Incorporating spatial and temporal variability in analyses of the relationship between biodiversity and ecosystem functioning

Matteo Tanadini
Lincoln College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Michaelmas 2016

Abstract

Thesis submitted for the degree of Doctor of Philosophy

Incorporating spatial and temporal variability in analyses of the relationship between biodiversity and ecosystem functioning

In the last few decades, a growing literature has examined how biodiversity influences ecosystem functioning. This body of work has greatly improved our understanding of ecosystem functioning and its modulation by biodiversity. In particular, there is nowadays large consensus that biodiversity increases ecosystem productivity, and stabilises ecosystems.

Early investigations were largely theoretical or involved simple experiments run in laboratory conditions, but over time biodiversity ecosystem-functioning experiments evolved to more realistic field experiments that better represent the real conditions found in natural ecosystems. In particular, these experiments are often run on larger spatial scales and over longer time frames allowing for the effect of environmental heterogeneity and temporal fluctuations to be explored.

The designs of these experiments evolved along with the questions addressed in this field of research. However, the analytical tools used in the analyses of these experiments followed a slightly different path. In particular, most of the metrics currently used to analyse biodiversity ecosystem functioning experiments are not entirely suited to properly deal with the complexity of modern designs as they make a number of assumptions that are not met any more.

In my thesis I developed a unified framework, based on the tailored use of Linear Mixed Effects Models, to analyse biodiversity-ecosystem functioning experiments such that the new complexities of these experiments can be taken into account. This thesis aimed to bring the focus of the analysis back to the biological interpretation of the results. I successfully applied my approach to several data sets.

The framework developed here is expected to improve greatly our understanding of ecosystem functioning and how biodiversity modulates it. It also sheds new light on past research in this field. The great flexibility of the new approach makes it possible to let these experiments to evolve such that new biological questions can be addressed.

Dedicato alla mia famiglia, Tania e Bianca
siete la cosa piú bella della mia vita

Acknowledgements

During my doctoral studies I had the great chance to meet and share opinions with a large number of people who had very different backgrounds and origins. This was extremely enriching. I want to thank all of them. In particular, I wish to thank my supervisor Andy Hector, whose enthusiasm for the transdisciplinary field of ecological statistics made this thesis possible.

I am very grateful to the people that read and commented on the chapters of this thesis and their extensive appendices. Their help has been invaluable. I hope to be able to return this favour. I am also thankful to all people who work at the Department of Plant Sciences and at Lincoln College. My family also deserves special recognition for their support during these years of intense work.

This research was founded by the *Berrow Foundation*. I particularly want to thank the *Marquise de Amodio* for giving me, and other Swiss students, the opportunity to study at Oxford University.

Contents

Chapter 1

General Introduction

1.1 Thesis scientific introduction

Biodiversity is declining

Biodiversity, the variability of living organisms represented in genes, species, populations, communities and ecosystems is rapidly declining (Naeem et al., 2012; Barnosky et al., 2011). This negative trend, observed worldwide, is attributed to several factors and their interactions (Hooper et al., 2012). The most important of these factors, such as habitat loss, habitat alteration and climate change, are tightly linked with human activities (Ceballos et al., 2015; Pimm et al., 2014). These factors play a major role in determining which species are lost, and ultimately, affect the global functioning of ecosystems. In the light of these changes, it is fundamental to better understand the consequences of species loss on ecosystems (Cardinale et al., 2012; Tilman et al., 2014).

Biodiversity ecosystem-functioning experiments

Experiments on the relationship between biodiversity and ecosystem functioning, hereafter diversity-function experiments, create artificial ecosystems of differing diversity by assembling organisms of different species into an experimental unit. These experiments are often carried out with primary producers such as grasses, trees and algae (Cardinale et al., 2012; Tilman et al., 2014). The assemblages constructed by experimenters range from ecosystems composed of just one species (monocultures) through a subset of some (or all) the possible species assemblages, often up to the full species compo-

sition. In this context, species richness is considered to be a surrogate of biodiversity. However, diversity-function experiments have also manipulated other aspects of biodiversity, including numbers of functional groups (Arenas et al., 2006) and levels of genetic diversity (Hughes and Stachowicz, 2004; Prieto et al., 2015). The experimental plant communities are then left to grow over time, which usually ranges from one growing season up to several years (Hector et al., 1999; Potvin and Gotelli, 2008; Hautier et al., 2015).

Ecosystem functions (or processes) are comprised of stocks and flows of energy and matter. These include productivity, carbon storage and nitrogen fixation among others. In diversity-function studies, productivity is probably the most commonly studied ecosystem function (Cardinale et al., 2012). Net primary productivity is approximated through measurements such as grass biomass, tree diameter (and its annual increases as seen in tree rings) or algal biomass (e.g. Vojtech et al. (2008); Philipson et al. (2012); Arenas et al. (2006)). The variability in the responses of these experimental communities over time, known as "temporal stability", can also be measured (Loreau, 2010). Temporal stability is determined to a large degree by how synchronous or asynchronous is the fluctuation in individual species biomass over time. If species are perfectly synchronous then the community is a simple sum of its parts and the whole system tracks the fluctuations of the individual species. On the other hand, if species responses are asynchronous then the out of phase fluctuations buffer the whole system response (Tilman et al., 2006). Ecosystem stability can also be measured over space (i.e. spatial stability; Loreau (2010)).

In diversity-function experiments, biomass is compared across species richness levels to assess whether ecosystem productivity and stability are affected by diversity. There is growing evidence that biodiversity increases productivity and stabilises ecosystems (Cardinale et al., 2012; Tilman et al., 2014). Two not mutually exclusive mechanisms have been proposed to explain the higher productivity of richer ecosystems: "Complementarity" and "selection effects". Complementarity refers to positive species interactions or resources partitioning among species that lead to greater productivity.

The selection effect indicates dominance of species with particular traits that affect ecosystem productivity (Huston, 1997; Tilman et al., 1997; Loreau and Hector, 2001). Analogously, greater stability of more diverse ecosystems has been attributed to several factors, species asynchrony among others (Loreau and de Mazancourt, 2008).

Studying ecosystem functioning is an essential component of being able to predict how ecosystem services, "nature's societal benefits", will evolve under the current figure of biodiversity loss (Isbell et al., submitted; Cardinale et al., 2012). It is fundamental to act now as ecosystems functioning and thus the associated services are likely to be affected gradually rather than directly after species losses (Isbell et al., submitted). This is a worrying prospect in that even ecosystem functions and services that persist in our world today may already be accruing debt, of which the full consequences have yet to be seen (Wearn et al., 2012).

First generation experiments

The first generation of diversity-function studies (e.g. the Ecotron by Naeem at Imperial College (1994)) were conducted in laboratories and addressed the general question of how species diversity influences ecosystem functioning (Tilman et al., 2014). These studies were complemented with intensive theoretical work often based on simulations (Tilman et al., 1997; Loreau, 2010).

These studies greatly improved our understanding of ecosystem functioning and how this is modulated by biodiversity (Tilman et al., 2014; McCann, 2000; Loreau et al., 2002). However, these experiments were criticised in several ways. In particular, they used random assembly of species mixtures (implying random loss of species) and were run in laboratory or greenhouse conditions that were not considered realistic (Huston, 1997; Loreau et al., 2002; Srivastava and Vellend, 2005; Duffy, 2009; Wardle and Jonsson, 2010).

