using BioMar database in conjunction with ITIS to produce species lists that are suitable for bootstrapping. Despite the fact that BioMar dataset is only a snapshot of the empirical data (far from being precise and complete), this analysis shows that it can, nevertheless, be used as a sampling pool for simulation of ecological communities.

To summarise, graphical analysis of the distributional properties of the taxon delimitation variation and unit variation (such as symmetry, modality, and range), using the SRD, shows that the different biodiversity source data are statistically similar. The analysis has

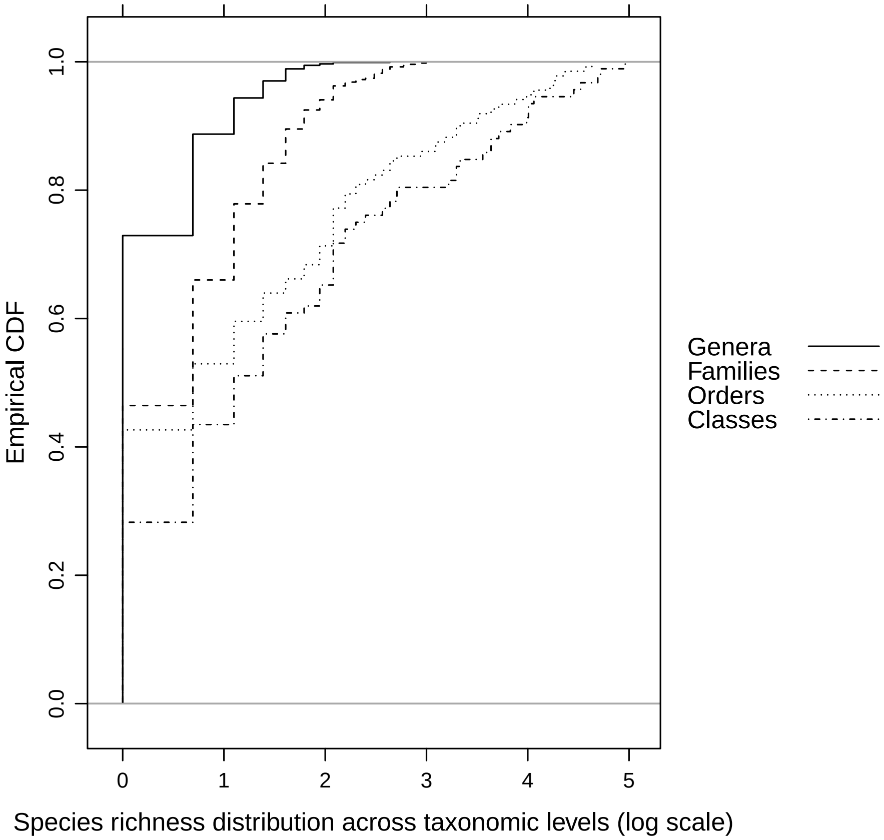


**Figure 4.7:** ECDF plot of species richness across taxonomic level in BioMar dataset

produced a break-down of statistical distribution of taxonomic richness by level among the data-sources. Based on these, it is now possible to construct simulated communities showing the same statistical properties by constraining stratified re-sampling: this will be described next.

## 4.3 Building a model for simulating communities

The communities are created by bootstrap resampling of existing biodiversity databases applying a combination of taxonomic sampling schemes (Table 4.1). In the light of Hillis' schemes, scheme 1 refers to a random sample from ITIS, scheme 2 – a random sample from BioMar, scheme 3 – sampling taxa from ITIS restricted to a "group of interest" derived from BioMar. Since both schemes, 1 and 2 are not desirable (many long branches for scheme 1 and underrepresented taxa for scheme 2), a combination of schemes was considered as the most feasible.

To specify stratified rules needed to build a model the following three steps were un-

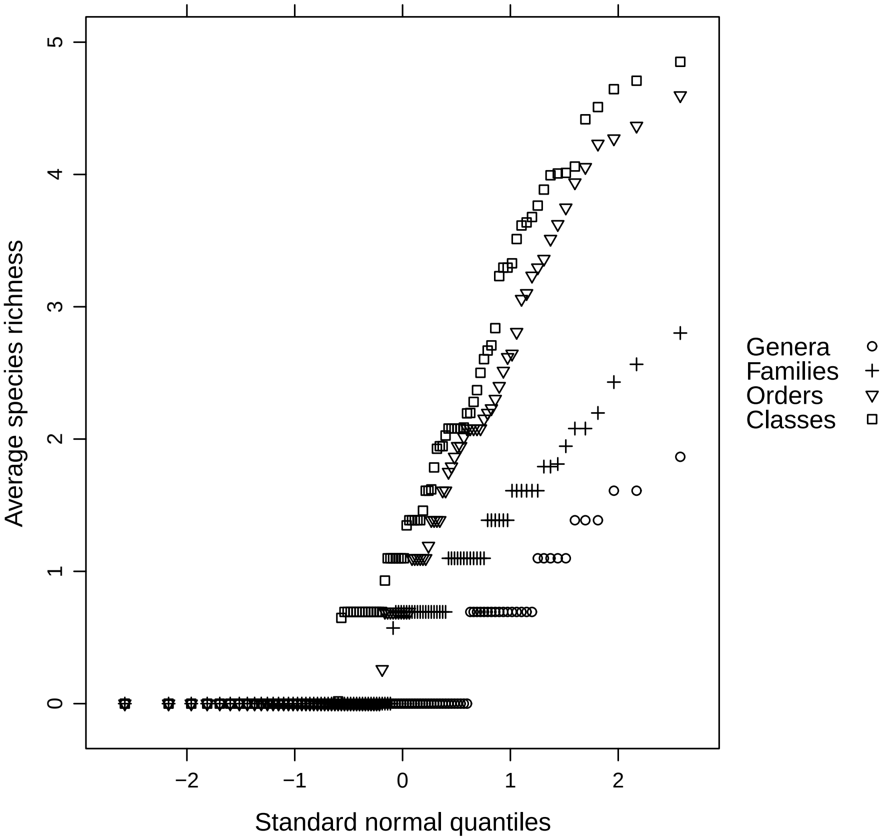


**Figure 4.8:** Normal Q-Q plot of SRD for different taxonomic levels in BioMar dataset. Quantiles from the standard normal distribution on the x-axis are plotted against quantiles of SRD in BioMar on the y-axis. Due to the large number of data points only 100 quantiles have been used for data visualisation

dertaken. First a taxonomic tree topology was generated for each community using the empirical probability distributions attributed to the unit variation – OiC, FiO, etc. Then species were selected by resampling with replacement from the ITIS database, until the topology was instantiated as a species list. Finally, species abundances were assigned by further resampling of ITIS, following a log-normal distribution. The following describes the community-simulation algorithm in more detail.

## 4.3.1 Establishing taxonomic topology

1. The complete set of taxonomic classes in the BioMar data base (total number 115) was identified as $\mathbf{C_T}$.

2. 1000 artificial communities were prepared as empty sets $\mathbf{W_i}$, each then being assigned a number, $N_c(i)$, of members of the class list $\mathbf{C_T}$, where $N_c = 1 \ldots 115$. $N_c(i)$ was

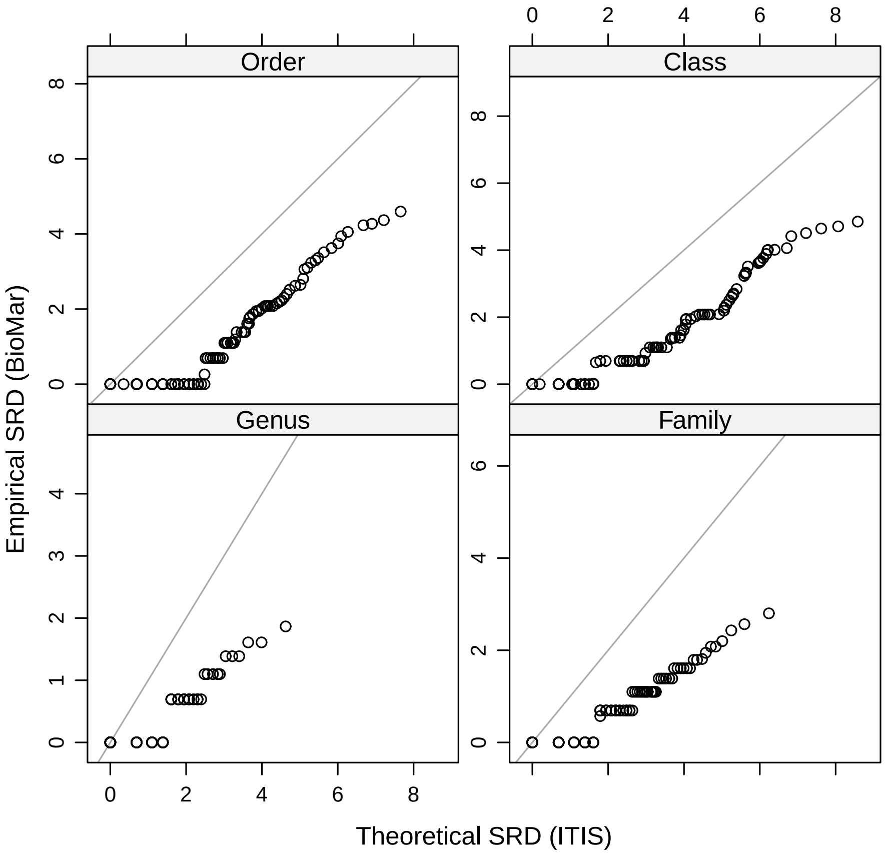


**Figure 4.9:** Two-sample Q-Q plot to compare Species Richness Distribution (SRD) in ITIS (x-axis) and BioMar (y-axis) after conditioning on taxonomic level. The Q-Q plots are non-linear and their curved pattern suggests that the quantiles in BioMar are more closely spaced than in ITIS

randomly generated for each community following a uniform distribution to give a spread of species richness among the synthetic communities.

The following steps were repeated for each community $\mathbf{W_i}$ in turn ($i = 1 \ldots 1000$).

3. For each taxonomic class $C_j(i)$, ($j = 1 \ldots N_C(i)$), of the community set of classes $\mathbf{C_W(i)}$, the number of taxonomic orders $N_O(i,j)$ in $C_j(i)$ was assigned by random sampling following the OiC (orders in class) distribution.

4. For each taxonomic order $O_k(i,j)$, ($k = 1 \ldots N_O(i,j)$), in each class $C_j(i)$, of the community set of orders $\mathbf{O_W(i)}$, the number of families $N_F(i,j,k)$ in $N_O(i,j)$ was assigned by random sampling following the FiO (families in order) distribution.

5. For each family $F_m(i,j,k)$ ($m = 1 \ldots N_F(i,j,k)$) in each order $N_O(i,j)$, of the community set of families $\mathbf{F_W(i)}$, the number of genera $N_G(i,j,k,m))$ in $F_m(i,j,k)$ was assigned by random sampling following the GiF (genera in family) distribution.

6. For each genus $G_g(i,j,k,m)$ $(g = 1 \ldots N_G(i,j,k,m))$ in each family $F_m(i,j,k)$, of the community set of genera $\mathbf{G_W}(i)$, the number of species $N_S(i,j,k,m,g)$ in $G_g(i,j,k,m)$ was assigned by random sampling following the SiG (species in genera) distribution.

This resulted in 1000 community taxonomic tree topologies, each described by a set of numbers of ramifications at each taxonomic level: $N_x(i)$ where $x = C, O, F, G, S$.

### 4.3.2 Establishing community from topology

Each community tree was then instantiated by selecting species from the ITIS database, such that they fit into the taxonomic tree to give the correct number of each taxon in the community, using the following algorithm:

For each class $C_j(i)$ of community $W(i)$, select from ITIS a set of $N_O(i,j)$ orders which are members of that class. For each of these orders $N_O(i,j)$, select from ITIS a set of $N_F(i,j,k)$ families which are members of the order. For each of these families $F_m(i,j,k)$, select from ITIS a set of $N_G(i,j,k,m)$ genus which are members of the family. For each of these genus $G_g(i,j,k,m)$, select from ITIS a set of $N_S(i,j,k,m,g)$ species which are members of the genus.

The process of stratified resampling is represented in 'plate notation' in the diagram of Figure 4.10, where the community taxonomic tree is specified. The sampling is a hierarchical process, starting with the outermost layer, it introduces the taxonomic resolution “classes”, then, working inwards, this is followed sequentially by “orders”, “families”, “genera”, and “species”. In the diagram, different shades of grey represent the number of ramifications at each taxonomic levels $N_x(i)$, with the darkest shade corresponding to the highest number of $x$. The number of taxonomic units decreases with the taxonomic resolution, thus reflecting the fact that there are more species than classes.

Resampling takes place at each resolution (stratum), where there are two possible outcomes: (a) continuation of the selection process which results in the complete taxonomic lineage for a specific organism (solid line) and (b) a termination (dashed line). To select a daughter taxon, the mother taxon need to be selected first. If in the process of sampling this mother taxon is not selected, the whole selection process on that particular lineage terminates. A single ecological community can, therefore, be described in terms of the full taxonomic topology of a pruned branch of the taxonomic tree. Each branch of this tree is correctly estimated if the lineage is complete.

To illustrate this, consider a class of annelid worms Polychaeta. By random sampling following the OiC distribution I choose family Arenicolidae within that class, then genus Arenicola and, finally, terminal taxa – species, e.g., *A. marina*. This is a complete lineage

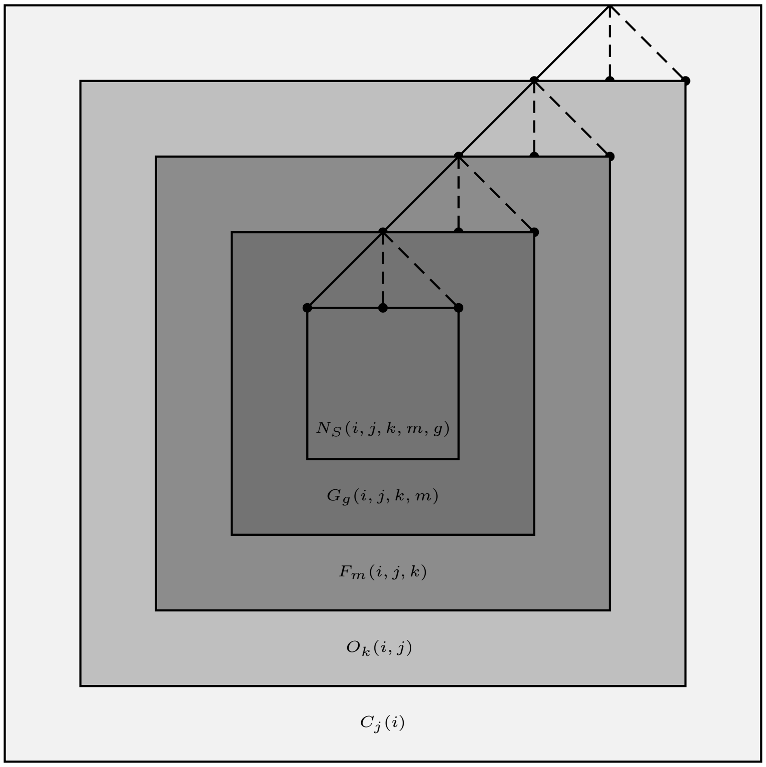


**Figure 4.10:** Plate notation: the process of stratified resampling with replacement from ITIS to instantiate the topology $\mathbf{W_i}$ as a species list $N_S(i,j,k,m,g)$

that contain representative at each taxonomic resolution; and, therefore, it gives a full taxonomic description of species *A.marina* or lugworm, a large marine worm.

The taxonomic sampling algorithm resulted in a set of 1000 communities, each comprising a species list, each differing in species richness and composition, but having distributions among taxa that match the BioMar dataset. Each species carried functional traits with it from the BIOTIC database, so that the simulated communities also had a representative distribution of functional traits. A detailed code implemented for community simulation is shown in Appendix 4.A.

### 4.3.3 Assigning species abundances

In the final stage, each species in each community was assigned a population abundance by attaching a probability weight. Species abundance distribution, being one of the few universal patterns in ecology (Morlon et al., 2009), facilitated at least a rough characterization of an ecological community.

In the BioMar dataset the abundances were recorded according to the SACFOR scale: (Ma-

**Table 4.4:** Models parameters for rank abundance distribution plot

| | par1 | par2 | par3 | Deviance | AIC | BIC |
|---|---|---|---|---|---|---|
| Null | | | | 2504.66 | 13428.37 | 13428.37 |
| Preemption | 0.00051608 | | | 842.30 | 11768.01 | 11774.44 |
| Lognormal | 0.42499 | 0.62712 | | 517.93 | 11445.64 | 11458.50 |
| Zipf | 0.0052543 | -0.44689 | | 540.94 | 11468.64 | 11481.50 |
| Mandelbrot | 0.066681 | -0.77576 | 212.56 | 182.16 | 11111.87 | 11131.16 |

rine Nature conservation review): where S was superabundant, A – abundant, C – common, F – frequent, O – occasional, R – rare, P – present. However, the major problem with this type of abundances lies in database design. Abundances were linked to individual records (one-to-many) rather than species, resulting in several records with different abundances for each species. Another (more time consuming) alternative is to record abundances in numeric form , if it was the case than it would be a straightforward issue to aggregate them using their numeric estimates. All this would make it possible to derive empirical species abundance distributions which are known to be well integrated within ecological models (McGill et al., 2007).

To amalgamate abundances across studies and sites the following heuristic algorithm was used:

(a) the modal abundance code was taken to characterize the whole group; e.g., if a species has records of its abundance as $A, A, S, C, R- > A$ the most frequent code is extracted; and

(b) if all abundance codes had equal frequency – the highest code was attributed, such as $S > A > C > F > O > R > P$ e.g., $S, A, C, F- > S$ or $S, S, A, A- > S$

The original intention was to match species abundance to trophic level through the BIOTIC database traits, but I found only 2% overlap in trophic traits between BioMar and BIOTIC species, which was insufficient to create realistic correlation. Thus, in this instance I allocated lognormal abundances to species at random following the log-normal distribution (Figure 4.11).

Analysing model parameters that were used to construct rank abundance species distributions (see Table 4.4), the lognormal model was the most effective.

## 4.4 Discussion and Conclusions

Ideally, empirical studies of real communities would supply the data to compare empirical biodiversity estimates, but the literature contains disappointingly little opportunity, cer-

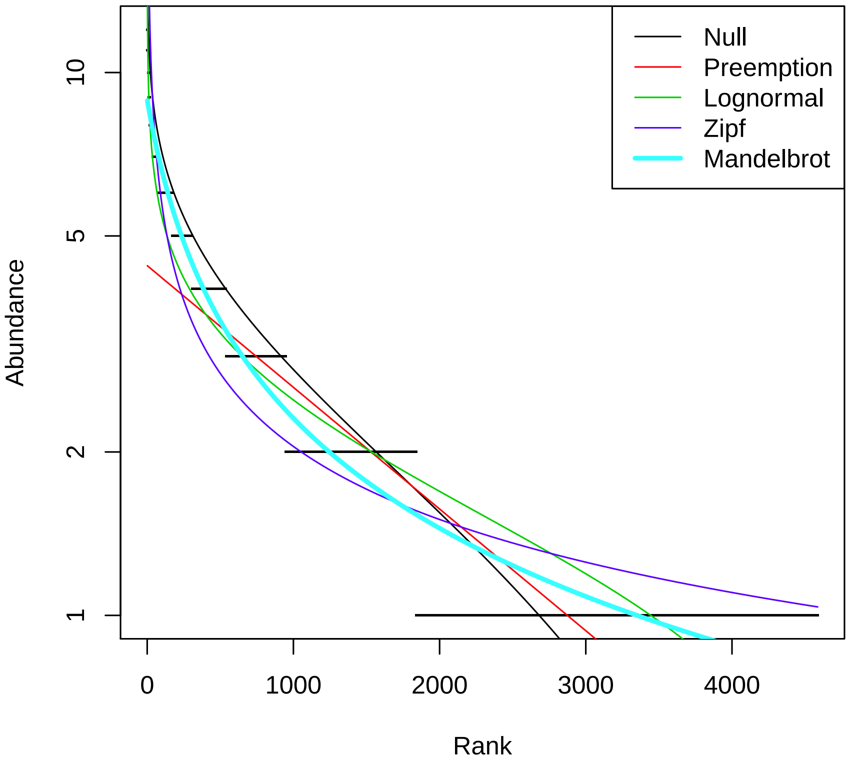


**Figure 4.11:** Rank abundance distribution plot (Whittaker plot). Different models fitted to BioMar data (Null), following Wilson (1991), show logarithmic species abundances against species rank orders

tainly not enough to perform an analysis with real data that is not affected by environmental co-variation. Lack of standardisation in methods and reporting of field studies account for some of this, but the wide range of biomes and size, and location of ecosystems, and the variety of purposes for empirical studies seems to preclude the kind of meta-analysis, commonly found in medical research, and needed if this study were to be empirical. Thus, here I have implemented a simulation algorithm to construct ecological communities from a list of species following their taxa-distributional properties. To generate artificial communities I start with a simple "null hypothesis" of random trait assignment, in which each species carries its own traits of taxonomic identity and functional role. In this model, as more species are randomly placed into a community, aggregate diversity increases.

So far, applications of simulation studies have been confined to individual levels of biodiversity typically attributed to a narrow study system (e.g., Hubalek, 2000; McGill, 2003; Pla, 2004; Chao et al., 2005; Bevilacqua et al., 2009). Here, in a departure from these studies I unite different components of biodiversity – "descriptors" represented at different "levels". Species lists have typically been used as a starting point for ecosystem simulations (e.g., Vellend, 2005; Villéger et al., 2008; Nipperess et al., 2010). While some of the studies were based on subsets of real species lists, (e.g., free-living marine nematodes, Clarke and Warwick, 1998, 2001), they all generate communities to some extent in a random manner. For instance, Villéger et al. (2008), while using uniformed statistical distribution to generate species and their abundances did not establish the reference to anything real.

Using these simulation studies for guidance, I have built artificial communities with taxonomic structures and distributions of species traits that statistically match real communities. Taxonomic structures were derived from distributional properties of the taxonomic units of near-shore temperate marine ecosystems that arise at different levels. However, taking a step further from random sampling, a more elaborate taxonomic sampling schemes discussed by Hillis (1998) were introduced into the sampling algorithm. For this purpose, once constructed, a taxonomic tree topology was then instantiated as species lists through a taxonomic sampling algorithm. In this way, model communities were composed such that they reflect the higher-taxon composition of real communities according to the BioMar database. The resulting correlation among species taxa causes a narrowing of taxonomic diversity both within and among communities. In line with Villéger et al. (2008), I used uniformed statistical distributions to generate species lists by their resampling. However, to ensure the plausibility of community composition, I obtained the set of empirical taxonomic probability distributions across all taxonomic levels.

There are a few unresolved points in the algorithm. For example, insufficient data were available to construct empirically-based species abundances, which are known to show correlation with trophic function (e.g., the exponential relation to trophic level in Sheldon and Parsons, 1967; Sheldon et al., 1972). It was unfortunately not possible to obtain sufficient data to incorporate this relationship into the simulated communities. Abundances were assigned by a second round of resampling which followed log-normal distribution to generate species abundances matching general empirical expectations. My original intention was to match species abundances to trophic level through the BIOTIC database traits, but having found only 2% overlap in trophic traits between BioMar and BIOTIC species it was clearly insufficient to have realistic correlation.

Thus, in these simulations, community structure characteristics were based on a log-normal abundance distribution, closely matching empirical data (see Figure 4.11), but with no correlation to functional traits. Additionally, the weak taxonomic invariance within the taxonomic units (Storch and Sizling, 2008) was implied by the use of empirical taxonomic probability distributions across all taxonomic units of equivalent level. In line with Storch and Sizling (2008), patterns in taxon delimitation variation had similar shapes at each taxonomic resolution.

The log-normal model parameters that were derived from rank abundance distribution justify the efficacy of the chosen distribution. As a result, simulated communities exhibit log-left-skewed species abundance distribution which implies many rare and a few common species – a phenomenon typically found in real communities. As it has been shown that sampling algorithm affects taxa-abundance distributions (McGill et al., 2007), this can be partially attributed to the specificity of the Hillis's sampling schemes.

Further comparison of the taxa-distributional properties in BioMar and ITIS, there is an evident shift from positive (low species richness) to negative (high species richness) skewness at high taxonomic levels. This is even more distinctive in BioMar dataset, which can be explained by a rapid acceleration of accumulation rate of species when the number of taxonomic units (e.g., classes) gets relatively small. In practical terms, it would imply that there are fewer species rich classes in BioMar comparing to ITIS. This finding clearly discards the alternative of using BioMar as a single biodiversity source for establishing a taxonomic tree topology and resampling in order to instantiate topology as a species list. The size of BioMar is simply not large enough and, therefore, it should be used in conjunction with ITIS.

To summarise, in this chapter I describe an algorithm used to generate synthetic ecological communities from species lists with the properties matching real ecological communities. To produce a taxonomic tree topology which could be then transferred into a species lists, the taxonomic structure of real communities was applied via a taxonomic sampling algorithm. Resulting synthetic communities are a suitable data source for examining biodiversity indicators to elicit patterns of variability among them.

## Summary

1. Artificial ecological communities were constructed from a list of species of coastal marine communities following their taxa-distributional properties and distributions of species traits;

2. Community structure characteristics were based on a log-normal abundance distribution, closely matching empirical data;

3. The taxa-distributional properties in BioMar and ITIS, demonstrated a shift from positive (low species richness) to negative (high species richness) skewness at high taxonomic levels;

4. A rapid acceleration of accumulation rate of species with the relatively small number of taxonomic units (e.g., classes) in BioMar database implied that it contained fewer species rich classes, thus discarding the alternative of using BioMar as a single biodiversity source; and

5. Knowledge of the empirical distributional properties of taxa among the categories of diversity allowed for realistic relations to be built into the test data.

# Appendices

## Appendix 4.A Simulation of hypothetical community

### Preparation of the file (SQL)

```
--ITIS database: a flat file of all taxonomic levels:

mysql> CREATE TABLE flatfile
SELECT O1.name AS tclass, O2.name AS torder, O4.name AS tfamily, O5.name AS tgenus, O6.name
     AS tspecies
FROM taxonomic_units1 AS O1
LEFT OUTER JOIN
taxonomic_units1 AS O2
ON O1.tsn = O2.parent_tsn
LEFT OUTER JOIN
taxonomic_units1 AS O3
ON O2.tsn = O3.parent_tsn
LEFT OUTER JOIN
taxonomic_units1 AS O4
ON O3.tsn = O4.parent_tsn
LEFT OUTER JOIN
taxonomic_units1 AS O5
ON O4.tsn = O5.parent_tsn
LEFT OUTER JOIN
taxonomic_units1 AS O6
ON O5.tsn = O6.parent_tsn;

--join ITIS and BioMar
mysql> CREATE TABLE ITIS.marinespecies SELECT tsn,
CONCAT(unit_name1,unit_name2) AS speciesmarine, Abundance, nAbnd
FROM ITIS.taxonomic_units, marine.SpeciesAbnd WHERE rank_id=220
AND (marine.SpeciesAbnd.GenericName=ITIS.taxonomic_units.unit_name1
AND marine.SpeciesAbnd.SpecificName=ITIS.taxonomic_units.unit_name2);

--this table is now ready to be used in simulation
mysql> CREATE TABLE flatfilemarine
SELECT flatfile.tclass, flatfile.torder, flatfile.tfamily, flatfile.tgenus,
flatfile.tspecies
FROM flatfile, marinespecies
WHERE tspecies=speciesmarine;
```

### Simulation (R)

```
attach(mydata2)
com<-list()
for (i in 1:1000){
tcla<-sample(tclass,sample(1:length(unique(tclass))),replace=T)
tord<-sample((torder[tclass%in%tcla]),sample(1:length(unique(torder[tclass%in%tcla]))),
    replace=T)
tfam<-sample((tfamily[torder%in%tord]),sample(1:length(unique(tfamily[torder%in%tord]))),
    replace=T)
tgen<-sample((tgenus[tfamily%in%tfam]),sample(1:length(unique(tgenus[tfamily%in%tfam]))),
    replace=T)
tspe<-sample((tspecies[tgenus%in%tgen]),sample(1:length(unique(tspecies[tgenus%in%tgen]))),
    replace=T)
com[[i]]<-list(Lclasses=list(classes=tcla,nclasses=length(tcla),uclasses=length(unique(tcla
    )),aclasses=table(factor(tcla))),
Lorders=list(orders=tord,norders=length(tord),uorders=length(unique(tord)),aorder=table(
    factor(tord))),
Lfamilies=list(families=tfam,nfamilies=length(tfam),ufamilies=length(unique(tfam)),
    afamilies=table(factor(tfam))),
Lgenera=list(genera=tgen,ngenera=length(tgen),ugenera=length(unique(tgen)),agenera=table(
    factor(tgen))),
Lspecies=list(species=tspe,nspecies=length(tspe),uspecies=length(unique(tspe)),aspecies=
    table(factor(tspe))))}
```

If `replace = FALSE`, function `sample` generates a random permutation of taxonomic elements. If `replace=TRUE`, accumulated abundance can be used for testing the performance of measures and indices of biodiversity.

# Chapter 5

# Biodiversity metric: multum in parvo

## 5.1 Introduction

This chapter is about finding a unified metric of biodiversity, which I seek to derive from a multidimensional set of measures of biodiversity. Two questions motivated the work presented in this chapter: firstly, can an optimal measure of biodiversity be constructed, and secondly, how closely is it approximated by the most commonly used measure – species richness? By optimal I mean capturing the maximum information about biodiversity in a compact form – multum in parvo. I seek, therefore, a measure with the maximum information density.

There is already plenty to choose from. The rapidly growing biodiversity literature offers a substantial "lexicon zoo" (Marcot, 2007) of biodiversity indices, leading some commentators to refer to a confusion of meaning (Hamilton, 2005) and to the presence of ambiguities (Weesie and van Andel, 2003). Biodiversity is often taken as a constellation of meanings and while some authors suggest that it can never be captured by a single number (Purvis and Hector, 2000; Mayer, 2006; Failing and Gregory, 2003), others attempt to find such an index (e.g., Mendes et al., 2008; Certain et al., 2011).

Using unifying aspects of biodiversity indices Mendes et al. (2008) suggested that a generalised entropy (also known as Tsallis entropy) adopted after Patil and Taillie (1982) has the potential to capture multiple aspects of biodiversity. While going beyond individual indices (such as Simpson or Shannon-Wiener) and considering aspects related to evenness, rarity, and dominance (Wilsey et al., 2005), this study has apparent limitation. It is still restricted to one facet of biodiversity outlined by Noss (1990) – structural, thus impeding the potential to describe different aspects of biodiversity.

The results of this chapter were presented at ICES Annual Science Conference, Nantes, France, September 20-24, 2010

Multiple facets of biodiversity, represented by this diversity of meanings and applications encompass a diversity of measures. Facets of biodiversity that are commonly considered include genetic and phenotypic variance, species numbers, ecosystem structural properties, and patterns of functional heterogeneity. I organise the broad spectrum of diversity metrics into three conceptually distinctive groups of measures. These include community structure and composition, taxonomic and functional diversity measures calculated for a set of simulated communities.

As a reflection of the multiple nature of the concept, May (1994) concluded that "biological diversity can be quantified in many different ways, at many different levels". In Chapter 2 I formalised this idea by recognising biodiversity as "difference" among the components of a biological system defined on a set of axes coinciding with the set of "descriptors" and "levels" over which these descriptors may measure (see Table 2.1). Recalling, "measures" (defined as a scalar combination of one descriptor at one level, see Definition 2.5) are the fundamental metrics from which all biodiversity indices are composed, the optimal estimator of biodiversity which I am seeking now must be some combination of measures. This combination of measures is then allowed to optimise for information content.

The present proliferation of metrics calls now for rationalisation and synthesis to identify which features of biodiversity are mathematically independent. Thereby, the irreducible (optimum) set of metrics which must be included to encompass total biodiversity, can be found. Implied in that goal is the identification of redundant metrics which are so mutually correlated that any one of them may be taken to approximate the others. Given the multidimensional nature of biodiversity, I attempt a reduction to the minimal set of metrics needed to describe biodiversity (often by default taken to be species richness) using a set of simulated communities.

Using this set of known total differences in biodiversity, the task amounts to finding those indices which maximise the measured differences. If diversity exists among $d$ characteristics, then the index conveying all the diversity with greatest efficiency will be formed from a set of $d$ orthogonal measures. Rank-ordering orthogonal axes of variation enables information density to be maximised by removing those axes with less than a statistically justified information content. This procedure is a description of Principal Component Analysis (PCA), which therefore provides the basis of my analysis. I apply it to a population of artificial communities generated by (bootstrap-like) resampling of real benthic community data. In other words, I assess index sensitivity to diversity within communities by measuring their sensitivity among communities.

The search for the maximum information density involves defining a necessary and sufficient (irreducible) set of metrics which best approximate total biodiversity. This practically amounts to an ordination among measures, constructing principal axes of variation and interpreting them in biological terms. A single measure estimate could then be calculated as a distance metric in the reduced space of principal axes. Comparison between species richness and this composite measure gives an indication of the proportion of unexplained variation in biodiversity metric space created by using the most common metric of biodiversity – species richness. Even though, it is commonly acknowledged that biodiversity has a broader meaning than species diversity (see, e.g., Krebs, 1998), the question I want to ask here is how much broader? More specifically, here I am attempting to quantify the proportion of missing variation in total biodiversity when I use species richness as a single surrogate.

The chapter is organised as follows. Indices of biodiversity and the procedure of standard ordination practice will be addressed in Section 5.2. Then, following this practice, potentially correlated indices of biodiversity will be transformed into their orthogonal linear combinations capitalising on any patterns and redundancies that emerge in the dimension reduction process in Section 5.3. This is followed by a discussion and a conclusion in Section 5.4.

## 5.2 Methods

Simulated biodiversity communities (Chapter 4) were used as the source of data for all the biodiversity metrics calculated here. For this, variables that were needed to calculate measures and indices were obtained either directly from simulated data (e.g., species richness and species abundance) or from external sources (e.g., functional traits). Reduction of a set of biodiversity metrics to a minimum set enables me to test the overall variability in biodiversity estimates and to contrast it against variability accounted by species richness alone. Three main questions define biodiversity: *(1) what things are in there; (2) how different are these things; and (3) how different are the things they do?* To address these questions, I need to form a complete set of distinct components of biodiversity.

### 5.2.1 Groups of measures and indices

Forming an optimum set is considered as a time consuming process (Grantham et al., 2009) and a large number of metrics is needed to ensure broad coverage of biodiversity aspects in it (Certain et al., 2011). Here, following Noss's classification I choose three conceptual groups of measures and indices to represent a spectrum of biodiversity aspects ranging from structural and compositional to taxonomic and functional aspects. For convenience, I gather all these metrics under the common groups A, B, and C, which relate to community structure, taxonomy, and function respectively. Different groups, related to both patterns and processes, allow me to account for the maximum amount of variability in biodiversity estimate across simulated communities by exploring the independence of these groups from one another and eliminating any redundancies. These three groups are intimately linked through shared common elements – species and calculated for communities that were simulated from species lists. In Table 5.1 I show components of each of the groups, give their references and define abbreviations.

**Table 5.1:** Groups of indices and measures of biodiversity

| Measure | Reference |
|---|---|
| Group A: Structure and Composition | |
| Simpson Index (SIMP) | |
| Pielou Index (PIEL) | |
| Jaccard Index (JACC) | |
| Sorensen Index (SORE) | |
| Chao-Jaccard Index (CHJA) | |
| Chao-Sorensen Index (CHSO) | |
| Shannon Index (SHAN) | |
| Turnover Index (TURN) | |
| Abundance (ABUN) | |
| Richness (RICH) | |
| Group B: Taxonomic diversity | |
| Taxonomic Diversity (see also $\Delta$) (DELT) | Clarke and Warwick 1998 |
| Taxonomic Distinctness (see also $\Delta^*$) (DSTR) | Clarke and Warwick 1998 |
| Variation in Taxonomic Distinctness (see also $\Lambda^+$) (LPLU) | Clarke and Warwick 2001 |
| Taxonomic Diversity (for presence, absence)(see also $\Delta^+$) (DPLU) | Clarke and Warwick 1998 |
| Taxonomic Diversity (accounting for species richness) (SPLU) | Clarke and Warwick 2001 |
| Group C: Functional diversity | |
| Functional Richness (FRIC) | |
| Functional Evenness (FEVE) | Villéger et al. 2008 |
| Functional Divergence (FDIV) | Mason et al. 2003; Villéger et al. 2008 |
| Functional Dispersion (FDIS) | Anderson 2006; Laliberté and Legendre 2010 |
| Quadratic Entropy (RAOQ) | Rao 1982; Botta-Dukát and Wilson 2005 |

### Group A: Community structure and composition

This first group conceptually relates to the first question that defines biodiversity – *what things are in there?* While structural diversity, as Noss defines it, refers to patterns of a system, compositional diversity relates to the identity and variety of species. Therefore, most measures and indices within this group are concerned with the community patterns formed by the evenness in relative abundance among species (Balvanera et al., 2005). This group is grounded on species-based estimates emphasising closely related concepts of species diversity – such as richness, heterogeneity, and equitability (evenness) (Peet, 1974).

Briefly addressing each of the concepts, I shall start off with species richness (RICH). Species richness is the most frequently applied measure of biodiversity, since there is no need to derive complex indices to express it, it is a widely understood measurable parameter. The popularity of species richness is more related to data availability (Fleishman et al., 2006), but the lack of universal agreement as to the definition of species leading to "mutually incompatible ways" (Mallet, 2007) is the major disadvantage.

Taking a conceptual perspective and further examining the utility and limitations of species richness, Fleishman et al. (2006) conclude that species richness by itself has a very limited information content. For example, it does not convey any information on rarity, endemism, and function, and does not distinguish between native and non-native species.

Following Good (1953), species richness combined with evenness forms the concept of heterogeneity which is as a measure of community organisation applied to species diversity. Finally, evenness of a biological community can be defined as "the degree to which abundances are divided equitably among the species present" (Routledge, 1983). Here it is worth noting, that abundance (ABUN) being a variable property of the community, is used to estimate diversity. When abundance fluctuates, so do species diversity indices that are derived from it.

Community structure and group composition also encompass other indices aiming to capture species diversity – such as Simpson (SIMP), Shannon (SHAN), Sorensen (SORE), Pielou (PIEL). Similarly to species richness, all these indices rely on assumption of taxonomic equality. Being a "traditional way" of quantifying biological diversity, these indices are well documented in the literature (see, e.g., Magurran, 2004, for a review of different methods), yet the understanding of which index of biodiversity represents maximum information content is still wanting.

Here, using Peet's conceptual framework for species diversity (richness, heterogeneity, and equitability), I have selected indices in such a way that each aspect of it was covered. All indices were calculated on a bootstrap resampled data using R-script in Appendix 5.A.

### Group B: Taxonomic diversity and distinctness indices

This group refers to the second question that defines biodiversity – *how different are the things?* This difference can come in several forms such as taxonomic, morphological or genetic (see, e.g., Faith, 1994). Indeed, community composed of species that are distantly related intuitively is more diverse, than community composed of similar species (Desrochers and Anand, 2004). Taxonomic relatedness indices based on a tree topology are believed to reflect this dissimilarity (Warwick and Clarke, 1995; Clarke and Warwick, 1998).

In contrast to closely related phylogenetic diversity measures by Faith (1994), measures of taxonomic diversity differ from phylogenetic measures by the fact that branch length is typically not included. Still, considered to be a fairly accurate representation of underlying phylogeny (see, e.g., Crozier, 1997), distances based on Linnaean taxonomy overcome an assumption of taxonomic equivalence between species. This is especially convenient since general application of phylogenetic methods is impeded by the lack of complete phylogenies. If taxonomic arrangement mirrors the topology of the evolutionary tree, then the sequences of the genes constitute the information content (Crozier, 1997). The practical link between taxonomy and phylogeny has been further strengthened by Crozier et al. (2005), who proposed an algorithm of inferring surrogate phylogenies from systematic nomenclature.

Based on Rao's idea of incorporating the difference between species chosen at random from community, Clarke and Warwick (1998) have proposed indices of taxonomic diversity and taxonomic distinctness (DELT – Δ and DSTR – Δ* respectively).

Given two biological entities $i$ and $j$ chosen at random from community and belonging to different taxonomic groups taxonomic diversity between them can be calculated as following:

$$\Delta = \frac{\sum\sum_{i<j} \omega_{ij} x_i x_j}{n(n-1)/2} \tag{5.1}$$

$$\Delta^* = \frac{\sum\sum_{i<j} \omega_{ij} x_i x_j}{\sum\sum_{i<j} x_i x_j} \tag{5.2}$$

where $x_i (i = 1, \ldots, s)$ is abundance (or presence/absence), n is the total number of individuals in the community and $\omega_{ij}$ is the matrix containing the species pairwise distances. These distances are taxonomic differences calculated as a distinctness weight of the path length linking any two entities. In both formulas, distance, established from taxonomic hierarchy, gives species that belong to the same genus 1, to the same family 2, etc.

In Formula 5.1 the taxonomic hierarchy is ignored when $\omega_{ij}$=1, thus Δ reduces to standard diversity indices shown in Group A. Index $\Delta^*$ (Formula 5.2) is a function of taxonomic relatedness of individuals which is invariant to a scale change in $x$. Distances are scaled to reflect the decrease in taxon richness at each level.

Formulas 5.1 and 5.2 can be used to derive several other indices such as variation in taxonomic distinctness (LPLU – $\Lambda^+$) or taxonomic diversity for presence/absence data (DPLU – $\Delta^+$).

An index of variation in taxonomic distinctness $\Lambda^+$ can be calculated as following:

$$\Lambda^+ = \frac{\sum\sum_{i<j}\omega_{ij}^2}{n(n-1)/2} - (\Delta^+)^2 \tag{5.3}$$

While Clarke and Warwick determine distances taxonomically, $\omega_{ij}$, can be any distance structure among entities. Izsak and Papp (2000), for example, calculated the species pairwise distances $\omega_{ij}$ from the distance matrix on feeding behaviour and the number of nodes that separate each pair of species. Alternatively Euclidean distance was proposed to be used to calculate distance matrix using species traits (Champely and Chessel, 2002), which links this approach to functional diversity indices I shall be addressing in Group C.

To calculate above mentioned indices, first I recreated taxonomic hierarchy for each of the simulated ecological communities. Then from this established hierarchy, I derived the pairwise taxonomic distances among species (using `taxa2dist{vegan}` Oksanen et al., 2010) and finally, implemented the taxonomic distances to calculate the taxonomic diversity indices (`taxondive{vegan}` Oksanen et al., 2010). Since some of the communities contained many rare species, to simplify computation abundances of all represented species were assigned to 1, missing species – 0. This algorithm is outlined in Appendix 5.B.

### Group C: Functional diversity indices

In contrast to groups A and B which are mostly related to description of community patterns, indices of this group are mostly concerned with community processes. This group of indices is therefore especially interesting in the recent trend of determining ecosystem processes and valuing ecosystem services (Daily and Dasgupta, 2007; Virginia and Wall, 2007). I address here the last question that defines biodiversity: *'How different are the things they do?'* by introducing a set of functional diversity indices.

Functional diversity can be defined as a value and range of functional traits (Diaz and Cabido, 2001; Tilman, 2007; Schleuter et al., 2010). Term trait requires further clarification. Functional trait is defined as the "characteristics of an organism that are considered relevant to its response to the environment and/or its effects on ecosystem functioning" (Diaz and Cabido, 2001). These traits, therefore, are components of functional diversity indices, which measure the distribution and the range of what organisms do in communities (Diaz and Cabido, 2001). Any pair of communities with a similar amount of species may demonstrate different amounts of diversity depending on how similar or dissimilar their functional traits.

There is a number of ways in which functional diversity can be estimated. It can be done either from species dendrogram as in Petchey and Gaston (2002) or species traits matrix as in Villéger et al. (2008). Since dendrogram-based functional diversity is calculated in a manner similar to taxonomic diversity, and, it is, therefore, associated with it. Here I am deliberately choosing for the latter option – measuring functional diversity from the species traits matrix.

I measure functional diversity from multiple traits derived from the biological traits information catalogue (BIOTIC, 2010). This dataset comprises about 400 species across 40 traits ($nt = 40$), and it is used in conjunction with Irish benthic marine database and the taxonomically structured species database described in Chapter 4. I considered functional traits such as food type, maximum size, habitat, biogeographic range or biozone. These traits were coded as continuous, ordinal, nominal, or binary traits. Unfortunately, the number of species for which functional traits were available in information catalogue was much lower than the number of species in dataset used to simulate ecological communities. This introduced communities that have very few species for which functional trait information was available. Due to missing values, none of the techniques of traits estimation (such as from average, by regression or interpolation from positive autocorrelation) outlined by Legendre and Legendre (1998) seemed to be appropriate.

To circumvent this problem individual species traits recorded in the information catalogue were generalized to the high taxonomic level (i.e., classes level $L^C$). By doing this, I have made an assumption that most of the species were representative to the classes they fall into in terms of their functional traits. So for instance, if I consider diverse class the Gastropoda that includes more than 35000 species, most of them still can be characterised as a small to medium sized animals feeding on detritus and algae regardless of the wide range of feeding strategies they employ.

Taking a multiple-trait approach and increasing taxonomic resolution from species level $L^S$ to $L^C$, an overlap between simulated and functional trait data increases. For each of the communities species-by-traits ($L^S \times t$) matrix has been generalized to a class-by-traits matrix ($L^C \times t$), standardised, and columns (traits) of interest were extracted. Using $L^C \times t$ matrix, I calculated the distance between each pair of classes based on their functional traits. The functional-niche space was defined by the $nt$ dimensional space, where each of the axes $n$ corresponds to a trait (t). Since some of the traits are categorical and others are quantitative, to consider them simultaneously, distance functions for mixed data need to be used. Gower (1971) proposed an appropriate measure – called Gower distance, which allows the simultaneous appearance of presence/absence, unordered categorical and quantitative variables.