Second generation experiments

Later generations of experiments aimed to extend this work into more natural settings and realistic situations (e.g. the Jena Experiment in Germany; Weigelt et al. (2010)). Larger and more complex experiments were designed (e.g. the Biodepth Experiment ran on eight European sites; Hector et al. (1999)). This new cohort of experiments were also ran for longer time frames (e.g. the Cedar Creek Experiment by Tilman in the USA; (1996)). Moving outside the laboratory also enabled researchers to investigate other types of ecosystems such as forests (e.g. the Sabah Biodiversity Experiment in Borneo; Hector et al. (2011), the Satakunta Forest Experiment in Finland by Koricheva; (2002), or the collaborative project BEF-China; `www.bef-china.de`).

The second generation of diversity-function experiments enabled researchers to conclusively resolve earlier debates, and consequently to address new and more specific questions. For example, Tilman and colleagues at Cedar Creek, demonstrated with compelling evidence that biodiversity has a stabilising effect on ecosystems (i.e. richer ecosystems are more stable; Tilman (1996); Tilman et al. (2006)). This provided the basis for other researchers to begin to better understand how this process works. Consequently, Loreau and de Mazancourt (2013) proposed three, not mutually exclusive, mechanisms that could explain the observed pattern (i.e. species asynchrony, differential speed of reaction to disturbances, and lower competition). This example shows the interplay between experimental and theoretical understanding, and how the increased understanding of these important ecological processes gained through experiments inevitably leads to more specific and harder to test hypotheses.

Tools used to analyse diversity-function experiments

The analysis of diversity-function experiments followed two different paths. On one hand, the analysis of the effects of biodiversity on ecosystem functioning (e.g. productivity) is carried out with a modern mainstream statistical technique, Linear Mixed

Effects Models[1] (Schmid et al., 2002). This approach makes it possible to take the complexity of the newer designs into account and provides a solid framework for formal testing of biological hypotheses (Schmid et al., 2002). Hereafter, I refer to this part of the analysis of diversity-function experiments as the "primary analysis".

On the other hand, other aspects of diversity-function studies are carried out with a wide variety of field-specific metrics. For example, various metrics have been created to compare the performance of mixtures to monocultures to test for overyielding[2] (Hector, 1998; Loreau, 1998). Other metrics have been devised to measure species asynchrony (Loreau and de Mazancourt, 2008; Gross et al., 2014) or ecosystem stability (Tilman et al., 1997; Carnus et al., 2015). These metrics were developed independently and do not have a shared framework of use and interpretation. These analyses, referred hereafter as "secondary analyses", are carried out independently with no formal connection among them nor with the primary analysis.

Knowledge gap

To reiterate, the analysis of diversity-function studies is composed by two disconnected parts, a primary analysis and the secondary analyses, that use two fundamentally different approaches. Furthermore, secondary analyses are carried out with a piecemeal approach, that uses a vast range of metrics developed independently. Strikingly, even closely related ecological concepts such as species asynchrony and ecosystem stability are analysed with metrics that can greatly differ in use and interpretation.

This situation is inconvenient as it makes the interpretation and the reporting of the results of diversity-function studies difficult. It is also not always straightforward to understand the practical and biological implications of these experiments (e.g. two ecosystems may have different stability values, however, in practice this difference is irrelevant). The difficulty in interpreting the results of diversity-function studies can

[1] This method enables the user to include both random and fixed effects into a single model. Fixed effects are quantities of direct interest (e.g. the effect of biodiversity on biomass), while random effects are used to take the grouping structure of the data into account (e.g. to account for the fact that trees are grouped in plots).

[2] Overyielding refers to the increase in yields observed in mixtures as compared to the weighted average of monocultures. Weights are based on the species relative abundances in the mixture.

have important real-world consequences as the ultimate scope of diversity-function studies is to provide policy-makers involved in ecosystem management with sound, understandable relevant information (Isbell et al., submitted). The difficulties in interpreting the results of diversity function studies are further discussed in Chapters 2 and 3.

Aim of the thesis

This thesis aims to bring together all the analyses of a diversity-function experiment into a shared framework with a strong focus on the biological interpretation of the results and ease in their communication. To achieve that, the advantages of the model-based framework used in the primary analysis are combined with the specificity of the metrics used in the secondary analysis.

1.2 Thesis structure

The structure of this thesis fulfils the requirements for a DPhil submitted by papers in the Department of Plant Sciences, as recognised by the University of Oxford. This thesis is composed of seven chapters: A general introduction, five "papers" (hereafter "main chapters") and a general discussion. The five main chapters are conceptually divided into three groups. The first two main chapters (2 and 3) present my new analytical approach to analyse diversity-functioning experiments. The fourth main chapter presents a case-study in a closely related field of research. Finally, in the main chapters 5 and 6 two diversity-function studies are analysed with my approach.

Chapter 2

This first methodological chapter presents how to extend the primary analysis to quantify and compare, in a formal way, the functioning of different ecosystems. For example, I show how to compute overyielding or the effect of legumes presence on ecosystem productivity.

Chapter3

This chapter, illustrates how to use the analytical framework of the primary analysis potentially to answer all questions addressed in secondary analyses. In particular, this chapter shows how to quantify temporal and spatial stability, and species asynchrony.

Chapter 4

This chapter discusses the results of a previously published study that addressed a specific form of ecosystem functioning: Agricultural productivity. In particular, the latter study inspected the effects of no tilling practices on crop yields. This chapter, although not formally a diversity-function study, highlights important practical aspects discussed in chapters 2 and 3. In particular, it highlights: i) The fundamental importance of an appropriate graphical visualisation (of the raw data and of the results); ii) the correct interpretation and reporting of ecological studies; and iii) the distinction between statistical and biological significance.

Chapter 5

This chapter shows the analysis of a diversity-function study run in an experimental forest in the Czech Republic. In this study the approach presented in the two first chapters is used. This chapter has been published in the *Journal of Applied Ecology* (Chamagne et al., 2016).

Chapter 6

This chapter shows the analysis of a large-scale diversity-function experiment: The Sabah Biodiversity Experiment (Hector et al., 2011). In particular, I applied the approach presented in chapters 2 and 3 to analyse the data collected during the first decade of this experiment. This chapter has been published in the *Proceedings of the Royal Society B* (Tuck et al., 2016).

General Discussion (Chapter 7)

Finally, I summarise the main results of this thesis and discuss them in a broader context. In particular, I discuss the improvements introduced in terms of biological understanding of ecosystem functioning and how past research may be re-evaluated. I also explain that this doctoral thesis represents the basis for a medium-term project aimed to further improve the analysis of diversity-function experiments.

Because of the *paper structure* of this thesis, some concepts are introduced several times in different chapters. In addition, in each main chapter we refer to other main chapters as "publication in preparation" (e.g. Tanadini and Hector, in prep. a).

Appendices

The five main chapters are associated with extensive appendices. Those associated with chapters 2 and 3 are fundamental to understand how the proposed approach can be put into practice. They show several worked-examples and explain the details of my method. In addition, they discuss and highlight shortcoming of the metrics currently-used in secondary analyses (i.e. secondary metrics). The appendices of Chapter 4 illustrate and support the claims made in the main text of the chapter. Finally, for the two last chapters (5 and 6), the appendices aim to make the analysis transparent and reproducible. For this reason, the appendices of chapters 2, 3 and 4 are presented in the body of this thesis, while those associated with chapters 5 and 6 are to be found in the electronic supporting material (Compact Disc).

Chapters' preface

Each chapter has a preface section that specifies further details (e.g. authors, context and target audience).

Chapter 2

A modern model-based framework to compare ecosystem functioning

2.1 Preface

Authors: Matteo Tanadini[1] & Andrew Hector[1]
1. Department of Plant Sciences, University of Oxford, South Parks Road, OX1 3RB.

Target readership: Researchers working in the filed of biodiversity ecosystem functioning experiments.