Using the $L^C \times t$ matrix I derived a range of functional diversity indices. These include functional richness (FRIC), functional evenness (FEVE), functional divergence (FDIV), functional dispersion (FDIS), and Rao Quadratic Entropy (RAOQ). I shall address each of them below.

Functional richness describes volume of the functional space occupied by the community, functional evenness regularity of the distribution of abundance in this volume calculated from a minimum spanning tree which links all taxonomic groups in the multidimensional functional space, and functional divergence – divergence in the distribution of abundance in this volume (Villéger et al., 2008).

To measure the functional space occupied by a community, Cornwell et al. (2006) proposed to use the convex hull volume, where FRIC is the minimum convex hull which includes all the classes. First the most extreme points were determined, then they were linked to form a convex hull and, finally, the volume was calculated. One of the important conditions for that is that number of taxonomic groups must be higher than the number of traits i.e., $nL^C > nt$. FEVE has been defined as the evenness of abundance distribution in a functional trait space (Mason et al., 2005). To transform taxonomic distribution in a multidimensional space (many traits) to one dimension Villéger et al. (2008) have suggested to use the concept of the minimum spanning tree and outlined a formula to do so (rewritten):

$$FEVE = \frac{\sum_{L^C=1}^{nL^C-1} (PEW_{L^C}, \frac{1}{nL^C-1}) - \frac{1}{nL^C-1}}{1 - \frac{1}{nL^C-1}} \tag{5.4}$$

where PEW is a partial weighted evenness, values of which vary across branches of the taxonomic tree $T$ and it is defined as:

$$PWE_{L^C} = \frac{EW_{L^C}}{\sum_{L^C=1}^{nL^C-1} EW_{L^C}} \tag{5.5}$$

and $EW_{L^C}$ is defined as $\frac{dist(i,j)}{\omega i + \omega j}$ with i and j being a pair of taxonomic units and $\omega$ their weight. FEVE defined in this way is constrained between 0 and 1, and as it is claimed by the authors it is not biased by species richness. FEVE will decrease if abundance is less evenly distributed or when functional distances between taxonomic units are less regular.

Functional divergence (FDIV) relates to how abundance is distributed within the volume of functional trait space occupied by taxonomic units. Villéger et al. (2008) suggest a novel approach of doing it. First, using convex hull approach the center of gravity of taxonomic units forming the vertices has been calculated, then the mean Euclidean distance to this centre is calculated ($\overline{dG}$). Using the sum of abundance-weighted deviances ($\Delta d$), FDIV may be calculated as following:

$$FDIV = \frac{\Delta d + \overline{dG}}{\Delta|d| + \overline{dG}} \tag{5.6}$$ [1]

This index ranges between 0 and 1, and it approaches 0 when highly abundant taxonomic units are very close to the centre of gravity.

Two indices that were calculated – FDIS (Laliberté and Legendre, 2010) and RAOQ (Botta-Dukát and Wilson, 2005) are closely related to each other. FDIS can be defined as the mean distance in multidimensional trait space of individual species (or other taxonomic units) to the centroid of all species. It has been suggested to use multivariate dispersion (Anderson, 2006) as a multidimensional index of functional dispersion and it can be used on any distance or dissimilarity measure.

Rao's quadratic entropy (Rao, 1982) incorporates both the relative abundances of species $p$ (or other taxonomic units) and a measure of the pairwise functional differences between them – $d_{ij}$:

$$RAOQ = \sum_{i=1}^{nL^C-1} \sum_{j=i+1}^{nL^C} d_{ij} p_i p_i \tag{5.7}$$

Villéger et al. (2008) claim that all functional diversity indices described above satisfy a list of *a priori* criteria advocated by Mason et al. (2003), to investigate these properties as well as their sensitivity they were included in principal components analysis. A detailed procedure for calculating indices is shown in Appendix 5.C. Species-by-trait matrix was calculated using R function (`dbFD{FD}` Laliberté and Shipley, 2010).

To summarize, using Noss classification I have calculated three conceptually distinct groups of measures and indices of biodiversity. These were calculated across a range of simulated ecological communities with an intention to form an optimum set of measures and indices of biodiversity, that describe multiple facets of biodiversity. Having done this, the relationship between the components of each group can be tested using multivariate techniques.

### 5.2.2 Dimension reduction

Reduction to a minimum set of biodiversity metrics allows testing the overall variability of the components of that set, each representing a facet of biodiversity. To form an optimum set, any patterns or redundancies that emerge in the dimension reduction process need to be capitalised and accounted for. Dimension reduction was achieved through principal component analysis. Potentially correlated biodiversity indicators were transformed into their orthogonal linear combinations following standard ordination practice.

[1] For details on how to compute this index see Villéger et al. (2008)

To do so I used the dataset generated in Chapter 4. This dataset comprises of 1000 rows, each corresponding to a simulated ecological community and columns – each occupied by a single measure of biodiversity. Principal components were based on a standardised correlation matrix and were calculated using a singular value decomposition of the centered data matrix (*prcomp*{*stats*}).

First, dimension reduction was achieved by eliminating all principal component axes contributing less than threshold variance. This variance corresponded to a three dimensional space indicated by a screeplot using broken stick model (Legendre and Legendre, 1998). Then, representatives from each dimension were selected and Manhattan distance between them was calculated to derive a single value indicator of biodiversity. This single value indicator corresponding to the scalar distance from the origin in the largest three principal components was plotted against species richness to show the contribution of species richness as a single measure of biodiversity.

Finally, in conjunction with PCA, I account for the relatedness between closely similar components within and between each group of measures via hierarchical clustering. This allows to further separate the contribution of individual components towards the overall variability in biodiversity metric space. I conducted clustering on the rotation matrix of variables using Manhattan distance.

## 5.3 Modelling results

I used multivariate analysis to investigate properties of indices and measures of biodiversity such as their correlation and membership over a range of simulated ecological communities. These include correlation matrix, principal component analysis and hierarchical clustering. Hierarchic organisation of simulated data tests for changes in indices and measures of biodiversity occurring both within and between taxonomic levels. Analysis, therefore, was not limited to species level: it was also possible where needed to test different combination of D and L.

### 5.3.1 Correlation matrix

Spearman correlation coefficients for pairs of biodiversity indices (Table 5.2) revealed a perfect (both positive and negative) relationship within several structure and composition indices (e.g., Simpson (SIMP), Pielou (PIEL), Shannon (SHAN), Jaccard (JACC), Sorensen (SORE), and Turnover (TURN)). Such correlations imply that these indices can be mathematically derived from one another, and so are mutually redundant.

Taxonomic diversity indices were weakly correlated ($\rho$ lies in the range $(-0.054;\ 0.072)$

**Table 5.2:** Correlation between measures of biodiversity calculated on 1000 simulated ecological communities. Three groups of measures/indicators can be distinguished: (A) community structure and composition – Richness, Pielou, Jaccard, Sorensen, Chao-Jaccard, Chao-Sorensen, Shannon, Turnover, and Abundance; (B) taxonomic diversity – D, Dstar, Lambda, Dplus, SDPlus; and (C) functional diversity – FEve, FDiv, FDis, RaoQ

| | SIMP | PIEL | JACC | SORE | CHJA | CHSO | SHAN | TURN | ABUN | RICH | DELT | DSTR | LPLU | DPLU | SPLU | FEVE | FDIV | FDIS | RAOQ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SIMP | 1.000 | | | | | | | | | | | | | | | | | | |
| PIEL | 1.000 | 1.000 | | | | | | | | | | | | | | | | | |
| JACC | 0.266 | 0.266 | 1.000 | | | | | | | | | | | | | | | | |
| SORE | 0.266 | 0.266 | 1.000 | 1.000 | | | | | | | | | | | | | | | |
| CHJA | 0.575 | 0.575 | 0.528 | 0.528 | 1.000 | | | | | | | | | | | | | | |
| CHSO | 0.575 | 0.575 | 0.528 | 0.528 | 1.000 | 1.000 | | | | | | | | | | | | | |
| SHAN | -1.000 | -1.000 | -0.266 | -0.266 | -0.575 | -0.575 | 1.000 | | | | | | | | | | | | |
| TURN | -0.266 | -0.266 | -1.000 | -1.000 | -0.528 | -0.528 | 0.266 | 1.000 | | | | | | | | | | | |
| ABUN | -0.132 | -0.132 | 0.295 | 0.295 | 0.165 | 0.165 | 0.132 | -0.295 | 1.000 | | | | | | | | | | |
| RICH | -0.038 | -0.038 | -0.016 | -0.016 | -0.054 | -0.054 | 0.038 | 0.016 | 0.064 | 1.000 | | | | | | | | | |
| DELT | 0.007 | 0.007 | -0.020 | -0.020 | -0.021 | -0.021 | -0.007 | 0.020 | -0.011 | 0.346 | 1.000 | | | | | | | | |
| DSTR | 0.013 | 0.013 | -0.023 | -0.023 | -0.020 | -0.020 | -0.013 | 0.023 | -0.018 | 0.308 | 0.996 | 1.000 | | | | | | | |
| LPLU | 0.027 | 0.027 | 0.021 | 0.021 | 0.072 | 0.072 | -0.027 | -0.021 | -0.028 | -0.273 | -0.303 | -0.296 | 1.000 | | | | | | |
| DPLU | 0.012 | 0.012 | -0.024 | -0.024 | -0.019 | -0.019 | -0.012 | 0.024 | -0.018 | 0.312 | 0.996 | 0.999 | -0.297 | 1.000 | | | | | |
| SPLU | -0.036 | -0.036 | -0.017 | -0.017 | -0.054 | -0.054 | 0.036 | 0.017 | 0.061 | 0.995 | 0.421 | 0.384 | -0.274 | 0.388 | 1.000 | | | | |
| FEVE | 0.040 | 0.040 | -0.005 | -0.005 | 0.007 | 0.007 | -0.040 | 0.005 | -0.013 | -0.157 | -0.207 | -0.199 | 0.019 | -0.200 | -0.169 | 1.000 | | | |
| FDIV | -0.012 | -0.012 | 0.020 | 0.020 | 0.022 | 0.022 | 0.012 | -0.020 | 0.028 | 0.081 | 0.162 | 0.157 | 0.015 | 0.157 | 0.093 | -0.608 | 1.000 | | |
| FDIS | 0.006 | 0.006 | 0.037 | 0.037 | 0.047 | 0.047 | -0.006 | -0.037 | 0.019 | 0.048 | 0.135 | 0.131 | -0.006 | 0.132 | 0.058 | -0.389 | 0.922 | 1.000 | |
| RAOQ | 0.002 | 0.002 | 0.044 | 0.044 | 0.051 | 0.051 | -0.002 | -0.044 | 0.013 | 0.057 | 0.138 | 0.133 | -0.009 | 0.135 | 0.067 | -0.436 | 0.920 | 0.985 | 1.000 |

with either the community structure or the function indices. The only exception was species richness (RICH), with which most taxonomic diversity indices had some degree of correlation, the largest being for SPLU $\rho$=0.995.

The functional evenness (FEVE) and functional divergence (FDIV) gave $\rho = -0.608$. A slightly lower value was found for FEVE against RAOQ and FDIS, $\rho = -0.436$, and $\rho = 0.985$ respectively.

To further investigate patterns between raw indices of biodiversity, a heatmap is constructed below (Figure 5.1).

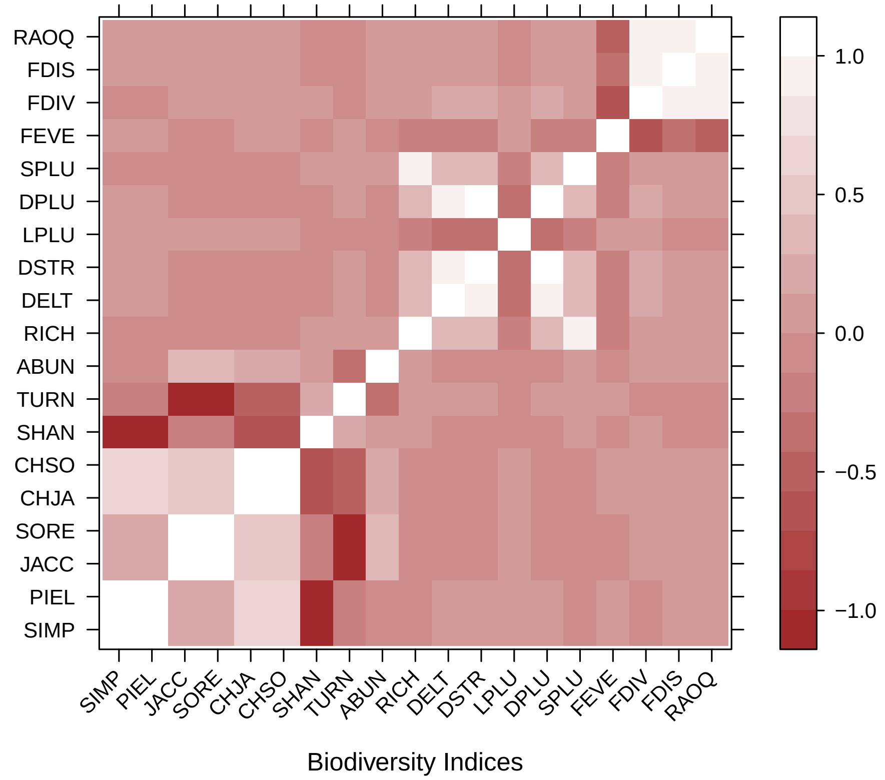


**Figure 5.1:** Spearman correlation between different biodiversity indices in a two-dimensional space represented as a heatmap. Dark shade indicates negative correlation, light shade – positive correlation

## 5.3.2 Principal component analysis

Principal component analysis (PCA) includes a number variables attributed to three conceptually distinct groups of indices: (A) community structure and composition; (B) taxonomic; and (C) functional diversity. The results of the PCA are usually discussed in terms

**Table 5.3:** Rotation matrix: correlations of each of the variables with the first principal axes

| | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|
| SIMP | -0.363 | 0.032 | -0.058 | 0.342 | -0.040 |
| PIEL | -0.363 | 0.032 | -0.058 | 0.342 | -0.040 |
| JACC | -0.326 | 0.050 | 0.043 | -0.422 | 0.037 |
| SORE | -0.323 | 0.051 | 0.045 | -0.427 | 0.036 |
| CHJA | -0.360 | 0.036 | 0.001 | 0.101 | -0.004 |
| CHSO | -0.378 | 0.036 | 0.011 | 0.047 | -0.006 |
| SHAN | 0.363 | -0.032 | 0.058 | -0.342 | 0.040 |
| TURN | 0.326 | -0.050 | -0.043 | 0.422 | -0.037 |
| ABUN | -0.045 | 0.024 | 0.020 | -0.278 | -0.072 |
| RICH | 0.042 | 0.275 | -0.192 | -0.052 | -0.590 |
| DELT | 0.044 | 0.395 | -0.278 | -0.003 | 0.297 |
| DSTR | 0.043 | 0.389 | -0.275 | 0.005 | 0.322 |
| LPLU | -0.031 | -0.168 | 0.161 | 0.055 | 0.182 |
| DPLU | 0.043 | 0.390 | -0.275 | 0.005 | 0.320 |
| SPLU | 0.042 | 0.298 | -0.204 | -0.047 | -0.550 |
| FEVE | -0.039 | -0.246 | -0.258 | -0.039 | 0.085 |
| FDIV | 0.032 | 0.311 | 0.440 | 0.067 | -0.003 |
| FDIS | 0.021 | 0.295 | 0.443 | 0.059 | 0.026 |
| RAOQ | 0.020 | 0.300 | 0.440 | 0.063 | 0.014 |

of component scores and rotation. Scores is a matrix that contains the original data in a rotated coordinate system. Rotation is the matrix of variable loadings i.e., a matrix whose columns contain the eigenvectors. The signs of the columns of the rotation matrix are arbitrary. Correlations of each of the variables with the principal axes are shown in Table 5.3.

While summing squares over any two principal components gives variance explained in two-dimensional space, summing the squares of each principal component gives the amount of variance on each axis. These new principal components are orthogonal by definition.

To show the fraction of total variance in the data as represented by each PC I plot the variances against the number of the principal component (see Figure 5.2).

The first three axes of the PCA (of 19 indices) accounted for 61.7% of total variation, the first five axes accounted for 82.3%. The community structure cluster of indices (see Figure 5.4) contributed almost exclusively to PC1, while taxonomic and functional diversity contributed approximately equally to PC2 and PC3 (Figure 5.3). The largest single proportion of variance among the top three PCs was explained by the functional diversity cluster (Figure 5.4) of indices. This means that functional diversity was the most variable kind of biodiversity among simulated communities.

The proportion of total diversity estimated by species richness alone was very variable

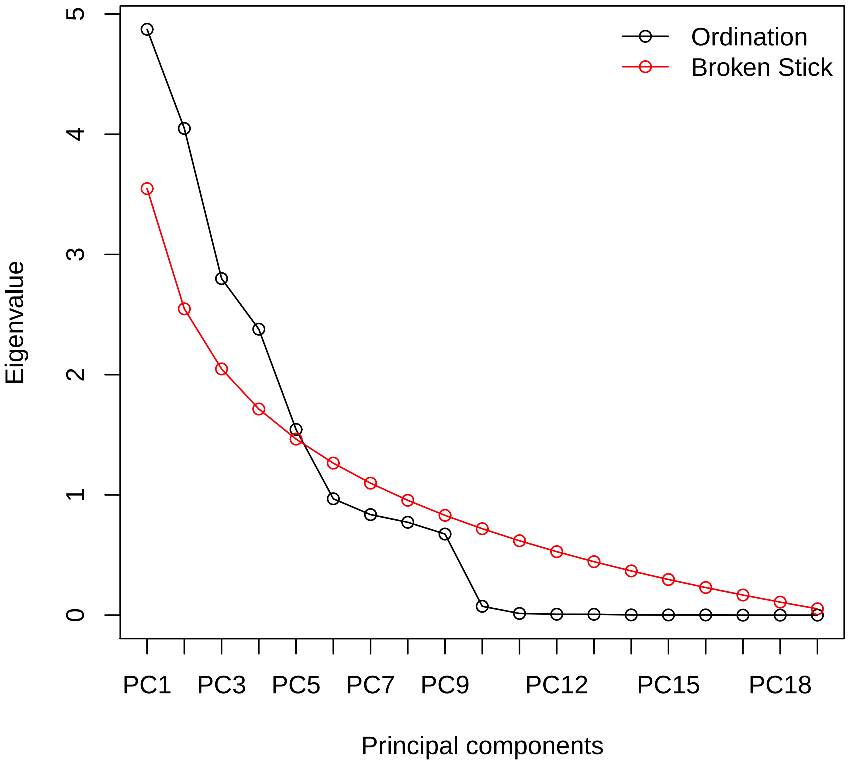


**Figure 5.2:** Screeplot shows the percentage of variance explained (y-axis) versus the principal components (x-axis). It indicates a clear separation in fraction of total variance. The point of separation is referred to as ”elbow”. Broken stick method suggests to retain first three to five axes

and typically low (Figure 5.5). Inevitably, this single measure – species richness – gives a minimum estimate of total biodiversity, but the extent to which information is lost when this is the sole measure of diversity is striking.

To summarise, the results of this analysis quantitatively demonstrate that conceptually different groups of measures of biodiversity remain persistently distinct and separated by different PC axes. These axes are orthogonal by construction, such that, a conclusion can be made that each of the distinct groups of measures is orthogonal to one another. It is suggested, therefore, that to give a comprehensive description by accounting for the largest proportion of variability in biodiversity metric space, indices that belong to different groups need to be considered included.

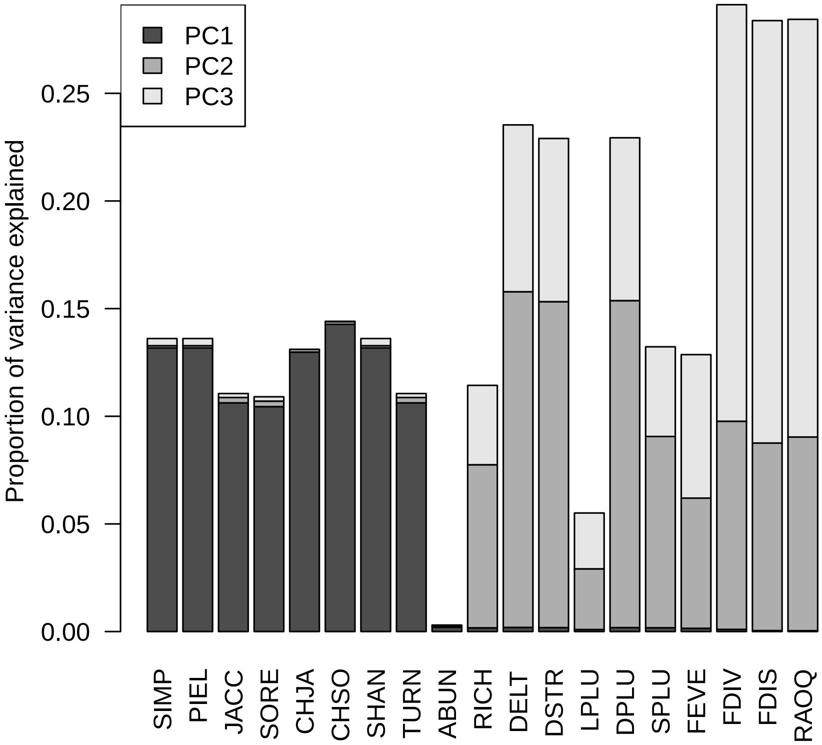


**Figure 5.3:** Variance explained by each variable in three dimensional space (61.7%). Different segments on the bar show contribution of the variables to corresponding principal components

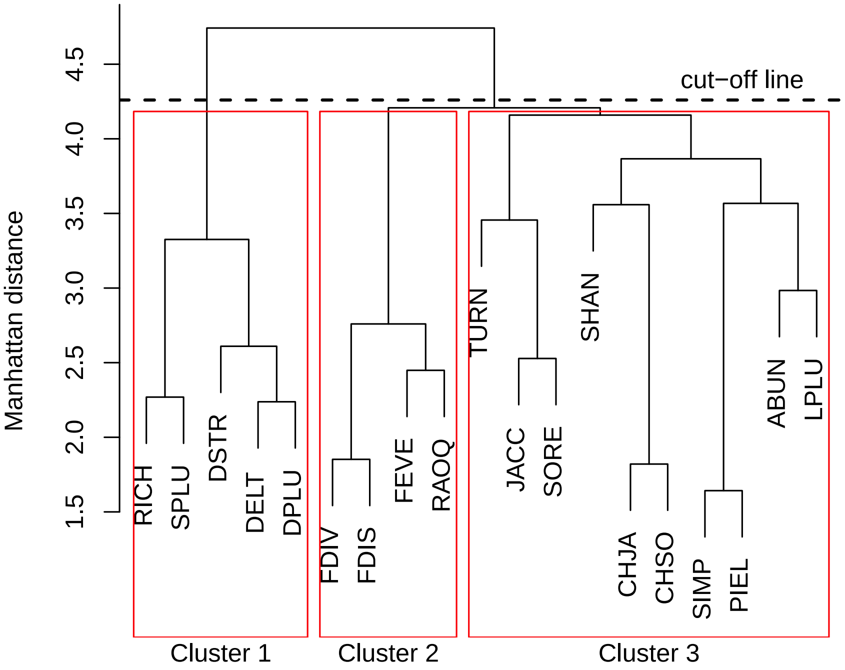


**Figure 5.4:** Hierarchical clustering of biodiversity indices using Manhattan distance. This dendrogram is calculated using the rotation matrix of variables, the columns of this matrix contain eigenvectors (principal components). Depending on a cut off point several clusters can be observed: Cluster 1 – taxonomic diversity; Cluster 2 – functional diversity; and Cluster 3 – community structure and composition

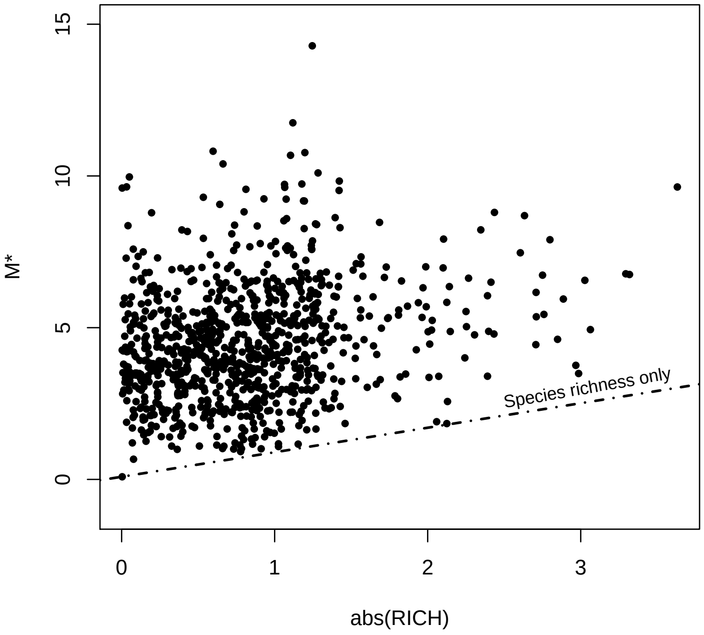


**Figure 5.5:** The y-axis is scalar distance from the origin in the largest three PC's ($M^*$), plotted against species richness (RICH) for 1000 synthetic communities shown here rank-ordered by species richness from left to right. Dashed line shows the contribution of species richness alone on the y-axis

## 5.4 Discussion and Conclusions

As conservation priorities move from single charismatic species to whole ecological communities and economists demand quantitative justifications for conservation, the need for a unifying measure of biodiversity mounts. At the start of this chapter I asked how well the simplest, most commonly used biodiversity index – species richness – meets this demand. If biodiversity is truly the aggregate of functional, structural, and taxonomic diversity, then Figure 5.5 shows species richness to be missing a significant portion of the information. This happened because in present simulations, functional, structural, and taxonomic variety were not simple correlates of species richness, as indeed they are not in real life.

Recalling that, in these simulations, community structure indices showed no correlation to functional traits but instead were based on a lognormal abundance distribution, closely matching empirical data (see Figure 4.11). Despite this obvious weakness, I see that the aggregated result of structural, taxonomic, and functional diversity (Figure 5.5), is that species richness no more than sets the lower limit of biodiversity, which varies above this limit in ways unrelated to species richness. In particular, Figure 5.3 shows that species richness contributes almost nothing to the first principal axis, which is dominated by the community structure indices: variation in structural diversity was independent of species richness. The lack of correlation between structural and functional indices does not weaken that finding. Community ecologists have long known about this (e.g., Magurran, 2004; Wilsey et al., 2005), which is why structural indices such as Simpson's are well supported. So my first message must be that species richness, quick and simple though it is, turns out to be a rather poor estimate of biodiversity as I have defined it.

The intuition of Franklin (1988) and Noss (1990) that biodiversity is essentially three dimensional, the axes being: structural, phylogenetic, and functional diversity is partially confirmed by this analysis. However, I also see built-in correlations among these traits of biodiversity, so the axes are unlikely to be strictly orthogonal. In simulated communities, structural aspects of biodiversity were found to be least correlated with function and taxonomy, but this is not surprising since abundances, from which structural indices were calculated, were generated by a process that was independent of the taxonomy and function.

Correlations among various types of biodiversity indices have often been reported (e.g., Mérigot et al., 2007; Heino, 2008). Gallardo et al. (2011) found significant correlations among some biodiversity indicators among river organisms, especially with Shannon diversity. They also discovered that these were correlated to environmental variation, particularly in relation to human disturbance, so substantial co-variation was likely. Indeed, Gallardo et al. (2011) comment that it is difficult to know the extent to which correlation among metrics is in fact co-variation with main environmental drivers. If this were known, then correlations may become a good indicator of environmental stress. The method of resampling deployed here, generates a population of communities which are entirely independent of environmental conditions, hence any remaining correlations among indices (as I found) are inherent.

These results raise important questions about priorities in biodiversity measurement. To the extent that function, taxonomy (or better still phylogeny), and community structure are found to be substantially independent, then any observed community structure may be produced from a wide variety of species (implying species substitutability). Further, many different communities, with different species compositions, could perform equivalent functions (also implying species substitutability). Thus, the strength of inherent correlations among the three major categories of biodiversity sheds light on species substitutability.

The practical consequence of this analysis is a parsimonious one. Faced with the urgent need to describe the rapidly declining diversity of life on earth, as comprehensively as possible but with limited resources, I see that no more than three well chosen indices are necessary. In the extreme of emergency cataloguing, I find that the simplest of all indices – species richness – performs poorly as a single surrogate for the three aspects of biodiversity, but of course it still may be the only practical option. When species, their phylogeny and significant functional traits are catalogued together in accessible databases, then field-collected species lists will serve as a key to estimating biodiversity in its fuller meaning. The need for this development sets an urgent goal for future biodiversity action.

## Summary

1. Nineteen taxonomic, functional, and compositional biodiversity indices over the population of 1000 communities were calculated;

2. Three main groups of indicators emerged from multivariate analysis: these were community composition (a surrogate for structure), taxonomic diversity (a surrogate for phylogenetic diversity), and functional diversity, based on ecological traits;

3. Species richness set the lower limit of biodiversity capturing only 21% of diversity among communities; and

4. Biodiversity has been shown to be irreducibly three dimensional.

# Appendices

## Appendix 5.A Community structure and composition indices

```
#Variables number of classes, orders, families, etc
x1<-vector("list",length(com))
for (i in seq(along=com)) {
for (j in seq(along=com[[i]])){
x1[[i]][[j]]<-rbind(com[[i]][[j]][2])
}
x1[[i]]<-do.call("cbind",x1[[i]])
}
x1<-matrix(unlist(x1),ncol=5)

#Number of unique classes, orders, families, etc
x2<-vector("list",length(com))
for (i in seq(along=com)) {
for (j in seq(along=com[[i]])){
x2[[i]][[j]]<-rbind(com[[i]][[j]][3])
}
x2[[i]]<-do.call("cbind",x2[[i]])}
x2<-matrix(unlist(x2),ncol=5)

#Abundance estimate
x3<-vector("list",length(com))
for (i in seq(along=com)){
for (j in seq(along=com[[i]])){
x3[[i]][[j]]<-fitdistr(com[[i]][[j]][[4]],"Poisson")$estimate
x3[[i]][[j]]<-rbind(x3[[i]][[j]])}}
x3<-matrix(unlist(x3),ncol=5)

#(2)Calculate indices

#(2.1)Community structure and composition
#Shannon-Wiener (Joshi et al 39)
index1<-index<-vector("list",length(com))
for (i in seq(along=com)) {
for (j in seq(along=com[[i]])){
index[[i]][[j]]<-sum(sapply(unlist(com[[i]][[j]][4]),function(x){x/unlist(com[[i]][[j]][3])
    })*log(sapply(unlist(com[[i]][[j]][4]),function(x){x/unlist(com[[i]][[j]][3])})))
index1[[i]][[j]]<-rbind(index[[i]][[j]])}}
#to convert list to matrix:
index1<-matrix(unlist(index1),ncol=5)

#Index of evenness/equitability or Pielou: (class level) H/log(S) (Joshi et al 39)
index2<-lapply(index1, function(x){x/log(sum(unlist(com[[1]][[5]][4])))})
index2<-matrix(unlist(index2),ncol=5)


#Jaccard's dissimilarity (Anderson 1...)
index3<-index<-vector("list",length(com))
for (i in seq(along=com[-1])){
for (j in seq(along=com[[i]])){
index[[i]][[j]]<-sum(unique(com[[i]][[j]][[1]])%in%unique(com[[i+1]][[j]][[1]]))/(com[[i
    ]][[j]][[2]]+com[[i+1]][[j]][[2]]-sum(unique(com[[i]][[j]][[1]])%in%unique(com[[i
    +1]][[j]][[1]])))
index3[[i]][[j]]<-rbind(index[[i]][[j]])}}
index3<-matrix(unlist(index3),ncol=5)

#Sorensen similarity (Moreno et al 52)
index4<-index<-vector("list",length(com))
for (i in seq(along=com[-1])){
```

```
for (j in seq(along=com[[i]])){
index[[i]][[j]]<-sum(unique(com[[i]][[j]][[1]])%in%unique(com[[i+1]][[j]][[1]]))/(com[[i
    ]][[j]][[2]]+com[[i+1]][[j]][[2]])
index4[[i]][[j]]<-rbind(index[[i]][[j]])
}}
index4<-matrix(unlist(index4),ncol=5)

#Chao-Jaccard (Moreno et al 52)
index5<-index<-vector("list",length(com))
for (i in seq(along=com[-1])){
for (j in seq(along=com[[i]])){
t1<-sum(table(factor(com[[i]][[j]][[1]][unique(com[[i]][[j]][[1]])%in%unique(com[[i+1]][[j
    ]][[1]])])))/length(com[[i]][[j]][[1]]))
t2<-sum(table(factor(com[[i+1]][[j]][[1]][unique(com[[i]][[j]][[1]])%in%unique(com[[i+1]][[
    j]][[1]])])))/length(com[[i+1]][[j]][[1]]))
index[[i]][[j]]<-(t1*t2)/(t1+t2-t1*t2)
index5[[i]][[j]]<-rbind(index[[i]][[j]])}}
index5<-matrix(unlist(index5),ncol=5)

#Chao-Sorensen (Moreno et al 52)
index6<-index<-vector("list",length(com))
for (i in seq(along=com[-1])){
for (j in seq(along=com[[i]])){
t1<-sum(table(factor(com[[i]][[j]][[1]][unique(com[[i]][[j]][[1]])%in%unique(com[[i+1]][[j
    ]][[1]])])))/length(com[[i]][[j]][[1]]))
t2<-sum(table(factor(com[[i+1]][[j]][[1]][unique(com[[i]][[j]][[1]])%in%unique(com[[i+1]][[
    j]][[1]])])))/length(com[[i+1]][[j]][[1]]))
index[[i]][[j]]<-(2*t1*t2)/(t1+t2)
index6[[i]][[j]]<-rbind(index[[i]][[j]])}}
index6<-matrix(unlist(index6),ncol=5)

#Shannon (Moreno et al 52)
index7<-index<-vector("list",length(com))
for (i in seq(along=com)) {
for (j in seq(along=com[[i]])){
index[[i]][[j]]<--1*sum(sapply(unlist(com[[i]][[j]][4]),function(x){x/unlist(com[[i]][[j
    ]][3])})*log(
sapply(unlist(com[[i]][[j]][4]),function(x){x/unlist(com[[i]][[j]][3])}),10
))
index7[[i]][[j]]<-rbind(index[[i]][[j]])}}
#to convert list to matrix:
index7<-matrix(unlist(index7),ncol=5)

#"Species"Turnover (Zamora et al. 79)
index8<-index<-vector("list",length(com))
for (i in seq(along=com[-1])){
for (j in seq(along=com[[i]])){
t1<-sum(unique(com[[i]][[j]][[1]])%in%unique(com[[i+1]][[j]][[1]]))
t2<-com[[i]][[j]][[2]]
t3<-com[[i+1]][[j]][[2]]
index[[i]][[j]]<-(t2+t3-2*t1)/(t2+t3-t1)
index8[[i]][[j]]<-rbind(index[[i]][[j]])}}
index8<-matrix(unlist(index8),ncol=5)

#cbind taxonomic indices at species level only [,1]. [,2-5] correspond to other taxonomic
    levels
Istructure<-data.frame(cbind(
index1=index1[,1],
index2=index2[,1],
index3=rbind(index3,NA)[,1],
index4=rbind(index4,NA)[,1],
index5=rbind(index5,NA)[,1],
index6=rbind(index6,NA)[,1],
index7=index7[,1],
index8=rbind(index8,NA)[,1],
x1=x1[,1],
x2=x2[,1],
x3[,1]
))
names(Istructure)<- c("ShWierner","Pielou","Jaccard","Sorensen","ChaoJaccard","ChaoSoren","
    Shannon","Turnover","Number","UniqueNumber","Abundance")
```

# Appendix 5.B Taxonomic diversity indices

```
##(2.2)Taxonomic diversity
`taxondive0` <-
    function (comm, dis, match.force = FALSE)
{
    binary <- FALSE
    comm <- as.matrix(comm)
    if (missing(dis)) {
        n <- ncol(comm)
        dis <- structure(rep(1, n * (n - 1)/2), Size = n, class = "dist")
    }
    dis <- as.dist(dis)
    if (match.force || attr(dis, "Size") != ncol(comm)) {
        if (match.force)
            message("Forced matching 'dis' labels and 'comm' names")
        else
            message("Dimensions do not match between 'comm' and 'dis'")
        if (all(colnames(comm) %in% labels(dis))) {
            dis <- as.matrix(dis)
            dis <- as.dist(dis[colnames(comm), colnames(comm)])
            message("Matched 'dis' labels by 'comm' names")
        } else {
            stop("Could not match names in 'dis' and 'comm'")
        }
        if (length(unique(colnames(comm))) != ncol(comm))
            stop("Names not in unique in 'comm': match wrong")
        if (length(unique(labels(dis))) != attr(dis, "Size"))
            warning("Labels not unique in 'dis': matching probably wrong")
    }
    del <- dstar <- dplus <- Ed <- Edstar <- edplus <- NULL
    if (!binary) {
        del <- apply(comm, 1, function(x) sum(as.dist(outer(x,
                                                            x)) * dis))
        dstar <- apply(comm, 1, function(x) sum(dis * (xx <- as.dist(outer(x,
                                                                          x))))/sum(xx))
        rs <- rowSums(comm)
        del <- del/rs/(rs - 1) * 2
        cs <- colSums(comm)
        tmp <- sum(as.dist(outer(cs, cs)) * dis)
        Ed <- tmp/sum(cs)/sum(cs - 1) * 2
        Edstar <- tmp/sum(cs)/(sum(cs) - 1) * 2
    }
    comm <- ifelse(comm > 0, 1, 0)
    dplus <- apply(comm, 1, function(x) sum(as.dist(outer(x,
                                                          x)) * dis))
    Lambda <- apply(comm, 1, function(x) sum(as.dist(outer(x,
                                                           x)) * dis^2))
    m <- rowSums(comm)
    dplus <- dplus/m/(m - 1) * 2
    Lambda <- Lambda/m/(m - 1) * 2 - dplus^2
    S <- attr(dis, "Size")
    omebar <- sum(dis)/S/(S - 1) * 2
    varome <- sum(dis^2)/S/(S - 1) * 2 - omebar^2
    omei <- rowSums(as.matrix(dis))/(S - 1)
    varomebar <- sum(omei^2)/S - omebar^2
    vardplus <- 2 * (S - m)/(m * (m - 1) * (S - 2) * (S - 3)) *
        ((S - m - 1) * varome + 2 * (S - 1) * (m - 2) * varomebar)
    out <- list(number = m, D = del, Dstar = dstar, Lambda = Lambda,
                Dplus = dplus, SDplus = m *
                dplus)
    class(out) <- "list"
    out
}

##without NA (5 rows missing)
result<-list()
for(i in seq(along=com) ){
hclass<-mydata2$tclass[mydata2$tspecies%in%com[[i]][[5]][[1]]]
horder<-mydata2$torder[mydata2$tspecies%in%com[[i]][[5]][[1]]]
hfamily<-mydata2$tfamily[mydata2$tspecies%in%com[[i]][[5]][[1]]]
hgenus<-mydata2$tgenus[mydata2$tspecies%in%com[[i]][[5]][[1]]]
hspecies<-mydata2$tspecies[mydata2$tspecies%in%com[[i]][[5]][[1]]]
hier<-cbind(hspecies,hgenus,hfamily,horder,hclass)
s<-nrow(hier)
s<-ifelse(s>1000,s/10,s) #take 10% only for those communities where n of species > 1000
taxa<-taxa2dist(hier[sample(s,replace=F),])
abd<-t(cbind(c(1:s),rep(1,s)))
taxdiv<-taxondive0(abd,taxa)
result[[i]]<-as.data.frame(taxdiv)[1,]}
Itaxonomy<-do.call(rbind,result)
```

# Appendix 5.C Functional diversity indices

```
##(2.3)Functional diversity
#import traitmatrix obtained from BIOTIC
traitmatrix<-read.table(file="./traitmatrix.csv",header=T,sep=",", na.strings=NA, strip.
    white=T)

#reshaping classes in com
comclass<-list()
for (i in seq(along=com) ){
comclass[i]<-list(com[[i]][[1]][[1]])}

#unique list of classes
#listclasses<-unique(mydata2$tclass) #115? classes present in mydata2
listclasses30<-traitmatrix$Class #30 classes that appear in traitmatrix

#obtaining abundance matrix
classabun<- t(sapply(comclass, function (x)
     table(factor(x, levels = listclasses30))))
classabun[classabun== 0] <- 1 #to give 1

#classabun[classabun== 0] <- NA

#calculating functional diversity indices
#gowdis(traitmatrix) #didnt use it
rownames(traitmatrix)<-colnames(classabun)
dbFD(traitmatrix, classabun)->Ifunction

#obtain a vector with T, F
which.NA <- apply(classabun, 1, function(x) all(is.na(x) ) )
Ifunction<-cbind(Ifunction$FEve,Ifunction$FDiv,Ifunction$FDis,Ifunction$RaoQ)
Ifunction[which.NA,]<-NA
colnames(Ifunction)<-c("FEve","FDiv","FDis","RaoQ")
```

# Chapter 6

# Implications: from theory to practice

## 6.1 Introduction

The purpose of this chapter is to draw conclusions from the previous five, developing a commentary on the significance and potential use of these to the practice of biodiversity metrification, in particular for economic valuation. The central hypothesis around which the whole study has been constructed is that biodiversity, at its most basic, is a measure of the degree of difference (i.e., the number of discrete differences) within a biological system. This was recognised as a measure of the total information content. Much of this information was understood to be random and therefore to have no interpretable effect. Taking inspiration from Gregory Bateson's writing Bateson (1972), it was recognised that only non-random "difference that makes a difference" information need be counted for economic valuation, on the grounds that only this information can be functional (i.e., show an effect). Thus, a further core hypothesis of this thesis is that the foundation of value in biodiversity is the strictly functional information content of the biological system. Further, since this information is by definition non-random, i.e., systematic pattern, it coincides with the scientific meaning of complexity. The main conclusion of the first two chapters of this thesis was therefore that measures of biological complexity, through identifying pattern in difference, can form the basis of a scientific measure of functional information.

It was further recognised that the foundation of value was the potential to cause a beneficial effect, so that only functional information need be counted in estimating potential value. The aim therefore became one of quantifying potentially valuable information from biodiversity metric data. The potential value accorded by this information is the scientific interpretation of "intrinsic" value, since, as potential, it is solely determined by the (information) properties of the biological system. For the potential value to be realised, it of course needs a valuer, since valuation describes a relationship between two (or more) entities. The theoretical developments of the first two chapters therefore set the goal of estimating the biological complexity (functional; meaningful) information content of a biological system, this being taken as the input to an economic valuation process.

The advantage of this approach to economic valuation is that it founds value on real measurable and intrinsic properties of systems, so is objective, in contrast with present opinion-based economic methods applied to biodiversity. This makes possible comparisons among systems, over time and among different studies in different locations – surely a requisite for international agreement on biodiversity action. The line of thinking, leading from difference to value has necessitated a mathematical precision of definition for biodiversity, which is in itself a useful development – having in Chapter 3, identified the confusion of terms and definitions confounding progress in biodiversity valuation.

Further, there are more academic advantages to understanding biodiversity as functional information. It provides a deep theoretical foundation for biodiversity which connects it consistently with philosophical ontology and physics – these together being the description of existence that should underpin all scientific subjects. It provides a concrete mathematical entity from which to calculate derived properties such as biodiversity index values from first principles. It enables the comparison of these values showing how they relate to one another quantitatively, again, from first principles. It places biodiversity in the broader context, as a measure of life itself, at all levels of biological organisation from molecule to the global "Gaia" system.

Finally, reiterating, this information-based approach provides a concrete, intrinsic, and system-independent currency for biodiversity (bits of information), comparable to the financial currencies used to quantify its estimated value. Indeed, with this, economic valuation of biodiversity can be summarised by the following steps: (a) identify the system under valuation; (b) estimate the functional information content of the system from available diversity data; and (c) convert this quantification of information into its financial equivalent via an information-money exchange rate.

There are still large gaps, so the procedure outlined is not immediately implementable, but it may now be regarded as a goal to guide further research. In this last chapter, I will look at potential benefits of pursuing such a goal, especially in the context recent moves to form international agreements on biodiversity preservation.

## 6.2 The history of Global action on biodiversity

If the world was not losing biological information through rapid decline in species number (May, 2011), study of biodiversity might be purely academic. However, to many, we are living through a time of biodiversity crisis (Crist, 2002), so that the primary goal for its academic study is that of finding the scientific basis for the means and justification of its conservation. This has, of course, made biodiversity a political and therefore an economic problem. Recognising the global span and trans-national causes and effects of biodiversity loss, this political economy arena has become international. The parallel development of political concerns alongside the science of biodiversity has strongly affected its conceptual development.

A brief overview of efforts by policy makers to conserve biodiversity, provides insights into the evolution of the concept of biological diversity since its first introduction nearly two decades ago. Here I concentrate on two main aspects of its development: the first one is the understanding of the concept of biodiversity and the second one is the understanding of the processes that underlie our efforts to protect it. The following review of the key events in chronological order, illustrates a gradual transition in thinking from ethical bio-wholism (the extension of rights and worth beyond humanity), via instrumentalism to the present anthropocentric utilitarian framework applied to biodiversity conservation. This is important so far as this wider (political economy) culture influences scientific concepts around biodiversity. If scientists need to reflect the public culture in their funding applications, then a close adherence to the prevalent thinking in the political economy seems very likely. The extent to which science is, or should be, free from such influences is beyond my scope, but where biodiversity research is motivated by the need to conserve, then it is reasonable to admit that policy direction exerts a strong influence.