Context: This chapter illustrates my approach to measure overyielding and related hypotheses. This chapter represents, along with chapter 3, the core of the thesis.

Article status: This article is in preparation to submit.

2.2 Main text

Introduction

Biodiversity ecosystem-functioning studies are modern experiments that aim to understand the effects of diversity on ecosystem functions (Cardinale et al., 2012). In these experiments several features of interest are analysed. A typical analysis quantifies the effects of biodiversity on productivity and ecosystem stability (Cardinale et al., 2012). Subsequently, other aspects of ecosystem functioning are quantified, generally in a disconnected manner. For example, temporal and spatial asynchrony are often measured to understand to what extent species respond differently to heterogeneous environments and climate fluctuations (Loreau, 2010). Quantifying asynchrony is fundamental to understand how ecosystems can be stabilised by species richness.

An important feature that is virtually always analysed in diversity-function experiments is overyielding, the surplus production of mixtures over the corresponding monocultures (Cardinale et al., 2012). This involves comparing the productivity of the full-species mixture with the weighted average of all monocultures (weights are based on species abundances in the full-species mixture). There are two, not mutually exclusive, hypotheses that try to explain the higher production of more diverse ecosystems, "selection effect" (simple dominance by individual species with particular traits that affect productivity) and "complementarity" (species positive interactions or niche partitioning; Tilman (1999)). Other questions regarding the comparison of ecosystem productivity often involve comparing different functional groups (e.g. effect of legumes or C4 plants). Hereafter we refer to these questions as "overyielding and related questions".

To quantify all these ecosystem properties and test biological hypotheses a very large number of metrics have been proposed over the last few decades. To take an example, to measure ecosystem stability we may use the coefficient of variation (CV), a simple measure that quantifies stability by looking at the variability of production while taking into account the magnitude of the production itself ($CV = sd/mean * 100$). However, other metrics have been proposed, CV2 is a slight modification of

CV ($CV2 = var/mean * 100$; Tilman et al. (1998)). There are also other metrics that differ fundamentally from CV (e.g. Loreau and de Mazancourt (2008)). A similar situation is encountered for the other metrics used to quantify spatial and temporal asynchrony or overyielding. The shared property of all the metrics used to quantify ecosystem properties is that they condense all the biological information of interest into one single value. In addition, all these metrics are specific and test a single well-defined ecological hypothesis.

These metrics are extremely useful and have improved our understanding of ecosystem functioning and its modulation by biodiversity. After decades of research, there is now clear evidence that biodiversity increases productivity and stabilises ecosystems (Tilman and Downing, 1996; Tilman et al., 2006). In particular, overyielding metrics, that are the focus of this publication, were essential to demonstrate that richer ecosystems are indeed more productive than monocultures. A companion publication will focus on the metrics used to measure two other important aspects of the analysis of diversity-function studies: Ecosystem stability and asynchrony (Tanadini and Hector, in prep.b).

Disadvantages of currently-used overyielding metrics

Although very useful, all the metrics used to measure overyielding and related questions looking at ecosystem productivity can have very different ranges, relative scales, and meanings. This is due to the fact that they were developed independently and do not use a shared framework. Furthermore, each question addressed requires its own separate analysis in such a way that information is not shared among them.

For example, we can look at two of the most widely used metrics of overyielding, the Relative Yield Total (RYT; see Hector (1998)) and the weighted average proportional deviation $\overline{D}$ proposed by Loreau (1998). The range of possible values obtained for RYT goes from 0 to infinity, where the value 1 indicates equal production in mix yield (the additive sum of yields in species monocultures weighted by their relative abundance in the mixture) and in the full-species mixture. On the other hand, the range of the metric proposed by Loreau goes from -1 to infinity where 0 implies equal production.

So the range of similar metrics that aim to measure the same biological process can be different. In addition, the meaning of the obtained values is different across metrics.

To analyse fully a data set several metrics may be required (Loreau, 1998). Therefore, due to the different ranges, meanings and scales of the obtained values, it can be difficult to compare results within and among studies.

For example, can a RYT value of 0.94 be considered different from the one obtained in another study that is 1.07 (i.e. slight overyielding)? Is this difference statistically significant? More importantly, is this difference biologically relevant? Currently it is not straightforward how this kind of question can be answered as for the vast majority of overyielding metrics there is no clear path to make statistical inference with the obtained values. Furthermore, it is not entirely clear how the biological relevance can be judged.

Some metrics display asymmetric ranges that makes the comparison of results among studies difficult. Indeed, is it not clear whether a deviation from the value one for RYT has the same meaning in both directions. Put simply, is a RYT value of 0.9 (which implies lower production in the the mixture) the opposite equivalent of 1.1[1]?

More practically, all these values are in a scale that is not straightforward to interpret. They are to be considered proportional relative deviations (Loreau, 1998), however, as we discussed above, they do not range from 0% to 100% as we would expect for a proportion. Consequently, their actual value is not easy to understand biologically.

The above example indicates that it may be difficult even for a field specialist to know all overyielding metrics available and to know exactly how to interpret the obtained values. It is also clear that comparing results within and among studies can be a challenging task. This implies that, for non-field specialists, it can be very hard to grasp fully the results of an analysis that uses these metrics. Finally, it is also worth mentioning that the final recipients of the results obtained in diversity-function

[1]It can be argued that mathematically we can obtain the opposite equivalent by dividing 1 by the obtained RYT value.

studies are policy-makers that may not have a scientific background and may struggle to understand the message conveyed from this scientific field.

To summarise, overyielding is currently analysed using several metrics that are different in terms of scales, ranges and interpretation. This is due to the fact that these metrics were developed independently and use different procedures to be computed. The practical consequences of this heterogeneity is that results can be hard to interpret and to compare within and among studies. This hinders the communication and spread of valuable scientific information among research fields and between science and society.

The interpretation issues are worsened because overyielding is just one of the many questions addressed when comparing ecosystem functioning. Other biological questions are often of interest (e.g. the effect of specific functional groups on productivity). For all the other questions that involve comparing the productivity of different ecosystems other methods are used. This implies that comparing results across biological questions is even harder. For example, we may want to compare whether the effect of overyielding is of the same order of magnitude as the effect of adding legumes to an assemblage. Combining information gleaned from different, but closely related analyses, is fundamental to better understand ecosystem functioning.

Post-hoc contrasts with modern Linear Mixed Effects Models

Ideally, to improve the above mentioned inconveniences, we need a single unifying framework where all questions regarding ecosystem productivity can be addressed. This implies that the same procedure is used in all cases, and thus, that the results are easy to compare within and among studies. A very convenient feature of this framework would be an easy to interpret scale, where the obtained values are straightforward to understand biologically.

We identified the use of modern Linear Mixed Effects Models, by which we mean those models where a variance parameter is explicitly estimated for each random effect, combined with post-hoc contrasts as a potential solution to solve the currently encountered issues. Contrasts have long been used to test specific hypotheses in diversity-function studies (Schmid et al., 2002). The workflow presented here is based on this

latter approach and further develops it to potentially include all questions regarding the comparison of ecosystem productivity, most importantly overyielding.

Post-hoc contrasts are a modern mainstream statistical technique used to test hypotheses that involve comparing groups of observations (Everitt and Hothorn, 2011). In other words, they are the usual path used to compare the mean value of levels of a categorical variable. These hypotheses can involve simply comparing the mean value of two groups (e.g. productivity of monoculture A and B), or can be complex and test very specific hypotheses (e.g. does the addition of legumes to two-species mixtures increase productivity more than the addition of other functional groups; Schmid et al. (2002)). Put simply, this technique allows us to compute the difference in means between groups of levels of a categorical variable and tells us whether this difference is significant.