The most prominent of state-sponsored international agreements on biodiversity was CBD (UN, 2011), which entered into force on 29 December 1993. It had three main objectives: the conservation of biodiversity, the sustainable use of the components of biological diversity and sharing of the benefits that arise out of the utilisation of genetic resources. Being the first global agreement on the conservation and sustainable use of biological diversity, it significantly acknowledges the fact that biodiversity sustains all life processes and contributes directly to human well-being. With its attempts to address global biodiversity loss, this biodiversity treaty gained rapid and widespread acceptance among those policy makers who were willing to accept that biodiversity was good and valuable in its own right.

Under the Convention, the “ecosystem approach to the conservation and sustainable use of biodiversity” was developed as a framework for action, in which all the goods and services provided by the biodiversity in ecosystems were to be considered. The ecosystem approach explicitly adopts the stance of humanity as “custodians” of nature through intergenerational equity, implicitly according ecosystems the “right” to exist in a healthy state. Further, by setting out the commitments to maintain the world’s ecological processes that are essential for human well-being, the link between biodiversity and ecosystem services was (although still indirectly) established. The statement “biological resources are the pillars upon which we build civilizations” clearly demonstrates an instrumental approach towards biodiversity.

For the first time it linked traditional conservation efforts to an economic goal of using biological resources sustainably. Thus, in this early inception, ethical, and instrumental justifications were used together to argue for conservation. The concept of ecosystem services was not formally linked to biodiversity until much later, when it appeared in The Millennium Ecosystem Assessment (Millenium Ecosystem Assessment, 2005).

The Convention's definition of biodiversity as "the variety of life on Earth and the natural patterns it forms" was, sadly, neither clear nor operational. Even now, nearly two decades after the CBD highlighted the lack of information and knowledge regarding biological diversity, it remains confused and obscure to most people (reviewed in Chapter 1). The original goals of the CBD have not been realised. Biodiversity, its value, and its threats, not only needed to be acknowledged but also well enough understood to take practical actions. Whilst it did raise the status of the issue, CBD contained no practically implementable tasks leading to biodiversity conservation (Harrop and Pritchard, 2011). Quite likely, one of the main reasons for this was the lack of quantifiability and precision in the definition of biodiversity and in its potential to be valued (Polski, 2005).

Ten years later, at the 2002 Johannesburg World Summit on Sustainable Development 190 countries agreed to "…achieve by 2010 a significant reduction of the current rate of of biodiversity loss at the global, regional, and national level …" (UNEP, 2011). The culmination point was achieved towards the end of 2002-2010 period, by declaring 2010 the "International Year of Biodiversity". This initiative, was essentially a retry to achieve the broad-defined objectives of halting or reducing the rate of biodiversity loss. Even then, it was clear that the strategic objectives for the decade were going to be hard to achieve. According to Mace et al. (2010), objectives need to be reformulated so as "to avoid undesired and dangerous biodiversity change and to strengthen the role of biodiversity in securing and enhancing the benefits that people derive from ecosystems". The same year many negative expectations were confirmed in The Global Biodiversity Outlook (GBO) (Secretariat of the Convention on Biological Diversity, 2010). At this point metrification of biodiversity had been adopted.

Drawing on a range of information sources to summarise the most up-to-date status and trends of biodiversity, the third edition (GBO-3) drew conclusions regarding the future of the Convention. It said "having reviewed all available evidence, including national reports submitted by Parties, this third edition of the Global Biodiversity Outlook concludes that the target has not been met". Even worse, in some cases biodiversity loss was estimated as intensifying, an observation which leading scientific commentators had already attributed to the vagueness of the indicators and important gaps in knowledge (see, e.g., Mace and Baillie, 2007).

To assess progress towards the 2010 targets the CBD selected a set of "22 headline indicators". Although these covered a broad set of potentially useful features to be maintained, this set was neither complete (Walpole et al., 2009), nor necessarily mutually compatible (Mace and Baillie, 2007). With some indicators being only weak proxies for biodiversity a measurement problem was evident. Many indicators were selected primarily for data availability (Mace et al., 2010). Since decisions on biodiversity management and conservation were by then largely based on indicators (Fischer et al., 2011), they needed to be linked explicitly to monitoring objectives (Jones et al., 2011). At this point it had become clear that the metrification of biodiversity was too confused and heterogeneous to support the practical aims of conservation. Accompanying this realisation, a narrowing of definition and goals to concentrate on the anthropocentric instrumental value of biodiversity came to dominate.

This was reflected in the The Millennium Ecosystem Assessment, which was particularly calling for indicators to fill the gaps of the biodiversity-ecosystem services link. The situation that needs to be addressed is clearly acute: it is claimed “60% of the Earth’s ecosystem services that have been examined have been degraded in the last 50 years”.

Further focusing on this more pronounced relationship between biodiversity-ecosystem services, The Economics of Ecosystems and Biodiversity report (TEEB, 2010) explored aspects of the economic significance of the global loss of biodiversity. This study, inspired by the recent success of the climate change studies, was initiated by the German Federal Ministry for the Environment and European Commission. The study leader, Pavan Sukhdev concluded that “the science of biodiversity and ecosystem is still evolving, their services to humanity only partially mapped and imperfectly understood”. He added that “you cannot manage what you do not measure”. The Economics of Ecosystems and Biodiversity report presented an approach intended to help decision makers to recognise and capture the values of ecosystem services and biodiversity, while recognising the plurality of the values and the techniques available for their assessment.

In its second phase, The Economics of Ecosystems and Biodiversity report built a valuation framework based entirely on knowledge of how ecosystems function and deliver services. If we are to manage our ecological security “we must measure ecosystems and biodiversity scientifically as well as economically”: biodiversity preservation was in this way framed in anthropocentric utilitarian terms, now separating scientific from economic measurement.

The report showed that on average one third of Earth’s habitats have been damaged by humans. This damage is heterogeneous, for example up to 85% of seas and oceans and more than 70% of Mediterranean shrubland have been affected. The report warned that in spite of growing awareness of the dangers, destruction of nature will “still continue on a large scale”, quoting the example that estimated species loss-rate is up to 10,000 times higher than natural.

Even though the understanding of the ways in which biodiversity and ecosystem services are linked has been increasing (Hooper et al., 2005), it is still very hard to scale this knowledge up beyond the level of small ecological experiments, since it is highly scale-dependent (Armsworth et al., 2004). The understanding of underlying mechanisms and lack of relevant biodiversity metrics to quantify these mechanisms still remain the major problem of the utilitarian biodiversity-ecosystem services approach (Feld et al., 2009). The hypothetical relationship between biodiversity and ecosystem functions underlying the services is frequently no more than implicit (Bengtsson, 1998) and it is affected by several exogenous factors (Danovaro and Pusceddu, 2007).

This led Spangenberg and Settele (2010) to conclude that the ecosystem service valuation delivers "context and method dependent price estimates, possibly several for the same service, based on a wide range of subjective, hypothetical, and partly questionable assumptions".

The conclusions made in GBO-3 have contributed to the formulation of the Strategic Plan 2011-2020 which was mostly concerned with the development of the post-2010 indicators. Similarly, new commitments signed by the Conference of the Parties in Nagoya (COP10, 2010) are urging a need to halt the loss of biodiversity and the degradation of ecosystem services by 2020 at the latest. The Aichi Biodiversity Targets aim to achieve this but stand as a set of aspirations, rather than concrete, practical steps. For example, Target 12 requires that "By 2020 the extinction of known threatened species has been prevented and their conservation status, particularly of those most in decline, has been improved and sustained." There is no indication for a scientific basis or means of achievement. The only quantified practical agreements achieved at Nagoya were commitments to increase the amount of the planet set aside for biodiversity protection to 17% of the land surface and 10% of the oceans. This constitutes a dramatic reduction in conceptual complexity in effect replacing ideas of biodiversity value and ecosystem services with a simple (perhaps naive) categorical requirement. In so far as the proportion of land and sea set-aside may be determined by political negotiation, this goal has no need of science at all. Taking all twenty targets together, this most recent commitment is still very similar to others, which sadly questions its feasibility. It differs from its previous counterparts by additional emphasis put on ecosystem services, thus making a link between biodiversity-ecosystem services even stronger. The multiple, overlapping aims of the Aichi Targets and their general lack of quantification are caused by the continuation of confusion (or at least multiple options) over the definition of and quantitative estimation of biodiversity (Feest et al., 2010). The relationship between biodiversity and value remains especially "kaleidoscopic".

From this history of international policy in relation to biodiversity, I conclude that both the aims and the chance of success in implementing them are dependent on a precise and operationally quantifiable definition of biodiversity, the lack of which has hampered progress. Successive efforts to gain a working level of precision have driven the policy sphere towards anthropocentric utilitarianism, defining biodiversity value by its ability to deliver ecosystem services – broadly defined as natural economic goods. The practical difficulty with this interpretation is that real quantitative relationships between biodiversity and ecosystem service are hard to come by. We seem to have failed to break through the fundamental conceptual challenge of understanding a) what biodiversity really is and b) what it is about biodiversity, that we value. It was in this context that my thesis was formulated as an attempt to obtain scientific answers to those two questions.

There are two main questions, then: one concerning the basic meaning of biodiversity that has not been redefined since it first came to prominence 20 years ago, the other concerning how biodiversity and ecosystem services link, this being required by the present economics framework of biodiversity assessment, which I find largely lacking scientific understanding. As predicted, in a reflection of the prevailing culture of political economy, recent public funding initiatives have boosted the development of ecosystem services analysis (e.g., NERC BESS – Natural Environment Research Council Biodiversity and Ecosystem Services Sustainability, in the UK). Study to understand the foundation of value in biodiversity remains neglected. Part of the explanation for this may be the need for public support for biodiversity research.

The lack of public awareness is often considered one of the most serious barriers in achieving the objectives of the Biodiversity Convention. Public participation is an essential part of forming biodiversity-related policies, it is argued (Fischer and Young, 2007). Public understanding of biodiversity is essential for public support, but the scientific community has long been criticised for failing to produce indicators of practical use to policy makers attempting to achieve targets (Balmford et al., 2005). A vociferous fraction of society, of course, supports conservation for its own sake, but widespread backing is sought via economic utilitarian arguments. What would be most useful to policy makers in this context is a clear and universal (so transferable) quantitative relation between biodiversity and the economic value of ecosystem services (Armsworth et al., 2004). The combined efforts of science and economics show little sign to-date of producing this relation. The only remaining alternative is to appeal to the direct value of biodiversity: not via ecosystem services, but via self-evident value in biodiversity. This does not necessarily constitute a return to the "deep ecology" ethical basis for valuing. My conclusions in Chapter 2 indicated that "biodiversity as information" is the raw material for ecosystem function – providing a theoretical, but mechanistic explanation for the link with ecosystem services. In the language of Environmental Economics, biocomplexity generates Indirect Use Value (IUV).

This is especially important in light of the new the EU strategic targets set to be achieved by 2050. According to this new vision, biodiversity and the ecosystem services it provides (the natural capital), must be protected, valued, and appropriately restored for *biodiversity intrinsic value* as well as for their contribution to human wellbeing and economic prosperity. The implication in this vision statement – that EU policy makers believe biodiversity to have intrinsic value – in turn implies that they understand biodiversity to be an independent entity and value to be one of its properties. This is clearly not compatible with understanding biodiversity as the numerical value of a scientific measure taken from a biological system. We would no more accord intrinsic value to that than to, say, a temperature.

The definition of biodiversity as functional bio-information can, however, be thought of in the way the EU have invisaged. This is because information is an independent entity with a potential to be valued as one of its intrinsic properties. For the first time this transforms biodiversity from a concept that is hard to specify in concrete terms, to a physical entity with intrinsic properties, so that inherent value may be instrumental, rather than simply an expression of ethical preferences.

Thus, the work of this thesis offers a new alternative to operationalise biodiversity at a policy level. It does so by suggesting a way to translate from the ecological properties of a system, as they are measured by ecologists, into a simply defined measure of a single valued property of the system. This is a high-level concept, implementation which requires a great deal of more detailed study. An early sketch of this development work was reported in Chapters 3-5 of my thesis.

## 6.3 Summarising the contribution of this study

By interpreting biodiversity as an estimator of the functional information content of a biological system, I have shown how it can be transformed from a loosely defined, context dependent concept to a specific, quantifiable inherent property of the system. This, in principle, enables its objective valuation via specification of the indirect use value. In identifying functional information at multiple levels of biological organisation and instantiated in different axes of variation, information-based biodiversity is explicitly recognised as multi-dimensional within a structure which I have formalised using the descriptor-level permutation matrix structure. To fully quantify biodiversity the total biological complexity at all levels of organisation in a biological system must therefore be identified. In practice, we tend to focus on the species-level and consider contributions from below species-level as necessary for specifying and enabling the functions of species (this assumption is for example implicit in the Noah's Ark problem formulation). Existing indicators of phylogenetic diversity, described in Chapters 2-3, serve as a proxy for functional information sub-species levels. Levels of organisation above that of species are represented by system structure indices (also described in Chapters 2-3), so that a combination of species, phylogenetic, and structural indicators may be a surrogate for whole biodiversity. The need for such an aggregate estimate of the system's functional information content follows from my aim of identifying a scientifically-based objective source of value inherent in the biodiversity of ecosystems. It has led to a search for patterns (indicating functionality) among the components of biodiversity in Chapters 2 and 3. An initial conclusion was that existing empirical data was distributed in such a way as to make this aggregation, at best very difficult. Given the data I was able to collect, it turned out to be impossible. Since it is important to see if, in principle, the formation of an aggregate measure representing "whole biodiversity" (functional information content) would be possible and useful, a simulation approach was taken. Chapter 5 demonstrated that using synthetic communities with natural statistical properties, patterns in ecological field-data could in principle yield an aggregate measure that captures a great deal of the diversity among systems that is missed by species diversity alone. I concluded that, in principle, a single measure (the scalar distance) of biodiversity as functional information was possible and would be useful for conservation decisions, which trade off against economic costs.

The aggregation of different kinds of biodiversity estimators is not, in fact, a new innovation. Ecologists, being aware of the problem of linking multiple measures to a single economic value have attempted to construct a unifying single-value index of biodiversity. Current literature, especially relating to the assessment of efficacy in biodiversity strategies, shows this. For instance, an index proposed by Certain et al. (2011) suggested the building of a set of biodiversity indicators by aggregating knowledge available within the Ecological Research Network – a framework that collates knowledge on biodiversity and the state on ecosystems from a network of experts (Henry et al., 2008). This new index was called the Nature Index, and it is certainly easy-to use for policy makers. However its major drawback is that it is based on expert opinions regarding reference states of biodiversity. Reference states are pointing towards different situations and, therefore, a unified scientific foundation is still lacking. Additionally, this index suffers from heterogeneity, which authors attribute to heterogeneities in knowledge available. A weighting-system to control for heterogeneities was used as a remedy, but, based on unknown states of ecosystems, the problem remained unresolved. Thus, the additional advantage of an aggregate based on the hypothesis of biodiversity measuring functional information is that it refers back to a single quantifiable property inherent in the system, rather than a heterogeneous set of reference points. My analysis from Chapters 2, 4, and 3 attempts to justify this choice by demonstrating a clear path of calculation from measurable system properties to the notion of information-based biodiversity as a single valued metric of system complexity, which Chapter 1 argued could be interpreted as the same as functional information content.

Acknowledging that multiple indicators arise from the fact that biodiversity is a multidimensional concept, implies that it must be measured in a multidimensional space. In reviewing the biodiversity literature, I found that most (by far) published quantitative studies concentrated on just one dimension or a very small subset of the potential space; overwhelmingly dominated by species richness. Further, although some recognition of multiple "facets" is apparent, surprisingly little appears in the literature to re-integrate single dimensions into the multidimensional concept of biodiversity. In analysis of empirical biodiversity reports, it was not possible to say which axes show largest variation, or even to say what the axes are, with any confidence. This was demonstrated by a meta-analysis review of the empirical biodiversity literature (Chapter 3), supported through constructing a relational database of measures of biodiversity. Though limited in sample size by time-constraints, this meta-review is the first organisation of biodiversity concepts using a unifying mathematical structure. The formal procedures of relational database design made concrete the theoretical classification of biodiversity metrics and their relations. The resulting design provides a suitable foundation for integrating biodiversity knowledge across the broadest range of metrics and concepts.

Making the first use of this formal structure, I showed that although the way biodiversity is measured varies in many practical ways, the species level and simple richness descriptor overwhelmingly dominate in frequency among reported scientific surveys. The numerical values found among biodiversity estimates reported across the literature were distributed with very large variance, but no patterns were found at the study-level. The variance did appear to decline as the number of included studies (year-by-year) increased, according with the expectation for a random variable. Because so few studies reported anything other than species-richness biodiversity estimates and because no patterns were seen in estimates among studies, meta-analysis found no informative trends. Following the lead of evidence-based medicine in which meta-review plays a central role, it seems that meta-review should provide valuable support for evidence-based conservation of biodiversity. The results of Chapter 3 suggest that this would require some coordination among empiricists to standardise methods and reporting as far as possible. Presently, whilst many commentators repeat that biodiversity is multi-dimensional, reported diversity in practice is overwhelmingly one-dimensional, but heterogeneous in the means of its collection and estimation. Apparently, there is too little desirable variety, and too much undesirable variety.

Having failed to create multidimensional diversity from published estimates, I explored the prospect of being, in principle, able to do this, if suitable data were available. To this end, data was simulated by a boot-strap method of resampling of real data taken from an example of a coastal marine community. The sampling method enabled the statistical properties of natural diversity to be preserved in the simulated data which was replicated into a population of simulated communities. Using this synthesised dataset, I was able to compare a wide variety of indices (many appearing very rarely in empirical literature) and to analyse them as a multidimensional descripton of biodiversity.

The results of the simulated-data study led to two main implications regarding biodiversity estimation. Firstly, three main groups of indicators emerged from mutivariate analysis, clarifying options for policy makers faced with choosing biodiversity indicators. These indicators address different ecosystem properties and comprise of (a) community structure and composition; (b) taxonomic or phylogenetic; and (c) functional diversity. Secondly, and based on this finding, a single scalar metric was suggested which combines the three Principal groups of measures. This new metric can be considered as an optimum estimate of biodiversity in the sense of being necessary and sufficient for capturing the main aspects of biodiversity in as compact a form as possible.

Application of the modelling tools allowed me to compare the performance of that most commonly used biodiversity estimator – species richness – with the information-based metric developed in my study. The results were striking: if species richness is taken to be the sole measure of biodiversity, then a large portion of biodiversity and its variation from one community to another is left unrecognised. This comparison strongly suggests that conceptual equivalence between species richness and biodiversity so commonly assumed, is not valid. Recognising this, recent researchers (e.g., Maclaurin and Sterelny, 2008) have suggested that species richness should be supplemented in various ways to become a good multipurpose measure of biodiversity; used alone, species richness is a poor predictor of the diversity of biological systems. I conclude from the present study that information about the phylogeny and ecological function should be recorded as supplementary information to species identity in compiling biodiversity databases. Using this, a-priori information, the richer multi-dimensional nature of biodiversity can be estimated from field-collected species lists. This way, the biodiversity databases can be used to translate rapid species identity studies, which show little of the intrinsic information resource of the system, into comprehensive and *comparable* estimates of the richness of those communities under study. This will provide a systematic means of comparing the efficacy of biodiversity strategies as we approach 2020.

To summarise, an important contributions of my study is a definition of biodiversity that enables translation of subjective and insubstantial notions of intrinsic value of biodiversity into objective concrete measures. Findings of this work, drawn on the knowledge from different domains, are the foundations of a technique that offers a more substantial method for valuing biological entities. This is done through a measure of information intensity constructed with data derived from biodiversity literature. Furthermore, not only can a formal framework for capturing biodiversity be made, but also a practical suggestion: if only global databases existed with the key species attributes, it would be possible to derive a unifying and comprehensive measure of biodiversity, even with just a species list. As a scalar distance in multivariate space, the resulting simple measure can be used by economists to justify policy shifts designed to conserve biodiversity.

## 6.4 De-emphasising abundance

As explained in Chapter 2, much of the literature on comparative biodiversity concerns the spatial distribution of organisms, a point emphasised by the fact that the distribution of organisms in space appeared in the foundation of the meaning of ecology. This has led to a strong emphasis on the relative abundance of species as the key metric of biodiversity in the literature. The main findings of my study, showing the importance of taxonomic and functional diversity questions the assumption that relative abundance should really be the largest influence on what eventually is called biodiversity. Several authors have introduced the topic by claiming that *intuitively*, a community with roughly equal abundances of species is more diverse than one where a single species makes up 90% of the total (e.g., Hoffmann and Hoffmann, 2008). This "intuition" is expressed quantitatively by the idea of equivalent numbers in entropy and entropy-like indices, where deviation from a uniform abundance distribution is explicitly used to calculate diversity (Jost, 2006).

Whilst premature to dismiss abundance-based measures and their derivatives such as traditional $\beta$-diversities, I certainly would challenge their supremacy as measures of biodiversity. This is important since a rather uncritical assumption seems to have built up, that any spot-measurement of relative abundances is biologically meaningful. Disquiet about such assumptions has recently re-emerged. For example, Magurran and Dornelas (2010) complained that "there is a pervasive view that habitats and assemblages are unchanging rather than acceptance that some change, including local species loss, is inevitable". If species loss is inevitable, how much should we invest in a measurement of relative species abundances at a single point in time – is this really a sound foundation for quantifying biodiversity? As Chapters 1-5 show, this question largely concerns the definition we choose for biodiversity and in particular, whether we mean an arbitrary description of the biological system, or, as I have attempted to, we try to quantify the meaningful information content of the system, recalling that "meaningful" in this context refers to non-transitory functional information, which in turn, I argued was the source of indirect use value. Recent simulations of ecological communities with realistic numbers of interacting species show that large fluctuations of relative abundance are likely to be ubiquitous features of *stable* systems, e.g., Rossberg et al. (2006). This phenomenon is well known to microbiologists, but macroscopic species tend to have such long generation times that their abundance dynamics often look static within the time-frame of a project grant. A few examples exist of datasets, sufficiently long-term to show the population behaviour in natural equilibrium in macroscopic communities. The former-Soviet countries provide some of the best illustrations, e.g., Evstaf'ev (2010) show 60-65 years of co-variation in dynamic of phytoplankton and Baikal omul (*Coregonus migratorius*) found in Baikal lake.

This understanding leads to the conclusion that the large number of definitions for $\beta$-

diversity, found for example in Tuomisto (2010a,b); Anderson et al. (2011), which were designed to describe field data, do not directly describe variation in the functional information diversity of a biological system. On the other hand, $\gamma$-diversity (also primarily designed to summarise field data) is a well defined concept which is compatible with and so may serve as an estimator for the information content coded within the biological system. If, however, $\beta$-diversity is defined following Whittaker (1960) with its original meaning of $\gamma$-/$\alpha$-diversity, then it is useful as a measure of how well the information content of an extensive space has been sampled. It is striking how the large literature concerning $\alpha$-, $\beta$-, and $\gamma$-diversities has concentrated on deriving scores for comparing communities based on the numerical distribution (common or rare) of species among them. Very little has been said about other, perhaps more meaningful, characters of community structure, such as foodweb connectance (Dunne et al., 2002). Genetic and functional variation in space, are even more neglected, other than through the surrogate of species identity. If my results in Chapter 5 are believed, then these are important omissions.

The practical importance of describing and understanding the spatial distribution of diversity, or information, is to be found in spatial planning for conservation. This in practice usually amounts to objective (quantitative) prioritising of landscape (or marine) areas, based on an evaluation of their biological importance or ecological utility (usually implied by one or more measures of biodiversity). Technical decision makers may be guided by a wide range of measures in prioritising areas to conserve, but economic and policy decisions typically require one simple, clear and unambiguous measure to place on the benefits side of the metaphorical scales in cost-benefit analysis. Hitherto, the best, perhaps only option for this has been species richness. Chapter 5 showed how this is only poorly related to the aggregate of multi-dimensional biodiversity, so past and present practice may present seriously misleading valuations to policy makers. A shift of understanding away from species counting, towards information content estimation, may redress this, but to make such an aim a practical reality sets a larger requirement for field data collection than the conventional approach. This however is not as big a problem as it first appears, because most components of bio-information are correlated with species identities and the ecological context created by the community that can be described in terms of a species list. A database containing phylogenetic and functional information for every species encountered would substantially fill the knowledge gap, reducing the field work requirement back to species counting. This is why one of my main recommendations in concluding this thesis is that biodiversity research should include international cooperation to build a comprehensive, accessible, database of species-indexed phylogenetic and functional data. Armed with such a database, the field-ecologist can sample an area, input their species list and obtain an estimate of biological information content based on genetic and functional information. Not only that, but sampling on a grid could turn a species-based biodiversity atlas into a map of biological information density. If, hypothetically, information can be valued in financial currency per bit, then a direct and objective economic value for the biodiversity of an area can be estimated. This was the suggestion with which I concluded Chapter 2.

## 6.5 Returning to the economic context

The original purpose of this work was to find how biodiversity can be represented consistently and unambiguously in cost-benefit analysis. The reason is that the cost-benefit equation (or inequality) formalises the economic means by which public decisions are made and justified in a modern democracy. Because quantitative comparisons require a common currency to be meaningful, biodiversity must be translated into monetary terms – hence the need for valuation.

Most economists have maintained their anthropocentric instrumentalist tradition in dealing with biodiversity (Lee, 2004). Biodiversity is acknowledged as essential, yet diminishing resource, leading to a now well established idea that biodiversity goods and services can be quantified in economic terms. Indeed, the focus has not been on valuing biodiversity as such (Pearce and Moran, 1995; Pearce, 2001), but rather on the economic values generated by resources and/or functions – so-called ecosystem services (Eppink and van den Bergh, 2007). In recent years a sizable literature has build up around different valuation approaches and their application (see, e.g., Pearce and Moran, 1995; Brown and Moran, 1993; Weitzman, 1998; Edwards and Abivardi, 1998; Brock and Xepapadeas, 2003; Christie et al., 2006; Nijkamp et al., 2008).

Neoclassical welfare economics is typically used to evaluate "the ecosystem" assuming that increasing the well-being (or utility) of individuals is the purpose of economic activity and the only role for biodiversity is its contribution to this. Biodiversity loss is therefore assessed in terms of cost-benefit analysis (Freeman, 2003). The substitutability of resources is intrinsic and fundamental to the philosophy of anthropocentric tradition. The value that economists place on biodiversity arises from solving a utility maximisation problem: comparing the consumption of biodiversity with an alternative resource or input. This value is, therefore, characterised as a single unit currency, which is then used to collapse the multiple social values onto a single measure (Bowker, 2004).

Economists typically assess the effect (only) of biodiversity on four services to humanity: material inputs, life support, amenity (including non-use values), and waste receptor services (Freeman, 2003), even though the ecological relations between biodiversity and services are poorly known (see, e.g., Bengtsson et al., 1997). The use-values of services may be revealed in a functioning market, in which case estimation using market-based methods (e.g., production function, replacement cost, etc.) is possible. Frequently (especially for indirect use-values) non-market-based methods of revealed preference are required (e.g., hedonic pricing and the travel cost method). Only hypothetical (stated preference methods), including contingent valuation and choice experiments, can estimate non-use values. Then using a reductionist approach the total economic value of biodiversity is an aggregate of various use and non-use values.

Although all these methods may deliver a useful information to policy makers on the monetary values of the services provided by biodiversity to individuals, there are a number of problems. Generally low level of public awareness and understanding of what biodiversity means (Christie et al., 2006), lead to a conclusion that revealed preferences fail for those biodiversity value categories that the general public is not informed about or has no experience with (Nijkamp et al., 2008). Valuation is not a systematic, market-based exercise, but "rather an ad-hoc search for values to plug into a common cost-benefit framework" (Brown and Moran, 1993). Additionally, none of the methods directly addresses the value of biodiversity itself. For biodiversity to be useful for economics, it needs to satisfy a set of necessary and sufficient criteria. These criteria include uniqueness, quantifiability, and invariance (both to context and observer), none of which are, at present, satisfied by interpreting biodiversity as an economic good.

A fifth service identified by Freeman (2003) introduces the value of biodiversity as a repository of genetic information. Information is seen by some as a primary source of the value in biodiversity, motivating phylogenetic information measurement (Faith, 1992, 1994; Faith et al., 2003), or counting species as information (Weikard, 2002). Closely related is the idea of biodiversity as an insurance against loss of ecosystem services (Baumgartner, 2007); quasi-option value (Arrow and Fisher, 1974; Henry, 1974) is taken to be the appropriate measure in such cases.

Practical economic applications have so far been limited to highly specific contexts (Brock and Xepapadeas, 2003), probably because links between future welfare and biological information are typically obscure. Assessing the direct welfare gained from biodiversity poses substantial problems for cost-benefit analysis since the specific attributes and components of biodiversity must each be identified for valuation. Nehring and Puppe (2002) describe species in terms of attribute sets, but they neglect the interdependence of species and the importance of system-level structures. More seriously, their economic valuation entails a subjective choice of species attributes: given the welfare economic position, these are selected for specific human welfare goals. A more objective approach uses genes as attributes (see Crozier's review, 1997) to generate inter-species distance measures, following the work of (Weitzman, 1992) and its elaboration into the "Noah's Ark Problem" (Weitzman, 1998). Genetic differences are aggregated into a dissimilarity index and it is assumed that the greater the dissimilarity, the more desirable (hence, valuable) the biological system to which they belong.

If the question is limited to one of choosing (from a set) which ecological community is to be preserved, then Weikard's application (2002) of the Noah's Ark problem at the species level can objectively guide decision makers. The “ecosystem” distance measure he proposed is effectively the complementarity measure demonstrated by Faith et al. (2004), but without the need for phylogenetics – an important advantage given our very incomplete knowledge. Weikard (2002) pragmatically replaces ecosystem information content with species counts, whilst acknowledging that the true information store lies at genetic, species, and system levels.

Thus, the usual economic outlook is to see valuation as an aggregation of people’s feeling about what is being valued: economic value derives its meaning from consumer theory. This presents well known problems in the case of natural goods, where the absence of a market precludes revealed preference (market behaviour) valuation. Two alternatives are presented to overcome this. In the first, valuation is taken to be purely subjective and estimated from survey-based stated preferences (e.g., contingent valuation). All the methods used for this suffer from well known biases, several of which arise from problems with defining what is to be valued, others concerning the psychology of subjective valuation. The second approach tries to focus on the instrumental value of biodiversity by identifying and then valuing (e.g., through replacement costs) the “services” rendered. This approach suffers from problems, discussed in Chapter 1, of defining the services and further of finding and quantifying functional relationships between biodiversity and ecosystem services. It is often relatively easy to determine a relation between particular organisms or ecosystem processes and ecosystem services, but proves very difficult when we look specifically at the diversity of systems (Armsworth et al., 2004). Again, we see the problems of definition and the intangibility of diversity, propagating into the valuation problem. The result so far has been a confusing array of different kinds of economic value for different kinds of biodiversity, all of which makes it very difficult to achieve consistent cost-benefit decisions that would be widely supported (Salles, 2011).

Irrespective of the difficulty, economic valuation of biodiversity seems to be necessary for effective conservation. According to the Convention on Biological Diversity’s Conference of the Parties (decision IV/10) “economic valuation of biodiversity and biological resources is an important tool for well-targeted and calibrated economic incentive measures”. This is further supported by the belief that “if we cannot express the value of biodiversity in economic terms, then decision makers will assume that it is unimportant” (see, e.g., Edwards and Abivardi, 1998). It seems that success in achieving biodiversity targets for 2020, directly depends on our understanding and ability to measure biodiversity in economic terms. This is the problem that set the scene for my work.

In Chapter 1, I presented an overview of what value meant in the case of biodiversity, this inevitably including a brief review of economic valuation methods. I concluded that the conventional methods, measuring value indirectly and subjectively, suffer from substantial problems that seem a long way from solution. The remaining option presented was to attempt a direct valuation of biodiversity, in and of itself. This too requires a single and specific definition of biodiversity and in addition, it must refer to a quantitative and intrinsic property of biological systems, rather than an intangible concept. I proposed that, hypothetically, if it were possible to express biodiversity as functional information, based on the elemental principles of information, then it would become a concrete property of the system, which is quantifiable in units of information (bits). If this could be substantiated, then the prospect of directly valuing biodiversity could be a practical solution to the valuation problem. Of course, direct valuation has always been possible in the restricted sense of subjective – opinion based values, but these are not scientific measures. They are not transferable and not a measure of biodiversity, but are only an estimate of public sentiments towards it – subject to changes in fashion and at the individual level, to the description and knowledge offered to the respondent. As subjective valuations, they can never command the status of "hard" measures of costs in the cost-benefit equation. Conversely, measuring the functional information that biodiversity represents would quantify its instrumental value via the functionality that gives rise to ecosystem services. For this reason, my work offers a way to solve the problems of indirect valuation by proposing a route to direct valuation which gives objective results; free from dependence on valuers. This measure may be placed alongside "hard costs" and represents the potential to be valued as an intrinsic property of a biological system. Given such a definition of biodiversity, valuation amounts only to finding an exchange rate between functional information (in general) and financial currencies.

This is a large project, by no means completed by the work reported here. What I have achieved towards it, is to define and explain the concept and steps needed to bring it into practice. I have shown that biodiversity can be understood as information and that the functional fraction of total information can be interpreted as complexity (identified as pattern) and so estimated through the statistical combination of multiple measures of biodiversity from field data. These measures must include representatives of the three main categories of biodiversity shown to be necessary and sufficient description (due to their orthogonality): phylogenetic, functional, and community structure diversity. This has been corroborated by Escarguel et al. (2011), who argued that there is an urgent need to complement the taxonomic dimension of biodiversity with other components including morphological, phylogenetic, and functional diversities. The combination of these into a single distance measure to estimate biological complexity, which is theoretically the foundation of ecosystem function (that in turn being the source of ecosystem services) gives a scalar estimate for use in cost-benefit analysis. The cost of collecting such a rich dataset from the field may be largely overcome by using existing information about the component parts of the system. These are mostly available in a variety of disparate sources – databases, keys, and other published descriptions of organisms. If all these were brought together in a database designed to construct a full picture of system complexity, then the field ecologist would need little more than a comprehensive species list, to be used as an index search of the database. Chapter 3 demonstrated how such a database would be constructed, making use of the logical relations among information entities. The entities were chosen from a matrix analysis of the components of biodiversity – the levels and descriptors. Thus, a formal structure representing biological complexity was operationalised into a computer model (the database design), for which ecological field data is the input and an estimate of biological complexity (i.e., functional information) is the output. A test of this procedure in Chapter 4 showed that phylogenetic, functional, and structural aspects of complexity were independent and each necessary in combination, but it also showed that indicators within each of these three categories were highly correlated and therefore mutually redundant (to a first approximation). The distance measure of the three principal axes showed considerably greater information content (via variation) than species richness alone. I therefore proposed that a composite measure based on the three main axes of biodiversity variation be used as an estimate of the concrete and transferable concept of biodiversity as functional information and that this would be a more justifiable measure to use in cost-benefit analysis for conservation decisions.

## 6.6 Further Progress of the idea

In the first chapter of this work I argued that the current definition of biodiversity is both imprecise and ambiguous. I also asked whether it is possible to quantify biodiversity in a such a way that it will provide grounds for an objective measure of biodiversity value to be used by economists. In this last chapter, I proposed that this can be achieved through developing the theoretical framework for understanding of "biodiversity as information" and constructing value from its functional effect, on the grounds that information causes function and function is the foundation of services. The translation between a concrete measure of functional information and the economic value realised from it has yet to be formulated. If the relationships between biological functions and ecosystem services were known in detail, then the way forward would be clear. Unfortunately, these relations, though better known than those connecting biodiversity to ecosystem services, are not fully described. One avenue of further development obviously lies in strictly defining and quantifying them, even if only in a restricted set of cases.

An important avenue for further research includes finding empirical evidence for correlation between the information-based measure of biodiversity and subjective economic value. This is useful because the gradient of this, as yet hypothetical, correlation could serve as an exchange-rate to translate between units of functional information and financial currency. A large meta-analysis of economic valuation literature would be needed to find the correlation, if it exists. The method developed here for calculating functional information estimates would have to be applied to a statistical population of real systems for which economic values have been published. This would constitute a new evidence-based study in ecological economics with direct policy implications.

This work has certainly highlighted the importance of functional and genetic diversity for understanding the potential for value. Neglect of these kinds of variety in favour of community composition seems to be missing important information that if it were quantified, would likely add to our motivation to conserve. Even though a few economists have proposed measures based on phylogenetic information (originating with the Noah's Ark problem), there has been little take-up of these ideas among ecologists. There remains no equivalent for functional information, which seems to be at an early stage of conceptual development. These themes may now be best developed through worked examples focussing on a particular system, using the principles developed here. The relatively high correlation among structural indices indicates that more effort spent on phylogenetic or functional information would yield greater rewards than concentration on different measures of abundance heterogeneity. Thus a system for which a great deal is already known about genetic composition and functional traits would be needed for such a worked example study.

The disappointing results of efforts to build composite biodiversity measures from published literature point to the need for an organisational overview for biodiversity research, with the aim of building a rich database of system diversities. This is different from the multitude of biodiversity databases currently available, which collect only species lists, sometimes with taxonomic details. A global database which organises all that is known of described species (including phylogenetic details) into functional information categories would be of great value in constructing a more comprehensive picture of biodiversity. This sets a goal for bioinformatics and the deployment of relational database systems in the service of global biodiversity knowledge. The present work demonstrates a prototype for this, which may be taken as a working proposal. Inclusion of geo-referenceing would elevate such an information system to the level needed to support the very ambitious goals of the CBD.

Further research should draw inspiration from my finding that biodiversity is indeed multidimensional and complex, but that it is not intractably so.

# Thesis Summary and Recommendations

1. **Essential Summary**. This thesis set out to find a comprehensive single-valued measure of biodiversity from which to derive an objective measure of value. The great variety of existing biodiversity measures was analysed through a decomposition into levels and descriptors, from which an information-maximising measure was sought. Meta-analysis of existing biodiversity literature failed to reveal patterns which could guide this aim. Instead, ecological communities were synthesized from resampling of

the combined data from three biodiversity data-sets representing species composition, taxonomy and ecological function, to create a population of realistic communities. A battery of biodiversity metrics, including genetic and functional measures was calculated for all the synthetic communities, from which it was found that three axes of diversity: community composition, taxonomy and function were necessary and sufficient to describe biodiversity comprehensively and that species richness captured only about one fifth of this. A theoretical argument accompanying this analysis demonstrated that biodiversity interpreted as functional information is the source of indirect use value and therefore the basis for objective valuation of biodiversity, which has little if any direct use. A relational database was design and prototyped to enable maximum use of existing knowledge for constructing comprehensive estimates of biodiversity for objective valuation. It is recommended that this template is used to create a global information resource to inform and enable quantitative achievement of international biodiversity targets.

2. **Formal Definition**. Biodiversity can be defined as an estimator of the functional information content of a biological system, where functional information is the information meeting both of the following criteria: (a) it constitutes a pattern of difference within the system and (b) the pattern has a consequence for ecological function. Pattern is defined here as non-transitory, compressible information, identified as the difference between the total information content and the Kolmogorov complexity of the system. Consequence means that a change in functional information results in a change of ecological function.

3. **Quantification**. Following this definition, biodiversity is contributed from every level of biological organisation from molecular to whole system. This results in it being multi-dimensional, in the sense that multiple descriptors at multiple levels combine to create biodiversity at all levels of system above the molecule. This work showed how biodiversity can be deconstructed and organised into a permutation matrix of descriptor × level measures, which may then be combined into arbitrary aggregating indices.

4. **Dimensionality**. Empirical analysis of simulated communities showed that three principle axes measure most of biodiversity, These are taxonomic (a surrogate for phylogeny), functional, and compositional (a surrogate for structural) diversity. Biodiversity is well represented in the three-dimensional space of these axes. When distances among simulated communities were measured in this space, they described more than four times the diversity content than species richness alone.

5. **Recommendation on Diversity Survey Standardisation**. Meta-review of biodiversity literature revealed lack of standardisation in definitions and methods such that patterns and inter-comparability were effectively impossible. Formal agreement

on basic measurement of biodiversity is needed to integrate the results of multiple studies. If achieved, this will significantly strengthen efforts to achieve quantitative targets of international biodiversity agreements.

6. **Recommendation for Global Biodiversity Database**. The difficulty of directly measuring comprehensive biodiversity in a single field study increases the importance of collating information and maintaining it in an openly accessible form so that basic field data can be augmented with existing knowledge. This requires a global relational database which combines phylogenetic, functional and community-level structural data. A design and prototype database were developed and described in this thesis.

7. **Recommendation for Valuation**. The work of this thesis offers a new alternative to operationalise biodiversity at a policy level. It does so by suggesting a way to translate from ecological field data, into a single valued property of the system which summarises its functional diversity. It offers a theoretical argument demonstrating that the functional diversity estimated by this measure is the biological source of indirect-use value and so is the objective basis for economic value of biodiversity in and of itself.

# References

Adami, C., Ofria, C., and Collier, T. C. (2000). Evolution of biological complexity. *Proceedings of the National Academy of Sciences of the United States of America*, 97(9):4463–4468.

Agrawal, R., Imielinski, T., and Swami, A. (1993). Database mining: a performance perspective. *Knowledge and Data Engineering, IEEE Transactions on*, 5(6):914 –925.

Alexandrou, M. A., Oliveira, C., Maillard, M., McGill, R. A. R., Newton, J., Creer, S., and Taylor, M. I. (2011). Competition and phylogeny determine community structure in mullerian co-mimics. *Nature*, 469(7328):84–88.

Anderson, M. J. (2006). Distance-based tests for homogeneity of multivariate dispersions. *Biometrics*, 62(1):245–253.

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., and Swenson, N. G. (2011). Navigating the multiple meanings of beta diversity: a roadmap for the practicing ecologist. *Ecology Letters*, 14(1):19–28.

Anderson, T. M. (2008). Plant compositional change over time increases with rainfall in serengeti grasslands. *Oikos*, 117:675–682.

Armsworth, P. R., Kendall, B. E., and Davis, F. W. (2004). An introduction to biodiversity concepts for environmental economists. *Resource and Energy Economics*, 26(2):115–136.

Arnqvist, G. and Wooster, D. (1995). Meta-analysis: synthesizing research findings in ecology and evolution. *Trends in Ecology and Evolution*, 10(6):236–240.

Arrow, K. J. and Fisher, A. (1974). Environmental preservation, uncertainty and irreversibility. *Quarterly Journal of Economics*, 89:312–319.

Attele, A. S., Wu, J. A., and Yuan, C.-S. (1999). Ginseng pharmacology: Multiple constituents and multiple actions. *Biochemical Pharmacology*, 58(11):1685–1693.

Aubert, M., Alard, D., and Bureau, F. (2003). Diversity of plant assemblages in managed temperate forests: a case study in Normandy (France). *Forest Ecology and Management*, 175(1-3):321–337.

Balmford, A., Bennun, L., ten Brink, B., Cooper, D., Côté, I. M., Crane, P., Dobson, A., Dudley, N., Dutton, I., Green, R. E., Gregory, R. D., Harrison, J., Kennedy, E. T., Kremen, C., Leader-Williams, N., Lovejoy, T. E., Mace, G., May, R., Mayaux, P., Morling, P., Phillips, J., Redford, K., Ricketts, T. H., Rodríguez, J. P., Sanjayan, M., Schei, P. J., van Jaarsveld, A. S., and Walther, B. A. (2005). The convention on biological diversity's 2010 target. *Science*, 307(5707):212–213.

Balmford, A., Lyon, A. J. E., and Lang, R. M. (2000). Testing the higher-taxon approach to conservation planning in a megadiverse group: the macrofungi. *Biological Conservation*, 93(2):209–217.

Balvanera, P., Kremen, C., and Martinez-Ramos, M. (2005). Applying community structure analysis to ecosystem function: Examples from pollination and carbon storage. *Ecological Applications*, 15(1):360–375.

Balvanera, P., Pfisterer, A. B., Buchmann, N., He, J. S., Nakashizuka, T., Raffaelli, D., and Schmid, B. (2006). Quantifying the evidence for biodiversity effects on ecosystem functioning and services. *Ecology Letters*, 9(10):1146–1156.

Bar-Yam, Y. (2004). Multiscale variety in complex systems. *Complexity*, 9(4):37–45.

Barbieri, M. (2007). *Biosemiotics: Information, Codes and Signs in Living Systems*. New York: Nova Science Publishers.

Barlow, J., Gardner, T. A., Araujo, I. S., Avila-Pires, T. C., Bonaldo, A. B., Costa, J. E., Esposito, M. C., Ferreira, L. V., Hawes, J., Hernandez, M. M., Hoogmoed, M. S., Leite, R. N., Lo-Man-Hung, N. F., Malcolm, J. R., Martins, M. B., Mestre, L. A. M., Miranda-Santos, R., Nunes-Gutjahr, A. L., Overal, W. L., Parry, L., Peters, S. L., Ribeiro-Junior, M. A., da Silva, M. N. F., Motta, C. d. S., and Peres, C. A. (2007). Quantifying the biodiversity value of tropical primary, secondary, and plantation forests. *Proceedings of the National Academy of Sciences of the United States of America*, 104(47):18555–18560.

Bates, M. (2005). Information and knowledge: an evolutionary framework for information science. *Information Research*, 10(4).

Bateson, G. (1972). Form, substance, and difference. In Bateson, G., editor, *Steps to an Ecology of Mind*, pages 448–466. University of Chicago Press.

Baumgartner, S. (2006). The ecological economics of biodiversity methods and policy applications. *Ecological Economics*, 59(1):181–182.

Baumgartner, S. (2007). The insurance value of biodiversity in the provision of ecosystem services. *Natural Res. Modeling*, 20:87–127.

Beaumont, N. J., Austen, M. C., Atkins, J. P., Burdon, D., Degraer, S., Dentinho, T. P., Derous, S., Holm, P., Horton, T., van Ierland, E., Marboe, A. H., Starkey, D. J., Townsend, M., and Zarzycki, T. (2007). Identification, definition and quantification of goods and services provided by marine biodiversity: Implications for the ecosystem approach. *Marine Pollution Bulletin*, 54(3):253–265.