This modern technique, which is mature and implemented in any statistical software suited for scientific publication, can be used in a wide variety of situations. Its use is possible even with complex modern models such as Linear Mixed Effect Models. Many diversity-function studies are best analysed within this framework as they have a repeated measure structure, with plots or trees measured repeatedly over time and have important design features such as blocks or sites.

A great advantage of this method is that we can test a wide variety of biological hypotheses and the obtained results are in the same scale as the response variable. In other words, if we were to quantify overyielding with this technique, then the obtained value would be the difference in productivity between all monocultures and the full-species mixture (e.g. $+10\ g/m^2$). This is very easy to interpret biologically, and we can also judge the practical relevance of that difference (a statistically significant difference of $10\ g/m^2$ may be irrelevant in practice). On top of that, we can easily obtain a p-value for any tested hypothesis and the estimated difference comes with a confidence interval. This latter point is fundamental to be able to compare results within and across studies.

Example 1

To briefly illustrate how this method works, we introduce here a real-data example. The data used comes from a diversity-function study using five perennial grass species found in European fertile meadows (Vojtech et al., 2008). For the sake of brevity the details of the experimental design are described in Appendix A and Vojtech et al. (2008). The response variable analysed in this study is the dry biomass harvested. All five monocultures were used as well as all ten two-species mixtures and the full-species mixture (i.e. the assemblage of the five species). There are thus 16 different species compositions (three- and four-species compositions were not created). All assemblages, were replicated five times (80 plots in total) and each of them was harvested eight times.

Appendix A shows an in-depth analysis of this data set. For the sake of brevity, we focus here on one single hypothesis. In particular, we are interested in testing whether putting two species together produces larger yields than the two corresponding monocultures. We want to test this hypothesis for each two-species mixture, as well as in general.

The procedure to test this hypothesis is straightforward and very simple to implement in any modern statistical software (Appendix A shows every step of the testing procedure for this hypothesis). Essentially, we just need a Linear Model, or any extension to it, that contains a categorical predictor coding for species composition. In this case, we fitted a Linear Mixed Effect Model where species composition was included as a fixed effect.

By using post-hoc contrasts, we can obtain a p-value for each of the ten two-species mixtures, testing whether there is a difference in yields between the mean of the two monocultures and the corresponding two-species mixture. In addition, we also get a p-value that tests the hypothesis that, in general, there is a difference between yields of the two monocultures and the corresponding two-species mixture.

The approach presented here strongly relies on the graphical representation of the results (Gelman, 2005; Bates, 2010). Therefore, instead of reporting the 11 estimated differences and the associated p-values, we reported the estimated contrasts along with their 95% confidence interval on a graph (here produced with the add-on **R** package

multcomp; R Core Team (2015); Hothorn et al. (2016)).

Each row on Figure 2.1 below represents the test where the average productivity of the two monocultures is compared to the corresponding two-species mixtures. For example, the first row tests whether the monocultures *Alopecurus pratensis* and *Anthoxanthum odoratum* (abbreviated to "Al" and "An") are more productive than the corresponding two-species mixture. The last row tests the general hypothesis that two-species mixtures are more productive than the corresponding monocultures.

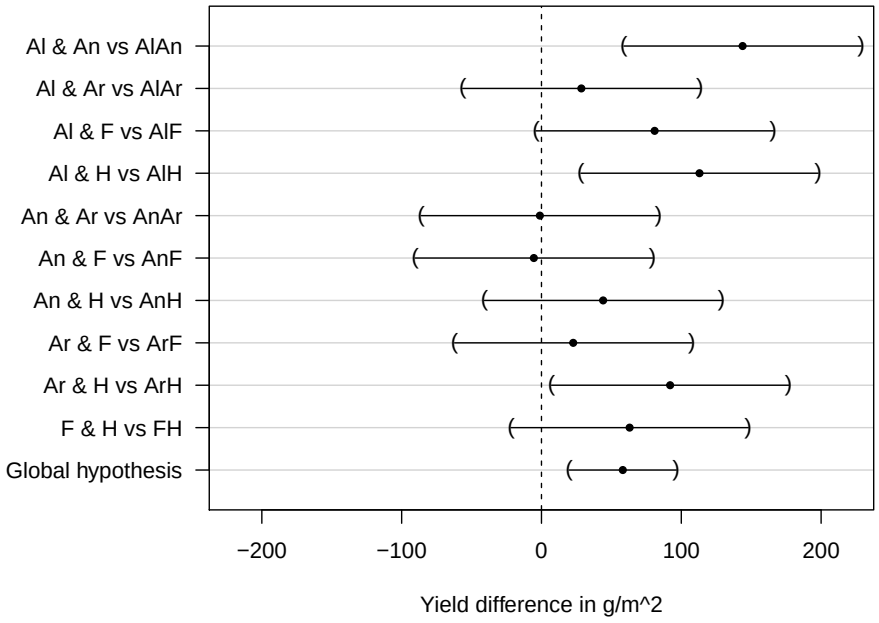


Figure 2.1: Fitted contrasts for all pairs of monocultures against the corresponding two-species mixtures (acronyms are defined in Vojtech et al. (2008)). Estimates are shown along with confidence intervals.

Figure 2.1 clearly indicates that most two-species mixtures yield more than the corresponding average of the monocultures. Indeed, the vast majority of the estimates lie on the right-hand side of this graph. As the confidence intervals indicate, only a few of these tests are statistically significant at the 5% percent level. The global hypothesis, illustrated in the last row, indicates that on average two-species mixtures yield about 60 g/m^2 more than the corresponding monocultures.

In this simple introductory example, for the sake of simplicity, we assumed that all species have the same abundance (50:50) in the two-species mixture. With the ap-

proach presented here, this assumption can easily be relaxed, so that the real observed proportions are used (see Appendix A, Section 3).

Interpretation of the estimated values

Figure 2.1 highlights some advantages of the approach presented here over currently-used metrics in terms of interpretation. To start with, the value obtained is indeed in the same scale as the response variable. For example, the first row (or contrast) which tests whether the two-species mixture `AlAn` is more productive than the average of the two corresponding monocultures, is quantified as an overproduction of 144 g/m^2. This is indeed the difference in measured yields between the average of the two monocultures and the corresponding two-species mixture. This result is very easy to interpret biologically and also to communicate. Note that if the response variable was transformed prior to fitting the Linear Mixed Effects Model (e.g. log-transformation), we could easily back-transform the obtained estimates. If we were to prefer to report our results in percentages, this would be straightforward to compute (i.e. 112%).

In addition, the range of the measure proposed here is symmetric. A value of +10 or -10 are the exact equivalent in the opposite directions (one indicates an higher productivity the other indicates a lower productivity). A value of zero intuitively indicates equal productivity, and the boundaries of this measure are again symmetric (i.e. potentially minus infinity and infinity).

The figure above highlights two other important aspects of this approach. First, most currently-used metrics condense a complex biological process into a single value. In this case, the classical approach would correspond to the last row of Figure 2.1, a single value that quantifies whether putting two species together yields more than the average of the corresponding monocultures. In reality, this effect may be different depending on what kind of species are put together. It may be that some functional groups tend to increase the biomass of the mixtures. Alternatively, we may have mixtures where productivity is larger than in the corresponding monocultures while others where it is lower. In this case, we would observe, on average, no effect. It is thus important to be able to break down hypotheses so that the whole picture is clearer. In

this example, we clearly see that mixtures seem to benefit in almost all situations, with no clear counterexample. Nevertheless, there is some non-negligible variation in the positive effects obtained by growing species together and we may want to dig deeper to better understand it.