Begon, M., Townsend, C., and Harper, J. (2006). *Ecology: From individuals to ecosystems*. Oxford: Blackwell, 4 edition.

Béné, C. and Doyen, L. (2008). Contribution values of biodiversity to ecosystem performances: A viability perspective. *Ecological Economics*, 68(1-2):14–23.

Bengtsson, J. (1998). Which species? What kind of diversity? Which ecosystem function? Some problems in studies of relations between biodiversity and ecosystem function. *Applied Soil Ecology*, 10(3):191–199.

Bengtsson, J., Jones, H., and Setl, H. (1997). The value of biodiversity. *Trends in Ecology and Evolution*, 12(9):334–336.

Berger, W. H. and Parker, F. L. (1970). Diversity of planktonic foraminifera in deep-sea sediments. *Science*, 168(3937):1345–1347.

Bevilacqua, S., Fraschetti, S., Terlizzi, A., and Boero, F. (2009). The use of taxonomic distinctness indices in assessing patterns of biodiversity in modular organisms. *Marine Ecology-an Evolutionary Perspective*, 30(2):151–163.

Bhagwat, S. A., Dudley, N. N., and Harrop, S. R. (2011). Religious following in biodiversity hotspots: challenges and opportunities for conservation and development. *Conservation Letters*, 4(3):234–240.

BIOTIC (2010). The biological traits information catalogue. Digital resource at http://www.marlin.ac.uk.

Bisby, F., Roskov, Y., Orrell, T., Nicolson, D., Paglinawan, L., Bailly, N., Kirk, P., Bourgoin, T., Baillargeon, G., and eds (2009). Species 2000 and ITIS Catalogue of Life: 2009 Annual Checklist. Digital resource at www.catalogueoflife.org/annual-checklist/2009/ Species 2000: Reading, UK.

Bitbol, M. and Luisi, P. (2004). Autopoiesis with or without cognition: defining life at its edge. *Journal of the Royal Society Interface*, 1(1):99–107.

Bongers, T., Alkemade, R., and Yeates, G. (1991). Interpretation of disturbance induced maturity decrease in marine nematode assemblages by means of the maturity index. *Marine Ecology Progress Series*, 76:135–142.

Botta-Dukát, Z. and Wilson, J. B. (2005). Rao's quadratic entropy as a measure of functional diversity based on multiple traits. *Journal of Vegetation Science*, 16(5):533–540.

Bowker, G. C. (2004). *Global Assemblages*, chapter Time, money and biodiversity, pages 107–123. Blackwell, Oxford, UK.

Brander, L. M., Van Beukering, P., and Cesar, H. S. J. (2007). The recreational value of coral reefs: A meta-analysis. *Ecological Economics*, 63(1):209–218.

Breusch, T. and Pagan, A. (1979). A simple test for a simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(1287–1294).

Brock, W. A. and Xepapadeas, A. (2003). Valuing biodiversity from an economic perspective: a unified economic, ecological, and genetic approach. *The American Economic Review*, 93(5):1597–1614.

Brown, K. and Moran, D. (1993). Valuing biodiversity: the scope and limitations of economic analysis. Technical report, Centre for Social and Economic Research on the Global Environment, University of East Anglia and University College London.

Brown, L. R. (1988). Biodiversity. chapter And today we're going to talk about biodiversity...that's right, biodiversity., pages 446–449. National Academy Press, Washington, D.C.

Camargo, J. A. (1992). Can dominance influence stability in competitive interactions? *Oikos*, 64(3):605–609.

Campos, D. and Fernando Isaza, J. (2009). A geometrical index for measuring species diversity. *Ecological Indicators*, 9(4):651–658.

Cano, J. M., Mäkinen, H. S., Leinonen, T., Freyhof, J., and Meriä, J. (2008). Extreme neutral genetic and morphological divergence supports classification of Adriatic three-spined stickleback (Gasterosteus aculeatus) populations as distinct conservation units. *Biological Conservation*, 141(4):1055–1066.

Cardinale, B. J., Srivastava, D. S., Duffy, J. E., Wright, J. P., Downing, A. L., Sankaran, M., and Jouseau, C. (2006). Effects of biodiversity on the functioning of trophic groups and ecosystems. *Nature*, 443(7114):989–992.

Caro, T. M. and O'Doherty, G. (1999). On the use of surrogate species in conservation biology. *Conservation Biology*, 13(4):805–814.

Caron-Lormier, G., Bohan, D. A., Hawes, C., Raybould, A., Haughton, A. J., and Humphry, R. W. (2009). How might we model an ecosystem? *Ecological Modelling*, 220(17):1935–1949.

Celko, J. (2004). *Joe Celko's Trees and Hierarchies in SQL for Smarties.* Morgan Kaufmann.

Certain, G., Skarpaas, O., Bjerke, J. W., Framstad, E., Lindholm, M., Nilsen, J. E., Norderhaug, A., Oug, E., Pedersen, H. C., Schartau, A. K., van der Meeren, G. I., Aslaksen, I., Engen, S., Garnasjordet, P. A., Kvaloy, P., Lillegard, M., Yoccoz, N. G., and Nybo, S. (2011). The nature index: A general framework for synthesizing knowledge on the state of biodiversity. *PLoS ONE*, 6(4).

Chaitin, G. (1990). *Information, Randomness and Incompleteness - Papers on Algorithmic Information Theory*, volume 8 of *Series in Computer Science.* World Scientific, Singapore, 2nd edition.

Champely, S. and Chessel, D. (2002). Measuring biological diversity using euclidean metrics. *Environmental and Ecological Statistics*, 9(2):167–177.

Chao, A., Chazdon, R. L., Colwell, R. K., and Shen, T. J. (2005). A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters*, 8(2):148–159.

Chargaff, E. (1978). *Heraclitean fire. Sketches from a Life before Nature.* The Rockefeller university press.

Chen, L., Xu, L., and Huang, H. (2007). Genetic diversity and population structure in vallisneria spinulosa (hydrocharitaceae). *Aquatic Botany*, 86(1):46–52.

Chen, W. (1976). The entity-relationship model-toward a unified view of data. In *ACM Transactions on Database Systems (TODS)*, pages 847–851.

Christie, M., Hanley, N., Warren, J., Murphy, K., Wright, R., and Hyde, T. (2006). Valuing the diversity of biodiversity. *Ecological Economics*, 58(2):304–317.

Clarke, K. R. and Warwick, R. M. (1998). A taxonomic distinctness index and its statistical properties. *Journal of Applied Ecology*, 35(4):523–531.

Clarke, K. R. and Warwick, R. M. (2001). A further biodiversity index applicable to species lists: variation in taxonomic distinctness. *Marine Ecology Progress Series*, 216:265–278.

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, 13:377–387.

Cohen, J. E., Jonsson, T., and Carpenter, S. R. (2003). Ecological community description using the food web, species abundance, and body size. *Proceedings of the National Academy of Sciences of the United States of America*, 100(4):1781–1786.

Colwell, R. K. and Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 345(1311):101–118.

Colyvan, M., Linquist, S., Grey, W., Griffiths, P. E., Odenbaugh, J., and Possingham, H. (2009). Philosophical issues in ecology: Recent trends and future directions. *Ecology and Society*, 14(2).

COP10 (2010). Tenth meeting of the conference of the parties to the convention on biological diversity. http://www.cbd.int/cop10/doc/.

Cornwell, W. K., Schwilk, D. W., and Ackerly, D. D. (2006). A trait-based test for habitat filtering: convex hull volume. *Ecology*, 87(6):1465–1471.

Cowling, R., Knight, A., Faith, D., Ferrier, S., Lombard, A., Driver, A., Rouget, M., Maze, K., and Desmet, P. (2004). Nature conservation requires more than a passion for species. *Conservation Biology*, 18(6):1674–1676.

Crist, E. (2002). Quantifying the biodiversity crisis. *Wild earth.*

Crozier, R. H. (1997). Preserving the information content of species: Genetic diversity, phylogeny, and conservation worth. *Annual Review of Ecology and Systematics*, 28:243–268.

Crozier, R. H., Dunnett, L. J., and Agapow, P.-M. (2005). Phylogenetic biodiversity assessment based on systematic nomenclature. *Evolutionary Bioinformatics*, 2005.

Daily, G. and Dasgupta, S. (2007). *Encyclopedia of Biodiversity*, volume 2, chapter Concept of ecosystem services, pages 353–362. Academic Press, 2nd edition.

Danovaro, R. and Pusceddu, A. (2007). Biodiversity and ecosystem functioning in coastal lagoons: Does microbial diversity play any role? *Estuarine, Coastal and Shelf Science*, 75(1-2):4–12.

de Groot, R., Alkemade, R., Braat, L., Hein, L., and Willemen, L. (2010). Challenges in integrating the concept of ecosystem services and values in landscape planning, management and decision making. *Ecological Complexity*, 7(3):260 – 272. Ecosystem Services - Bridging Ecology, Economy and Social Sciences.

de Ruiter, P. C., Wolters, V., and Moore, J. C., editors (2005). *Dynamic food webs. Multispecies assembalges, ecosystem development and environmental change.* Elsevier.

Dean, J., van Dooren, K., and Weinstein, P. (2011). Does biodiversity improve mental health in urban settings? *Medical Hypotheses*, 76(6):877–880.

DeLong, D. C. (1996). Defining biodiversity. *Wildlife society bulletin*, 24(4):738–749.

Desrochers, R. E. and Anand, M. (2004). From traditional diversity indices to taxonomic diversity indices. *International Journal of Ecology and Environmental Sciences*, 30:85–92.

Diaz, S. and Cabido, M. (2001). Vive la difference: plant functional diversity matters to ecosystem processes. *Trends in Ecology and Evolution*, 16(11):646–655.

Dickersin, K., Scherer, R., and Lefebvre, C. (1994). Systematic reviews - identifying relevant studies for systematic reviews. *British Medical Journal*, 309(6964):1286–1291.

Dixon, P. M. (2002). *Bootstrap Resampling*, volume 1, pages 212–220. Wiley.

Dunne, J., Williams, R., and Martinez, N. (2002). Food-web structure and network theory: The role of connectance and size. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12917–12922.

Edwards, P. J. and Abivardi, C. (1998). The value of biodiversity: Where ecology and economy blend. *Biological Conservation*, 83(3):239–246.

Egger, M., Smith, G. D., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315(7109):629–634.

Ehrenfeld, D. (1988). Biodiversity. Number 24, chapter Why put a value on biodiversity?, pages 212–216. National Academy Press, Washington, D.C.

Elliot, M. J., Bull, H. T., Pulford, C. I., Shadbolt, N. R., and Smith, W. (1995). Constructive knowledge engineering. *Knowledge-Based Systems*, 8(5):259 – 267.

Englund, G., Sarnelle, O., and Cooper, S. D. (1999). The importance of data-selection criteria: meta-analyses of stream predation experiments. *Ecology*, 80(4):1132–1141.

Eppink, F. V. and van den Bergh, J. C. J. M. (2007). Ecological theories and indicators in economic models of biodiversity loss and conservation: A critical review. *Ecological Economics*, 61(2-3):284–293.

Erwin, P. M., López-Legentil, S., and Schuhmann, P. W. (2010). The pharmaceutical value of marine biodiversity for anti-cancer drug discovery. *Ecological Economics*, 70(2):445–451.

Escarguel, G., Fara, E., Brayard, A., and Legendrer, S. (2011). Biodiversity is not (and never has been) a bed of roses! *Comptes Rendus Biologies*, 334(5-6):351 – 359. Biodiversity in face of human activities / La biodiversite face aux activites humaines.

Evstaf'ev, V. K. (2010). Analiz mnogoletnej dinamiki osnovnyh zven'ev troficheskoj seti v pelagiali ozera bajkal. *Izvestija Irkutskogo Gosudarstvennogo Universiteta*, 3(1):3–11.

Failing, L. and Gregory, R. (2003). Ten common mistakes in designing biodiversity indicators for forest policy. *Journal of Environmental Management*, 68(2):121 – 132.

Faith, D., Reid, C., and Hunter, J. (2004). Integrating phylogenetic diversity, complementarity, and endemism for conservation assessment. *Conservation Biology*, 18(1):255–261.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61(1):1–10.

Faith, D. P. (1994). Phylogenetic pattern and the quantification of organismal biodiversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 345(1311):45–58.

Faith, D. P., Carter, G., Cassis, G., Ferrier, S., and Wilkie, L. (2003). Complementarity, biodiversity viability analysis, and policy-based algorithms for conservation. *Environmental Science and Policy*, 6(3):311–328.

Feest, A., Aldred, T. D., and Jedamzik, K. (2010). Biodiversity quality: A paradigm for biodiversity. *Ecological Indicators*, 10(6):1077–1082.

Feld, C. K., da Silva, P. M., Sousa, J. P., de Bello, F., Bugter, R., Grandin, U., Hering, D., Lavorel, S., Mountford, O., Pardo, I., Partel, M., Rombke, J., Sandin, L., Jones, K. B., and Harrison, P. (2009). Indicators of biodiversity and ecosystem services: a synthesis across ecosystems and spatial scales. *Oikos*, 118(12):1862–1871.

Felton, A., Knight, E., Wood, J., Zammit, C., and Lindenmayer, D. (2010). A meta-analysis of fauna and flora species richness and abundance in plantations and pasture lands. *Biological Conservation*, 143(3):545–554.

Ferrero, T. J., Debenham, N. J., and Lambshead, P. J. D. (2008). The nematodes of the thames estuary: Assemblage structure and biodiversity, with a test of attrill's linear model. *Estuarine, Coastal and Shelf Science*, 79(3):409–418.

Fischer, A., Bednar-Friedl, B., Langers, F., Dobrovodská, M., Geamana, N., Skogen, K., and Dumortier, M. (2011). Universal criteria for species conservation priorities? findings from a survey of public views across europe. *Biological Conservation*, 144(3):998–1007.

Fischer, A. and Young, J. C. (2007). Understanding mental constructs of biodiversity: Implications for biodiversity management and conservation. *Biological Conservation*, 136(2):271–282.

Fisher, B. and Christopher, T. (2007). Poverty and biodiversity: Measuring the overlap of human poverty and the biodiversity hotspots. *Ecological Economics*, 62(1):93–101.

Fleishman, E., Noss, R. F., and Noon, B. R. (2006). Utility and limitations of species richness metrics for conservation planning. *Ecological Indicators*, 6(3):543–553.

Floridi, L. (2003). Information. In Floridi, L., editor, *The Blackwell Guide to the Philosophy of Computing and Information*, pages 40–61. Blackwell Publishing Ltd.

Floridi, L. (2005). Is semantic information meaningful data? *Philosophy and Phenomenological Research*, 70(2):351–370.

Floridi, L. (2011). Semantic conceptions of information. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Spring 2011 edition.

Fox, J. (2011). Formalizing knowledge and expertise: where have we been and where are we going? *The Knowledge Engineering Review*, 26(Special Issue 01):5–10.

Franklin, J. (1988). Biodiversity. chapter Structural and functional diversity in temperate forests., pages 166–175. National Academy Press, Washington, D.C.

Freeman, M. I. (2003). *The Measurement of Environmental and Resource Values: Theory and Methods*. Resources for the Future, Washington, DC, 2nd edition.

Gallardo, B., Gascón, S., Quintana, X., and Comín, F. A. (2011). How to choose a biodiversity indicator - redundancy and complementarity of biodiversity metrics in a freshwater ecosystem. *Ecological Indicators*, In Press, Corrected Proof.

Gambi, C., Vanreusel, A., and Danovaro, R. (2003). Biodiversity of nematode assemblages from deep-sea sediments of the atacama slope and trench (south pacific ocean). *Deep Sea Research Part I: Oceanographic Research Papers*, 50(1):103–117.

Gamito, S. and Furtado, R. (2009). Feeding diversity in macroinvertebrate communities: A contribution to estimate the ecological status in shallow waters. *Ecological Indicators*, 9(5):1009–1019.

Garcia-Molina, H., Ullman, J. D., and Widom, J. (2008). *Database Systems: The Complete Book (2nd Edition)*. Prentice Hall.

Gaston, K. (1996). *Biodiversity: a biology of numbers and difference*. Blackwell Science.

Gates, S. (2002). Review of methodology of quantitative reviews using meta-analysis in ecology. *Journal of Animal Ecology*, 71(4):547–557.

GBIF (2010). Global Biodiversity Information Facility. http://www.gbif.org/.

Gell-Mann, M. and Lloyd, S. (1996). Information measures, effective complexity, and total information. *Complexity*, 2(1):44–52.

Gell-Mann, M. and Lloyd, S. (2003). Effective complexity. In Gell-Mann, M. and Tsallis, C., editors, *Nonextensive Entropy - Interdisciplinary Applications*. Oxford University Press.

Ghilarov, A. (1996). What does 'biodiversity' mean – scientific problem or convenient myth? *Trends in Ecology and Evolution*, 11(7):304–306.

Glowka, L., Burhenne-Guilmin, F., and Synge, H. (1994). *A Guide to the Convention on Biological Diversity*. Gland: IUCN.

Goldstein, P. Z. (1999). Functional ecosystems and biodiversity buzzwords. *Conservation Biology*, 13(2):247–255.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3/4):237–264.

Gowdy, J. M. (2000). Terms and concepts in ecological economics. *Wildlife Society Bulletin*, 28(1):26–33.

Gower, J. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871.

Grantham, H. S., Wilson, K. A., Moilanen, A., Rebelo, T., and Possingham, H. P. (2009). Delaying conservation actions for improved knowledge: how long should we wait? *Ecology Letters*, 12(4):293–301.

Graudal, N. A., Galloe, A. M., and Garred, P. (1998). Effects of sodium restriction on blood pressure, renin, aldosterone, catecholamines, cholesterols, and triglyceride - a meta-analysis. *Jama-Journal Of The American Medical Association*, 279(17):1383–1391.

Graybeal, A. (1998). Is it better to add taxa or characters to a difficult phylogenetic problem? *Systematic Biology*, 47(1):9–17.

Gregorius, H.-R. and Gillet, E. M. (2008). Generalized simpson-diversity. *Ecological Modelling*, 211(1-2):90–96.

Gurevitch, J. and Hedges, L. V. (1999). Statistical issues in ecological meta-analyses. *Ecology*, 80(4):1142–1149.

Hamilton, A. J. (2005). Species diversity or biodiversity? *Journal of Environmental Management*, 75(1):89–92.

Harrop, S. R. and Pritchard, D. J. (2011). A hard instrument goes soft: The implications of the convention on biological diversity's current trajectory. *Global Environmental Change*, 21(2):474–480.

He, F. and Hubbell, S. P. (2011). Species-area relationships always overestimate extinction rates from habitat loss. *Nature*, 473(7347):368–371.

Heino, J. (2008). Patterns of functional biodiversity and function-environment relationships in lake littoral macroinvertebrates. *Limnology and Oceanography*, 53:1446–1455.

Henry, C. (1974). Option values in the economics of irreplaceable assets. *The Review of Economic Studies*, 41(89–104).

Henry, L.-A. and Roberts, J. M. (2007). Biodiversity and ecological composition of macrobenthos on cold-water coral mounds and adjacent off-mound habitat in the bathyal porcupine seabight, ne atlantic. *Deep Sea Research Part I: Oceanographic Research Papers*, 54(4):654–672.

Henry, P. Y., Lengyel, S., Nowicki, P., Julliard, R., Clobert, J., Celik, T., Gruber, B., Schmeller, D., Babij, V., and Henle, K. (2008). Integrating ongoing biodiversity monitoring: potential benefits and methods. *Biodiversity And Conservation*, 17(14):3357–3382.

Hillis, D. M. (1998). Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Systematic Biology*, 47(1):3–8.

Hoffmann, S. and Hoffmann, A. (2008). Is there a "true" diversity? *Ecological Economics*, 65(2):213–215.

Hooper, D. U., Chapin, F. S., Ewel, J. J., Hector, A., Inchausti, P., Lavorel, S., Lawton, J. H., Lodge, D. M., Loreau, M., Naeem, S., Schmid, B., Setala, H., Symstad, A. J., Vandermeer, J., and Wardle, D. A. (2005). Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. *Ecological Monographs*, 75(1):3–35.

Hoover, K. D. and Perez, S. J. (1999). Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *Econometrics Journal*, 2:167–191.

Hubalek, Z. (2000). Measures of species diversity in ecology: an evaluation. *Folia Zoologica*, 49(4):241–260.

Huggett, A. J. (2005). The concept and utility of 'ecological thresholds' in biodiversity conservation. *Biological Conservation*, 124(3):301–310.

Izsak, J. and Papp, L. (2000). A link between ecological diversity indices and measures of biodiversity. *Ecological Modelling*, 130:151–156.

Jiang, Y., Kang, M., Zhu, Y., and Xu, G. (2007). Plant biodiversity patterns on helan mountain, china. *Acta Oecologica*, 32(2):125–133.

Jones, J. P. G., Collen, B., Atkinson, G., Baxter, P. W. J., Bubb, P., Illian, J. B., Katzner, T. E., Keane, A., Loh, J., Mcdonald-Madden, E., Nicholson, E., Pereira, H. M., Possingham, H. P., Pullin, A. S., Rodrigues, A. S. L., Ruiz-Gutierrez, V., Sommerville, M., and Milner-Gulland, E. J. (2011). The why, what, and how of global biodiversity indicators beyond the 2010 target. *Conservation Biology*, 25(3):450–457.

Joshi, P. C., Kumar, K., and Arya, M. (2008). Assessment of insect diversity along an altitudinal gradient in pinderi forests of western himalaya, india. *Journal of Asia-Pacific Entomology*, 11(1):5–11.

Jost, L. (2006). Entropy and diversity. *Oikos*, 113(2):363–375.

Kim, J. (1996). General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing numbers of taxa. *Systematic Biology*, 45(3):363–374.

Kim, K. C. and Byrne, L. B. (2006). Biodiversity loss and the taxonomic bottleneck: emerging biodiversity science. *Ecological Research*, 21(6):794–810.

King, I. (2009). The need for the incorporation of phylogeny in the measurement of biological diversity, with special reference to ecosystem functioning research. *BioEssays*, 31(1):107–116.

Koleff, P., Gaston, K. J., and Lennon, J. J. (2003). Measuring beta diversity for presence-absence data. *Journal of Animal Ecology*, 72(3):367–382.

Kosman, E. (2003). Nei's gene diversity and the index of average differences are identical measures of diversity within populations. *Plant Pathology*, 52(5):533–535.

Krebs, C. (1972). *Ecology: The experimental analysis of distribution and abundance*. Harper and Row, New York, USA.

Krebs, C. (1998). *Ecological Methodology (2nd Edition)*. Benjamin Cummings.

Laliberté, E. and Legendre, P. (2010). A distance-based framework for measuring functional diversity from multiple traits. *Ecology*, 91:299–305.

Laliberté, E. and Shipley, B. (2010). *FD: measuring functional diversity from multiple traits, and other tools for functional ecology*. R package version 1.0-9.

Lambertini, C., Gustafsson, M. H. G., Frydenberg, J., Speranza, M., and Brix, H. (2008). Genetic diversity patterns in phragmites australis at the population, regional and continental scales. *Aquatic Botany*, 88(2):160–170.

Lecerf, A. and Richardson, J. S. (2010). Biodiversity - ecosystem function research: insights gained from streams. *River Research and Applications*, 26(1):45–54.

Lee, K. (2004). There is biodiversity and biodiversity. In Oksanen, M, P. J., editor, *Philosophy and Biodiversity*, pages 152–171. Cambridge University Press, Cambridge, UK.

Legendre, P. and Legendre, L. (1998). *Numerical Ecology, Volume 20, Second Edition (Developments in Environmental Modelling).* Elsevier Science.

Looijen, R. C. and van Andel, J. (1999). Ecological communities: conceptual problems and definitions. *Perspectives in Plant Ecology, Evolution and Systematics*, 2(2):210–222.

Loreau, M., Naeem, S., Inchausti, P., Bengtsson, J., Grime, J. P., Hector, A., Hooper, D. U., Huston, M. A., Raffaelli, D., Schmid, B., Tilman, D., and Wardle, D. A. (2001). Biodiversity and Ecosystem Functioning: Current Knowledge and Future Challenges. *Science*, 294(5543):804–808.

Mace, G. M. and Baillie, J. E. M. (2007). The 2010 biodiversity indicators: Challenges for science and policy. *Conservation Biology*, 21(6):1406–1413.

Mace, G. M., Cramer, W., DÌaz, S., Faith, D. P., Larigauderie, A., Le Prestre, P., Palmer, M., Perrings, C., Scholes, R. J., Walpole, M., Walther, B. A., Watson, J. E., and Mooney, H. A. (2010). Biodiversity targets after 2010. *Current Opinion in Environmental Sustainability*, 2(1-2):3–8.

Mace, G. M., Gittleman, J. L., and Purvis, A. (2003). Preserving the tree of life. *Science*, 300(5626):1707–1709.

Maclaurin, J. and Sterelny, K. (2008). *What is biodiversity?* The University of Chicago Press.

Magurran, A. (2004). *Measuring Biological Diversity.* Blackwell Publishing.

Magurran, A. E. and Dornelas, M. (2010). Biological diversity in a changing world. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 365(1558):3593–3597.

Mallet, J. (2007). *Encyclopedia of Biodiversity*, volume 5, chapter Concept of species, pages 427–440. Academic Press, 2nd edition.

Marcot, B. G. (2007). Biodiversity and the lexicon zoo. *Forest Ecology and Management*, 246(1):4–13.

Margalef, D. (1958). Information theory in ecology. *General Systems Yearbook*, 3:36–71.

Mason, N., Mouillot, D., Lee, W., and Wilson, J. (2005). Functional richness, functional evenness and functional divergence: The primary components of functional diversity. *Oikos*, 111(1):112–118.

Mason, N. W. H., MacGillivray, K., Steel, J. B., and Wilson, J. B. (2003). An index of functional diversity. *Journal of Vegetation Science*, 14(4):571–578.

Maturana, H. R. and Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living*, volume 42. D. Reidel Pub. Co. (Dordrecht, Holland and Boston).

May, R. M. (1994). Conceptual aspects of the quantification of the extent of biological diversity. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 345(1311):13–20.

May, R. M. (2011). Why should we be concerned about loss of biodiversity. *Comptes Rendus Biologies*, In Press, Corrected Proof.

Mayer, P. (2006). Biodiversity - the appreciation of different thought styles and values helps to clarify the term. *Restoration Ecology*, 14(1):105–111.

McAllister, J. (2003). Effective complexity as a measure of information content. *Philosophy of Science*, 70(2):302–307.

McGill, B. J. (2003). Does mother nature really prefer rare species or are log-left-skewed sads a sampling artefact? *Ecology Letters*, 6(8):766–773.

McGill, B. J. (2011). Linking biodiversity patterns by autocorrelated random sampling. *American Journal of Botany*, 98(3):481–502.

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., Hulbert, A. H., Magurran, A. E., Marquet, P. A., Maurer, B. A., Ostling, A., Soykan, C. U., Ugland, K. I., and White, E. P. (2007). Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, 10:995–1015.

McGraw, J. B. (2001). Evidence for decline in stature of american ginseng plants from herbarium specimens. *Biological Conservation*, 98(1):25–32.

McKee, J. K., Sciulli, P. W., Fooce, C. D., and Waite, T. A. (2004). Forecasting global biodiversity threats associated with human population growth. *Biological Conservation*, 115(1):161–164.

Meinard, Y. and Grill, P. (2011). The economic valuation of biodiversity as an abstract good. *Ecological Economics*, In Press, Corrected Proof.

Mellin, C., Delean, S., Caley, J., Edgar, G., Meekan, M., Pitcher, R., Przeslawski, R., Williams, A., and Bradshaw, C. (2011). Effectiveness of Biological Surrogates for Predicting Patterns of Marine Biodiversity: A Global Meta-Analysis. *PLOS ONE*, 6(6).

Mendes, R. S., Evangelista, L. R., Thomaz, S. M., Agostinho, A. A., and Gomes, L. C. (2008). A unified index to measure ecological diversity and species rarity. *Ecography*, 31(4):450–456.

Mérigot, B., Bertrand, J., Mazouni, N., C., M., Durbec, J.-P., and Gaertner, J. (2007). A multi-component analysis of species diversity of groundfish assemblages on the continental shelf of the Gulf of Lions (north-western Mediterranean Sea). *Estuarine, Coastal and Shelf Science*, 73:123–136.

Millenium Ecosystem Assessment (2005). Ecosystem and Human Well-Being. Biodiverisity Synthesis. Technical report, World Resources Institute, Washington, DC.

Mittelbach, G. G., Steiner, C. F., Scheiner, S. M., Gross, K. L., Reynolds, H. L., Waide, R. B., Willig, M. R., Dodson, S. I., and Gough, L. (2001). What is the observed relationship between species richness and productivity? *Ecology*, 82(9):2381–2396.

Mooers, A. (2007). Conservation biology - the diversity of biodiversity. *Nature*, 445(7129):717–718.

Moreno, C. E., Guevara, R., Sánchez-Rojas, G., Téllez, D., and Verdú, J. (2008). Community level patterns in diverse systems: A case study of litter fauna in a mexican pine-oak forest using higher taxa surrogates and re-sampling methods. *Acta Oecologica*, 33(1):73–84.

Morlon, H., White, E. P., Etienne, R. S., Green, J. L., Ostling, A., Alonso, D., Enquist, B. J., He, F. L., Hurlbert, A., Magurran, A. E., Maurer, B. A., McGill, B. J., Olff, H., Storch, D., and Zillio, T. (2009). Taking species abundance distributions beyond individuals. *Ecology Letters*, 12(6):488–501.

Naeem, S. and Wright, J. P. (2003). Disentangling biodiversity effects on ecosystem functioning: deriving solutions to a seemingly insurmountable problem. *Ecology Letters*, 6(6):567–579.

Nehring, K. and Puppe, C. (2002). A theory of diversity. *Econometrica*, 70(3):1155–1198.

Nei, M. and Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, 76(10):5269–5273.

Nekola, J. and White, P. (1999). The distance decay of similarity in biogeography and ecology. *Journal of Biogeography*, 26(4):867–878.

Neuman, Y. (2008). *Reviving the Living: Meaning Making in Living Systems*, volume 6 of *Studies in Multidisciplinarity*. Elsevier, Amsterdam.

Neumann, R. P., Kitchin, R., and Thrift, N. (2009). Biodiversity. In *International Encyclopedia of Human Geography*, pages 308–313. Elsevier, Oxford.

Nijkamp, P., Vindigni, G., and Nunes, P. A. L. D. (2008). Economic valuation of biodiversity: A comparative study. *Ecological Economics*, 67(2):217–231.

Nipperess, D. A., Faith, D. P., and Barton, K. (2010). Resemblance in phylogenetic diversity among ecological assemblages. *Journal of Vegetation Science*, 21(5).

Norton, B. (1994). On what we should save: The role of culture in determining conservation targets. In Forey, P., Humphries, C., and Vane-Wright, R., editors, *Systematics and Conservation Evaluation*, pages 23–40. Clarendon Press, Oxford University Press for the Systematics Association.

Noss, R. F. (1990). Indicators for monitoring biodiversity - a hierarchical approach. *Conservation Biology*, 4(4):355–364.

Nunes, P. and van den Bergh, J. C. (2001). Economic valuation of biodiversity: sense or nonsense? *Ecological Economics*, 39(2):203–222.

O'Gorman, E. J. and Emmerson, M. C. (2009). Perturbations to trophic interactions and the stability of complex food webs. *Proceedings of the National Academy of Sciences of the United States of America*, 106(32):13393–13398.

O'Gorman, E. J., Yearsley, J. M., Crowe, T. P., Emmerson, M. C., Jacob, U., and Petchey, O. L. (2011). Loss of functionally unique species may gradually undermine ecosystems. *Proceedings of the Royal Society B: Biological Sciences*, 278(1713):1886–1893.

Ohsawa, M. (2004). Species richness of cerambycidae in larch plantations and natural broad-leaved forests of the central mountainous region of japan. *Forest Ecology and Management*, 189(1-3):375–385.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., and Wagner, H. (2010). *vegan: Community Ecology Package*. R package version 1.17-2.

Ordonez, C. and García-García, J. (2008). Referential integrity quality metrics. *Decision Support Systems*, 44(2):495–508.

Osenberg, C. W., Sarnelle, O., Cooper, S. D., and Holt, R. D. (1999). Resolving ecological questions through meta-analysis: goals, metrics, and models. *Ecology*, 80(4):1105–1117.

Oxbrough, A. G., Gittings, T., O'Halloran, J., Giller, P. S., and Kelly, T. C. (2007). Biodiversity of the ground-dwelling spider fauna of afforestation habitats. *Agriculture, Ecosystems & Environment*, 120(2-4):433–441.

Paillet, Y., Bergès, L., Hjältén, J., Ódor, P., Avon, C., Bernhardt-Römermann, M., Bijlsma, R.-J., De Bruyn, L., Fuhr, M., Grandin, U., Kanka, R., Lundin, L., Luque, S., Magura, T., Matesanz, S., Mészáros, I., Sebastià, M. T., Schmidt, W., Standovár, T., Tóthmérész, B., Uotila, A., Valladares, F., Vellak, K., and Virtanen, R. (2010). Biodiversity differences between managed and unmanaged forests: Meta-analysis of species richness in europe. *Conservation Biology*, 24(1):101–112.

Patil, G. P. and Taillie, C. (1982). Diversity as a concept and its measurement. *Journal of the American Statistical Association*, 77(379):548–561.

Pearce, D. (2001). Valuing biological diversity: issues and overview. In *In: OECD: Valuation of Biodiversity Benefits; Selected Studies*, pages 27–44. Paris, OECD.

Pearce, D. and Moran, D. (1995). *The economic value of biodiversity*. Earthscan Publications Ltd, London.

Peet, R. K. (1974). The measurement of species diversity. *Annual Review of Ecology and Systematics*, 5:285–307.

Perrings, C. and Pearce, D. (1994). Threshold effects and incentives for the conservation of biodiversity. *Environmental and Resource Economics*, 4(1):13–28.

Perry, N. (2010). The ecological importance of species and the noah's ark problem. *Ecological Economics*, 69(3):478–485.

Petchey, O. L. and Gaston, K. J. (2002). Functional diversity (FD), species richness and community composition. *Ecology Letters*, 5(3):402–411.

Peters, R. H. (1983). *The Ecological Implications of Body Size (Cambridge Studies in Ecology)*. Cambridge University Press.

Picton, B., Emblow, C., Morrow, C., Sides, E., Tierney, P., McGrath, D., McGeough, G., McCrea, M., Dinneen, P., Falvey, J., Dempsey, S., Dowse, J., and Costello, M. J. (1992). Marine sites, habitats and species data collected during the BioMar survey of Ireland. Digital resource.

Pielou, E. C. (1969). *An introduction to mathematical ecology*. Wiley-Interscience, New York.

Pielou, E. C. (1975). *Ecological diversity*. Wiley, New York.

Pindyck, R. S. and Rubinfeld, D. L. (1998). *Econometric models and economic forecasts*. Irwin, McGraw-Hill.

Pla, L. (2004). Bootstrap confidence intervals for the shannon biodiversity index: A simulation study. *Journal of Agricultural, Biological, and Environmental Statistics*, 9:42–56. 10.1198/1085711043136.

Platt, H. and Lamberhead, P. (1985). Neutral model analysis of patterns of marine benthic species diversity. *Marine Ecology Progress Series*, 24:75–81.

Polski, M. (2005). The institutional economics of biodiversity, biological materials, and bioprospecting. *Ecological Economics*, 53(4):543–557.

Prieto-Benitez, S. and Mendez, M. (2011). Effects of land management on the abundance and richness of spiders (Araneae): A meta-analysis. *Biological Conservation*, 144(2):683–691.

Purvis, A. and Hector, A. (2000). Getting the measure of biodiversity. *Nature*, 405(6783):212–219.

Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21(1):24–43.

Redford, K. H. and Richter, B. D. (1999). Conservation of biodiversity in a world of use. *Conservation Biology*, 13(6):1246–1256.

Reed, D. H. and Frankham, R. (2001). How closely correlated are molecular and quantitative measures of genetic variation? a meta-analysis. *Evolution*, 55(6):1095–1103.

Reed, D. H. and Frankham, R. (2003). Correlation between fitness and genetic diversity. *Conservation Biology*, 17(1):230–237.

Ressurreiçaõ, A., Gibbons, J., Dentinho, T. P., Kaiser, M., Santos, R. S., and Edwards-Jones, G. (2011). Economic valuation of species loss in the open sea. *Ecological Economics*, 70(4):729–739.

Richardson, L. and Loomis, J. (2009). The total economic value of threatened, endangered and rare species: An updated meta-analysis. *Ecological Economics*, 68(5):1535–1548.

Ricotta, C. (2005a). A note on functional diversity measures. *Basic and Applied Ecology*, 6(5):479–486.

Ricotta, C. (2005b). Through the jungle of biological diversity. *Acta Biotheoretica*, 53:29–38.

Rossberg, A. G., Matsuda, H., Amemiya, T., and Itoh, K. (2006). Food webs: Experts consuming families of experts. *Journal of Theoretical Biology*, 241(3):552–563.

Routledge, R. D. (1983). Evenness indices: Are any admissible? *Oikos*, 40(1):149–151.

Salles, J.-M. (2011). Valuing biodiversity and ecosystem services: Why put economic values on nature? *Comptes Rendus Biologies*, 334(5-6):469–482.

Sanders, H. L. (1968). Marine benthic diversity: A comparative study. *The American Naturalist*, 102(925):243.

Sarkar, S. and Margules, C. (2002). Operationalizing biodiversity for conservation planning. *Journal of Biosciences*, 27(4):299–308.

Schleuter, D., Daufresne, M., Massol, F., and Argillier, C. (2010). A user's guide to functional diversity indices. *Ecological Monographs*, 80(3):469–484.

Schrödinger, E. (1944). What is Life? The physical aspects of the living cell. http://home.att.net/ p.caimi/schrodinger.html.

Secretariat of the Convention on Biological Diversity (2010). Global biodiversity outlook 3. http://gbo3.cbd.int/.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3,4):379–423,623–656.

Shannon, C. and Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.

Sharma, R. C. and Rawat, J. S. (2009). Monitoring of aquatic macroinvertebrates as bioindicator for assessing the health of wetlands: A case study in the central himalayas, india. *Ecological Indicators*, 9(1):118–128.

Sheldon, R. W. and Parsons, T. R. (1967). A continuous size spectrum for particulate matter in the sea. *Journal of the Fisheries Research Board of Canada*, 24:909–915.

Sheldon, R. W., Prakash, A., and Sutcliffe, W. H. (1972). The size distribution of particles in the ocean. *Limnology and Oceanography*, 17:327–340.

Smith, J. M. (2000). The concept of information in biology. *Philosophy of Science*, 67(2):177–194.

Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content, and its application to analyses of the vegetation on danish commons. *Biologiske skrifter/Det Kongelige Danske Videnskabernes Selskab*, 5:1–34.

Spangenberg, J. H. and Settele, J. (2010). Precisely incorrect? Monetising the value of ecosystem services. *Ecological Complexity*, 7(3):327 – 337. Ecosystem Services - Bridging Ecology, Economy and Social Sciences.

Srivastava, D. S. and Vellend, M. (2005). Biodiversity-ecosystem function research: Is it relevant to conservation? *Annual Review of Ecology, Evolution, and Systematics*, 36(1):267–294.

Storch, D. and Sizling, A. L. (2008). The concept of taxon invariance in ecology: Do diversity patterns vary with changes in taxonomic resolution? *Folia Geobotanica*, 43(3):329–344.

Strohbach, M. W., Haase, D., and Kabisch, N. (2009). Birds and the city: Urban biodiversity, land use, and socioeconomics. *Ecology and Society*, 14(2).

Studer, R., Benjamins, V. R., and Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data and Knowledge Engineering*, 25(1-2):161 – 197.

Suneetha, M. (2010). Sustainability issues for biodiversity business. *Sustainability Science*, 5(1):79–87.

Sutton, A., Abrams, K. R., Jones, D. R., Sheldon, T. A., and Song, F. (2000). *Methods for meta-analysis in medical research*. John Wiley and Sons, ltd.

Swingland, I. R. (2007). Definition of biodiversity. In Levin, S. A., editor, *Encyclopedia of Biodiversity*, volume 1, pages 377–390. Academic Press, 2nd edition.

Szostak, J. W. (2003). Functional information: Molecular messages. *Nature*, 423(6941):689–689.

Tan, G., Gyllenhaal, C., and Soejarto, D. D. (2006). Biodiversity as a source of anticancer drugs. *Current drug targets*, 7(3):265–277.

Tanksley, S. D. and McCouch, S. R. (1997). Seed banks and molecular maps: Unlocking genetic potential from the wild. *Science*, 277(5329):1063–1066.

Taylor, E. B. and McPhail, J. D. (1999). Evolutionary history of an adaptive radiation in species pairs of threespine sticklebacks (gasterosteus): insights from mitochondrial dna. *Biological Journal of the Linnean Society*, 66(3):271–291.

TEEB (2010). *The Economics of Ecosystems and Biodiversity: Mainstreaming the Economics of Nature: A synthesis of the approach, conclusions and recommendations of TEEB*. TEEB.

Teorey, T. J., Yang, D., and Fry, J. P. (1986). A logical design methodology for relational databases using the extended entity-relationship model. *ACM Computing Surveys*, 18:197–222.

Tian, Z., Chen, W., Zhao, C., Chen, Y., and Zheng, B. (2007). Plant biodiversity and its conservation strategy in the in-undation and resettlement districts of the Yangtze Three Gorges, China. *Acta Ecologica Sinica*, 27(8):3110–3118.

Tilman, D. (2007). *Encyclopedia of Biodiversity*, volume 3, chapter Functional diversity, pages 109–120. Academic Press, 2nd edition.

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(1):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(1):23–45.

Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., and Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine*, 358(3):252–260.

UN (2011). (United Nations). Convention on Biological Diversity. http://www.cbd.int/convention/text/.

UNEP (2011). Report on the sixth meeting of the conference of the parties to the convention on biological diversity (unep/cbd/cop/6/20/part 2) strategic plan decision vi/26. http://www.cbd.int/decision/cop/?id=7200.

Valentine, J. (2003). Architectures of biological complexity. *Integrative and comparative biology*, 43(1):99–103.

Vanderwel, M. C., Malcolm, J. R., and Mills, S. C. (2007). A Meta-Analysis of Bird Responses to Uniform Partial Harvesting across North America. *Conservation Biology*, 21(5):1230–1240.

Vellend, M. (2005). Species diversity and genetic diversity: Parallel processes and correlated patterns. *The American Naturalist*, 166(2):199–215.

Verbeek, M. (2006). *A guide to modern econometrics*. John Wiley and Sons, ltd, 2nd edition.

Villar, J., Carroli, G., and Belizan, J. M. (1995). Predictive ability of meta analyses of randomized controlled trials. *Lancet*, 345(8952):772–776.

Villéger, S., Mason, N. W. H., and Mouillot, D. (2008). New multidimensional functional diversity indices for a multifaceted framework in functional ecology. *Ecology*, 89(8):2290–2301.

Virginia, R. A. and Wall, D. H. (2007). *Encyclopedia of Biodiversity*, volume 2, chapter Principles of ecosystem function, pages 345–352. Academic Press, 2nd edition.

Walpole, M., Almond, R. E. A., Besançon, C., Butchart, S. H. M., Campbell-Lendrum, D., Carr, G. M., Collen, B., Collette, L., Davidson, N. C., Dulloo, E., Fazel, A. M., Galloway, J. N., Gill, M., Goverse, T., Hockings, M., Leaman, D. J., Morgan, D. H. W., Revenga, C., Rickwood, C. J., Schutyser, F., Simons, S., Stattersfield, A. J., Tyrrell, T. D., Vié, J.-C., and Zimsky, M. (2009). Tracking progress toward the 2010 biodiversity target and beyond. *Science*, 325(5947):1503–1504.

Warwick, R. M. and Clarke, K. R. (1995). New “biodiversity” measures reveal a decrease in taxonomic distinctness with increasing stress. *Marine Ecology-Progress Series*, 129(1-3):301–305.

Watson, R. T. (2006). *Data management : databases and organizations*. J. Wiley, Hoboken, NJ.

WBD (2010). World Biodiversity Datatbase. http://nlbif.eti.uva.nl/bis/index.php.

WCED (1987). *Our Common Future (Oxford Paperback Reference)*. Oxford University Press, USA.

Weesie, P. and van Andel, J. (2003). On biodiversity and its valuation. The CDS Research Report Series., Centre of Development Studies, The University of Groningen, the Netherlands.

Weikard, H.-P. (2002). Diversity functions and the value of biodiversity. *Land Economics*, 78(1):20–27.

Weitzman, M. L. (1992). On diversity. *The Quarterly Journal of Economics*, 107(2):363–405.

Weitzman, M. L. (1993). What to Preserve? An Application of Diversity Theory to Crane Conservation. *The Quarterly Journal of Economics*, 108(1):157–183.

Weitzman, M. L. (1998). The Noah’s Ark Problem. *Econometrica*, 66(6):1279–1298.

Whittaker, R. H. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs*, 30:279–338.

Wilsey, B. J., Chalcraft, D. R., Bowles, C. M., and Willig, M. R. (2005). Relationships among indices suggest that richness is an incomplete surrogate for grassland biodiversity. *Ecology*, 86(5):1178–1184.

Wilson, E., editor (1988a). *Biodiversity*. National Academy of Sciences and the Smithsonian Institution.

Wilson, E. (1988b). Biodiversity. Number 1, chapter The current state of biological diversity, pages 3–20. National Academy Press, Washington, D.C.

Wilson, J. B. (1991). Methods for fitting dominance/diversity curves. *Journal of Vegetation Science*, 2(1):35–46.

Yang, Z. and Goldman, N. (1997). Are big trees indeed easy? *Trends in Ecology & Evolution*, 12(9):357–357.

Zamora, J., Verdù, J., and Galante, E. (2007). Species richness in mediterranean agroecosystems: Spatial and temporal analysis for biodiversity conservation. *Biological Conservation*, 134(1):113–121.