The approach presented here is well-suited to work in conjunction with graphs. A well-chosen graph conveys information much more efficiently than text or tables (Sarkar, 2008). The table corresponding to Figure 2.1 would have 11 rows and at least four columns (i.e. test name, estimated effect, lower limit, upper limit). Such a table is very likely less efficient at conveying information.

Interestingly, the estimated gains in productivity (i.e. the estimated values of the contrasts) are reported with confidence intervals. This is very appealing as it agrees with modern trends in statistics (Wasserstein and Lazar, 2016; Gelman, 2005; Pinheiro and Bates, 2006). Indeed, this gives us more information, as the biological relevance of an effect can easily be judged (Cox, 1982). For example, it may be that the overproduction of the mixture `AlAn`, although statistically significant (see confidence intervals), is practically or biologically irrelevant. Viceversa, it may be that the overproduction of `AlF`, which is not statistically significant at the 5% level, is biologically relevant. In addition, confidence intervals are also fundamental to be able to compare results within and across studies. This is of vital importance for meta-analyses and other studies that try linking the results of different experiments.

Flexibility of the proposed approach

As it is shown in Appendix A, the approach presented here allows to test all currently addressed questions into this single unifying framework. To cite a few examples, we tested for overyielding, transgressive overyielding[2], and we tested whether a specific functional group contributes more to mixtures than other groups.

This method is extremely flexible as it allows to test many hypotheses. This makes it possible to test for study-specific questions while using the same framework. For example, we may test whether the crops previously used on the test fields affect the

[2]This term indicates that the full-species mixture outperforms the best performing monoculture.

results or we may want to compare sites of a multi-sites study.

Undoubtedly, the greatest advantage of this flexibility is that we can test newly-emerged questions without the need for a specific measure to be published (which can take considerable long time). Given a well-defined biological question, there is only one way to test it using post-hoc contrasts. Indeed, in contrast to current practices, this method has no redundancy. For example, to quantify overyielding we fit a contrast where the average productivity of all monocultures, weighted for their abundance in the mixture, is subtracted to the full-species composition productivity[3]. Because of the uniqueness of testing a hypothesis any new biological hypothesis can be quickly formalised into a test.

Some biological questions addressed in these studies hypothesise that one group of assemblages is more productive than another group. So we are interested not simply in testing for a difference, but for superiority. This very easy to implement with one-sided tests, which are part of any post-hoc package.

Some metrics that compare ecosystem productivity are based on sorted data (i.e. in each composition the response variable is measured at population level) (Hector et al., 2011; Vojtech et al., 2008). Our approach is naturally suited to test hypotheses that use sorted data (see Appendix A).

An advantage over current metrics is that we better deal with the designs of current diversity-function experiments. For example, most modern studies have a time component (i.e. the variable of interest is repeatedly measured over time). Therefore, we may want to measure overyielding at each time point to see how it changes. Or we may want to measure whether growth differs in different (functional) groups. This is very simple to implement with our approach, to give an example in Appendix A we tested whether growth in species compositions containing a particular species is faster than in all the other species compositions.

Put simply, this approach enables the user to test, in a unique and non-redundant way, a large pool of questions. This property is fundamental to be able to deal with

[3]Note that it does not matter which term is subtracted (here the weighted average of the monocultures), results are identical up to their sign.

currently addressed questions, study-specific questions and future questions in a single framework.

Advantages of the model-based framework

One of the greatest advantages of this approach is the fact that we are working on a model-based framework. This implies that our analysis can efficiently take into account the effect of any other relevant factor that may confound the obtained results. Indeed, post-hoc contrasts can be applied to any categorical variable of a Linear Model.

Many modern diversity-function experiments are run in realistic conditions outside the laboratory (i.e. environmental heterogeneity and temporal fluctuations are present). This implies that the experimental units (e.g. plots) may slightly differ from each other in terms of fertility or other factors. Some plots may be placed in less sunny regions of the experimental field (shadow is known to influence productivity; Fridley (2003)). Spatial heterogeneity may also be present in the form of soil fertility, soil moisture or pathogen load (Fridley, 2002). All these uncontrolled variables may influence yields and thus add noise to the effects of interest (i.e. the effect of species richness on productivity).

Fortunately, with the approach presented here we can include these confounding variables into our model, and thus, control for their effects. This approach is referred as "conditional", as we can condition (i.e. control) for the effects of the confounding variables. For example, when measuring overyielding we can remove the confounding effect of soil fertility. Currently-used metrics cannot take into account any other predictor and are thus said to use a "marginal" approach.

A model-based framework allows us to include design variables such as block or site in the modelling phase. There may indeed be some variability brought in by these design elements that we wish to control for. These types of predictors are usually included as random effects as we are usually not interested in their actual estimates but rather in their variability.

Another design element that can be included as a random effect, if required, is the experimental unit, in other words, the plot (Schmid et al., 2002). Indeed, most

diversity-function experiments measure repeatedly each plot, this yields a repeated-measures design type. Obviously, all measurements of one particular plot cannot be considered independent from each other. By including this element as a random effect in our model, we account for the grouped structure of the data, and we also control for the variability of this element (Bates, 2010). Therefore, all estimates obtained for other parameters of the model (e.g. species composition effects) are more precise.

As mentioned above, there is often a time component in these studies. Because most of them are run in realistic conditions, there may be important differences among time points (i.e. harvests, e.g. Vojtech et al. (2008)). Again, we may want to model and control for the variability generated by climatic fluctuations (this can be done with either a random or a fixed effect).

The advantages of using a model-based framework over the use of marginal metrics to compare the productivity of ecosystems are manifold. These do not limit themselves to the fact that we can account for the effect of external factors, and thus better estimates of the quantities of interest. Indeed, this approach enables us also to account for other relevant information such as sampling effort. To give a real example, we can look at the data set analysed in Appendix A. In this study, due to logistic constrains, different areas were harvested at different time points. Not surprisingly, smaller-area samples turned out to be less precise measurements of ecosystem productivity. We can simply include this important information in any parametric model by providing weights (i.e. weighted regression; see Appendix A section 3). The model that included weights was able to show that transgressive overyielding is present. This is due to increased power of the model based-approach.

Fitting a comprehensive model means that we can quantitatively compare the magnitude of the different effects. For example, the model fitted to this data presented in Appendix A shows that the differences in yields across years are fairly variable and are of the same magnitude as the observed biodiversity effects. This information is fundamental to discuss the practical relevance of the estimated effects. Other studies have shown that the effects of diversity on productivity are weaker than other real-world ef-

fects such as soil fertility (Fridley, 2002, 2003). Alternatively, we can also compare the variability of random effects (as quantified in the modelling phase) and overyielding (see Appendix A). Indeed, they lie on the same scale. In Appendix A we showed that the variability among plots is negligible compared to the effect of overyielding, however the variability within plots over time (i.e. residual variance) is not. This type of practical interpretation of the results is very valuable to make evidence-based recommendations to policy makers.

Another advantage of the model-based framework is that we do not need to restrict ourselves to the analysis of continuous traits such as biomass or size. Up to this point, we assumed that the ecosystem function of interest is productivity in grassland studies. However, there are many other ecosystem functions that biologists may want to quantify. We may want to understand how "invasible" an ecosystem is (as quantified in terms of non-target species found in each plot) and whether biodiversity can influence that. In that case, two possible response variables to be measured are i) the number of invading plants/trees or ii) the number of invading species. To model that data, we may want to use a Generalised Linear Model with Poisson family (as they are count data; Faraway (2005)).

In the last few decades, several forest experiments manipulating biodiversity have been set up (Hector et al., 2011; Bruelheide et al., 2014; Potvin and Gotelli, 2008). In these experiments trees are monitored for growth, measured as diameter increase, but also for survival. In that particular case, the response variable is binary (i.e. alive/dead) and could be modelled using a Generalised Linear Model with binomial family. Trees survival may be an interesting ecosystem function to measure and understand. For example, we may want to understand whether more biodiverse plots show lower mortalities. This may be economically and practically important as replanting schemes can have very low survival rates (Hector et al., 2011).

In the past, other model-based approaches that can include external factors have been proposed to compare the productivity of ecosystems (Bell et al., 2009; Connolly et al., 2013). Unfortunately, neither of them can directly be used to estimate overyield-

ing or transgressive overyielding, which are possibly the most common question regarding ecosystem productivity. Nevertheless, these two approaches represent the building blocks on which our method lies and are thus further discussed.

Connolly and colleagues (2013) use a non-linear model to quantify pairwise species interactions while controlling for confounding variables. This approach suffers from the fact that Non-Linear Models can be hard to fit and finding starting values, which are required, can be a great challenge (Venables and Ripley, 2013). This is especially true for non-Linear Mixed Effects Models (Pinheiro and Bates, 2006). The results of this approach are not on the same scale as the response variable and they can be hard to interpret as they cannot be compared to other estimated effects. The element shared with our approach is the fact that the user is enabled to control for the effect of nuisance variables (e.g. soil fertility).

On the other hand, Bell proposed to fit a Linear Model followed by a sequential ANOVA testing procedure (Bell et al., 2009). This publication formalised and further developed an approach that has long been used to analyse diversity-function experiment (Schmid et al., 2002). Put simply, the ANOVA testing procedure is used to test main effects (e.g. species composition) and contrasts within those (e.g. the effect of legumes).

The use of contrasts allows tests for a wide variety of questions that, roughly speaking, can be divided into two groups. Put simply, we can test for non-linearities (e.g. for a non-linear effect of species diversity) and any other test that implies dividing categorical variable into two groups (e.g. the effect of legume presence within species composition).

This approach has the great advantage that there is a clear inferential path for obtaining p-values (Schmid et al., 2002). This is a practically relevant advantage over other ecosystem metrics for which there is no formal way to compute p-values (e.g. $\overline{D}$; Loreau (1998)). In addition, variance components can be derived so that we can compare the relevance of different estimated effects (e.g. compare overyielding with the among-plots variability; Wood (2006)).

However, as Bell and colleagues acknowledged (2009), this approach suffers from an important shortcoming: The fitted models are virtually always overdetermined. This statistical term indicates those models where not all parameters can be estimated. Appendix B expands on this topics further with examples. In diversity-function studies this issue often arises because both species composition and species richness are included in the model as fixed effects. This makes the model non-identifiable and thus overdetermined. Put simply, the problem arises because each level of species composition can only exist within a certain level the species richness (e.g. monoculture A only exists within species richness 1) and therefore the effect of these two predictors cannot be separated. It is worth mentioning that other situations commonly encountered in diversity-function models can lead to models that are overdetermined (see Appendix B).

A practically-relevant inconvenience of overdetermined models is that the estimated coefficients loose their meaning. For example, the species richness coefficient, when considered as a continuous variable of a model that also contains species composition, cannot be interpreted biologically (see Appendix B). Being unable to interpret regression coefficients is very inconvenient as modern statistics advice us to focus on estimates and confidence intervals rather than on simple testing (i.e. p-values; Wasserstein and Lazar (2016); Gelman (2005)). However, it is worth mentioning that, as long as sequential sums of squares are used, the p-value associated with the regression coefficient species richness is correct.

On the other hand, if a meaningful regression coefficient is needed, the only solution is to fit another model where species composition is not present. This is clearly inconvenient as it leads to fitting several models. Moreover, the p-values corresponding to species richness do not match in the two models, as the residual variance used for testing is different (see Appendix B). This is a real disadvantage as there is no way to get a meaningful regression coefficient along with a corresponding correct p-value.

A great advantage of having a model-based framework is that we can control for the effects of confounding variables. For example, we may want to remove the

nuisance effect of soil fertility when testing for the effect of a particular functional group. To do that when using sequential sum of squares, we should introduce the soil fertility term first into the model. Unfortunately, this would make the design non-orthogonal (Schmid et al., 2002). In that case, it is advised to try several orders of the tested variables (Schmid et al., 2002; Venables and Ripley, 2013). Different orders do not give equal results, which in turn, would make the interpretation and reporting of the results difficult. Note also that because the model is overdetermined not all orders can be used. For example species richness, cannot enter the model after species composition as there would be no variance left to explain (Schmid et al., 2002).

In the context of diversity-function studies the main disadvantage of the approach proposed by Bell (2009) is that it does not enable us to compare specific levels of the species composition factor (or any other factor present in the model). This method only allows us to test hypotheses that involve dividing species compositions into two groups. We can test for the effect of a particular functional group (e.g. legumes presence), but not for a simple hypothesis such as comparing the productivity of two monocultures. As a matter of fact, none of the 11 hypotheses tested in Appendix A can be addressed with the approach proposed by Bell.

Our approach follows a slightly different rationale. An identifiable (non-overdetermined) Linear (Mixed) Model is fitted to the data and then post-hoc contrasts are estimated. The term contrast here is defined as its classical statistical meaning, a linear combination of model parameters whose coefficients sum up to zero (Venables and Ripley, 2013). To make a very simple example of a contrast, we can look at the difference in productivity between two monocultures of species A and B and the corresponding two-species mixture (AB). The corresponding contrast is simply

$$C_1 = \frac{1}{2} * \mu_A + \frac{1}{2} * \mu_B + (-1) * \mu_{AB}$$

where μ_i is the mean productivity of the species i as estimated from the fitted model. $\frac{1}{2}$ and (-1) are the contrast coefficients. One of the greatest advantages of this approach is that we can modify the contrast coefficients to test particular hypothesis.

In this example, we may want to test the weighted mean of the two species. If their biomass ratio in the two-species mixture was 80:20 then the contrasts would be.

$$C_2 = \frac{8}{10} * \mu_A + \frac{2}{10} * \mu_B + (-1) * \mu_{AB}$$

This allows us to test one of the main important question addressed in diversity-function studies, overyielding. On top of that, two contrasts can be used to construct a third contrast. Let us imagine that we computed contrast C_2 above, and we found an overyielding of the mixture of 100 g/m^2, at the same time another contrast fitted with two other species, say C and D yielded 113 g/m^2. It may happen that we want to compare formally this two overyielding results and test whether the difference of 13 g/m^2 between the two contrast is statistically different. This can be easily done with our approach (Appendix B; section 4).

This great flexibility in hypothesis testing should be taken into account when interpreting and reporting the results of these tests. Indeed, it should be made clear what contrasts fitted to the data were defined a-priori and which ones where defined a-posteriori (after looking at the data). When the number of fitted contrasts is large, multiple-testing procedures to correct p-values and confidence intervals can be used (Everitt and Hothorn, 2011). The confidence intervals reported on Figure 2.1 are corrected for number of tests performed and also according to their joint multivariate distribution (Hothorn et al., 2016). Other approaches exist to deal with multiple testing (e.g. Rosenthal and Rosnow (1985)).

To summarise, the main advantages of the approach presented here over other model-based approaches are: i) Its flexibility in terms of kind of hypotheses tested; and ii) that our approach makes it possible to obtain meaningful regression coefficients that can be interpreted biologically.

The main advantage of the our model-based approach over marginal metrics such as RYT or $\overline{D}$ is that we can control and remove the confounding effect of external factors within a solid a validated inferential framework (Schmid et al., 2002). We can also take into account sampling effort (e.g. area harvested) or control the variability

brought in by design elements such as plots, blocks or sites. While using a model we can include other omnipresent design elements such as harvest time. Marginal metrics assume that productivity is measured only once (or that it can safely be aggregated to one single meaningful measurement). This is nowadays rarely encountered as most modern studies have a repeated-measures structure. In other words, using a model-based framework we can properly deal with the complexity of modern diversity-function experiments. These experiments are nowadays run outside the laboratory in realistic settings where spatial and temporal variability do exist and can influence the results obtained (Chamagne et al., 2016; Tuck et al., 2016). In addition, this framework can deal with essentially any type of data. This will allow us to analyse a larger set of ecosystem functions than is currently considered and prevents us having a (biased) uni-dimensional view on ecosystem functioning.

Procedure to fit post-hoc contrasts

The procedure used to fit post-hoc contrasts is surprisingly simple and much faster than currently-used metrics. Essentially, the user needs to fit a model where the variable of interest (i.e. species composition) is taken as a fixed effect. If this variable is found to have a significant effect, we can test any hypotheses of interest[4]. Subsequently, to fit a post-hoc contrast we just need to set up a vector of contrasts (see Appendix A).

This way of proceeding is much faster and less error prone than the one used by marginal metrics. To estimate currently-used metrics we may need to perform several steps where averages of groups of observations are estimated. To estimate a simple metric such as overyielding, we may need to average within plot (over time), then within species composition (over plots) and finally average the monocultures. At this point we can start estimating the actual overyielding metric. Finally, we may want to estimate confidence intervals with, for example, bootstrapping techniques (Vojtech et al., 2008). In this case, at least five steps are needed. Each step requires time and increases the probability of introducing errors.

[4]If this predictor is not significant, then there is no evidence for differences among levels of the predictor and the analysis is finished.

On the contrary, post-hoc packages carry out all these steps automatically. Furthermore, if errors were to be present in the procedure of these packages, they are much more likely to be found and corrected. This is especially true for open-source packages such as **R**.

Using dynamic documents (Appendices A and B are such) we can easily report each stage of our analysis (Xie, 2015; Leisch, 2002). A biological hypothesis tested with post-hoc contrasts can be fully reported with a few lines of code (see Appendix A). This implies that there is a gain in transparency, reproducibility and a sharing of coding skills (i.e. it may be very useful to show how a new biological question is actually tested). Marginal metrics may involve the repeated use of control structures such as *for* loops. Although this can be reported in dynamic documents too, it is often hard for the reader to fully understand what is actually done in control structures written by others.

Model parametrisation for diversity-function experiments

It is important to note that to be able to use post-hoc contrasts we cannot use an overdetermined model. This implies that the two biodiversity predictors, species composition and species richness, cannot be included into a model at the same time. A model containing both these predictors is overdetermined and thus non-identifiable. Appendix B shows that the parameters yielded by overdetermined models cannot be interpreted. Post-hoc tests can only be carried out in a model that contains species composition as fixed effect, but not species richness.

Nevertheless, our approach enables the user to test for the effect of species richness (see Section 5 in Appendix A). It also allow us for other predictors not explicitly present in the model (as they would make it overdetermined) as Legume presence. This latter predictor is fully nested into species composition (i.e. a given species composition contains or does not contain legumes). Because one factor is nested in the other, the model becomes non-identifiable (further discussed in Appendix B).

There are four possible options to include species composition and species richness into a model. Appendix B discusses these four options in depth, with their pros and

cons. In particular, we discuss the implications of taking the two biodiversity factors as fixed, as random effects or a mixture of the two. Although, the content of this appendix is somehow technical, its focus is on the biological interpretation of the obtained results. In particular, i) we discourage using models where both predictors are taken as fixed effects if the obtained models coefficients have to interpreted biologically, ii) we recommend taking both predictors as random effects when their relative effects need to be compared(Gelman, 2005), iii) we recommend fitting models with species composition as a random and species richness as a fixed effect when the main interest lays in the effect of diversity or when there is low replication of species compositions and finally iv) we recommend fitting a model with just species composition as a fixed effect when we are interested in comparing the effects of groups of assemblages (e.g. overyielding or legumes presence).

Appendix B further discusses other aspects of fitting overdetermined models. In particular, Section 3 shows that although the parameters of a overdetermined model are biologically meaningless, fitting such models may be required to test for the non-linear effect of species richness. In that section several ways to test for non-linear effects are discussed.

Appendix B discusses important aspects of model parametrisation in diversity-function studies and highlights the importance of fitting identifiable models. The approach presented in this publication requires that the model used is identifiable (non-overdetermined). In the context of diversity-function studies this implies fitting a model with species composition and leaving out species richness.

Discussion

The approach presented here has the great advantage of being very easy to interpret, as the results are in the same scale of the response variable and can easily be transformed into other commonly used quantities such as percentages. In addition, the range of the variable is symmetric and intuitive. Furthermore, in compliance with the modern view of applied statistics, a strong attention is put on the graphical visualisation of the results (Pinheiro and Bates, 2006; Gelman, 2005; Sarkar, 2008). This is certainly

beneficial to the interpretation and sharing of the information yielded in these studies.

The intense use of graphics and confidence intervals to report results makes it possible to compare them within and across studies. At the same time, results are very easy to communicate within and among fields. This is also due to the fact that the method used here, post-hoc contrasts, is not field-specific and known by just a restricted number of people as it is the case for field-specific metrics.

The greatest practical advantage of this approach is that it can deal with the real complexity of modern diversity-function studies. Indeed, thanks to the model-based framework used, we can account for the confounding of covariates, but also include fundamental design variables or the sampling effort used. This has two main advantages, on one side we can better estimate the biological process of interest while on the hand we can also judge the practical relevance of the estimated effects.

The vast majority of the marginal metrics used to compare the productivity of ecosystems assume ideal, nowadays unrealistic, assumptions. Modern diversity-function experiments are run in more realistic conditions that the first generation of laboratory experiments, where everything was controlled.

This situation was greatly improved by two publications (Connolly et al., 2013; Bell et al., 2009) that formalised the previously used concept of model-based framework (Schmid et al., 2002). however, these are not appropriate to measure important quantities such as overyielding or transgressive overyielding.

The approach presented here can easily deal with the questions currently addressed in diversity-function studies. To cite a few examples, using post-hoc contrasts, it is straightforward to test for overyielding, for the effect of a particular functional group, as well as for the effect of species richness (although this is not explicitly present in the model; see section 5 in Appendix A). In addition, study-specific questions can easily be addressed using the same framework, but most importantly any new biological question can be turned into a post-hoc contrast in a simple and unequivocal manner.

Having a model-based framework that relies on an established technique brings the great advantage that there is a clear inferential path, and that modern and powerful

implementations are available along with a wide range of literature and documentation. On top of that, the procedure used is fast and less error-prone.

Finally, essentially any kind of data can be analysed using exactly the same framework. This makes it possible to analyse a wider variety of ecosystem functions, and not just focus on ecosystem productivity. This is supposed to greatly improve our understanding of ecosystem functioning. As the results are easy to compare even across types of data, we also expect to have a better understanding of how functions relate to each other. Furthermore, our approach is well-suited to be used simultaneously with several functions (multi-functionality, Lefcheck et al. (2015)).

Conceptually, the method presented here takes the best parts of each of the other approaches discussed in this publication. As marginal metrics do, it tries to condense the information contained in the data to yield one or few meaningful values. It takes the conditional approach introduced by Connolly (2013) and extends it, so that the effect of other covariates can be removed. It is inspired by the method proposed by Bell (2009) to enable a clear inferential path (p-values and confidence intervals). On top of these helpful features, our method is vastly more flexible and powerful, and its results are easier to interpret biologically than currently used metrics.

Conclusions

In this publication, we propose to modify and modernise the concept used in the comparison of ecosystem productivity. Most currently-used metrics aim to condense complex biological processes into a single meaningful value. This is done while making many assumptions that are no longer met in modern diversity-function experiments as they are not run in controlled conditions any more.

We propose to use a mainstream statistical method, post-hoc contrasts, to better deal with realistic conditions encountered in modern diversity-function experiments. The tailored use of this established technique makes it possible to address, within a single unified framework and in a non-redundant way, a large variety of questions. As shown with real data, this model-based approach is more powerful than marginal metrics.

The most important practical advantage of this approach is a straightforward and intuitive interpretation of the results. This is expected to improve the sharing of information within and among scientific fields. We expect a great improvement of communication between scientists and policy-makers which in turn should boost the implementation of evidence-based policies.

2.3 Appendices

Appendix A, starting at page 35, shows how the method presented in the main text can be used. Several examples are shown. Appendix B, starting at page 65, illustrates the concept of overdetermined models with examples. In particular, the implications of fitting overdetermined models in the context of diversity-functions studies are discussed. Appendix A is particularly relevant for this chapter.

Note on all the appendices of this thesis All appendices in this thesis have been produced with the software *knitr* (Xie, 2016). This latter combines the typesetting language LaTeX and the statistical software **R** to produce dynamic documents (here pdf documents; Xie (2015)). Every Appendix is produced in a stand-alone **R**-session. This implies that package versions may differ among appendices. Therefore, to enable full reproducibility, each appendix has its own bibliography section.

All appendices pages are internally numbered at the top of the page (page number within two dashes starting at page 2). Any reference to pages or sections found in the appendices of this thesis refer to the internal appendix page numbering. Nevertheless, the thesis numbering is also present at the bottom of each page of all appendices.

Appendix A: Examples

Chapter 2

Contents

1 Introduction

In the main text of this publication we argue that post-hoc contrasts can be used to compare productivity of ecosystems differing in their species composition. In this appendix we introduce this statistical technique, we show a wide variety of examples and discuss the its advantages and disadvantages.

Post-hoc contrasts are used to test any hypothesis that involves comparing levels of a categorical variable. These tests are used in the framework of Linear Models and their extensions. Put simply, if a categorical variable present in a linear model is found to be significant, this implies that at least one level differs from the other. However, the p-value associated with that categorical variable does not tell us which level or levels differ from the others. For that reason, post-hoc tests, also named post-hoc contrasts, are used to test specific hypotheses that involve comparing levels of a categorical variable.

In the context of diversity-function studies, this technique is particularly useful to compare levels of the *species composition* factor. The simplest example involves comparing the productivity of species A and B to test whether they differ. However, as we will show in this document, post-hoc contrasts can be used to test very specific and complex hypotheses.

After this brief introduction, in Section 2, we are going to show several examples of post-hoc contrasts with a relatively simple model that can be considered representative of the analysis of diversity-function studies. Subsequently, Section 3 shows how population-level data can be analysed with post-hoc contrasts. Section 4 shows the use of post-hoc contrasts with a complex model. Section 5 show how to test for the effect of species richness. Section 6 discusses the versatility of this approach in terms of kind of data that can be analysed. Section 7 shows how to take into account important study peculiarities such as sampling effort. Finally, Section 8 makes some important remarks about this document and on the tailored use of post-hoc contrasts to analyse diversity-function studies.

2 Using post-hoc contrasts with a simple model

2.1 The Zurich data set

In this section, we are going to use a real data set that comes from a diversity-function experiment on five perennial grass species (Poaceae). In this experiment, which was run in Zurich (Switzerland) five grass species found in European fertile meadows were left to grow over a period of five years (2004-2008). These species were grown in monocoltures, all ten two-species mixtures and the full mixture (i.e. five species). Each one of the 16 species compositions was replicated five times (i.e. with five plots each). The biomass of each of the 80 plots was measured eight times over a period of five years. More details about this study can be found in Vojtech et al. (2008). A detail of this study that is worth mentioning, is that the surface harvested in years 2004, 2005 and 2006 was 0.25 m^2, while later in years 2007 and 2008 only 0.09 m^2 were harvested.

To deal with that, we multiplied the measured biomasses so that all observations refer to 1 m^2.

The first step of any analysis using our approach is the graphical visualisation of the data. Here, we start with a graph that gives us an overall view of the data while highlighting the different species compositions. We will then produce further tailored graphs when testing specific hypotheses. For the sake of brevity, the **R**-code used to produce graphs is not shown here. However, these lines of code can be found in the **R**-code file attached.

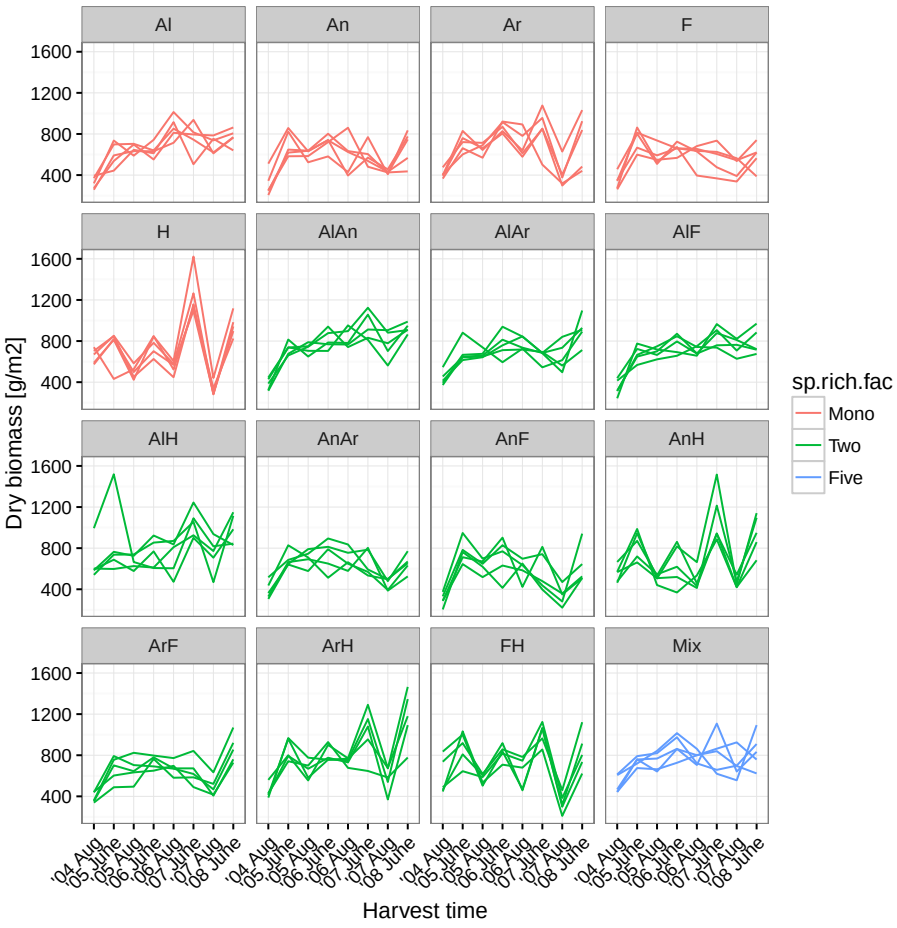


Figure 1: Biomass of the Zurich experiment plotted against time. Each species composition (see panels) has been replicated five times. Species richness is highlighted with colours.

This graph gives a good overall impression of the performance over time of the 16 species compositions used (acronyms are defined in Vojtech et al. (2008)). To produce this graph the add-on package *ggplot2* was used (Wickham and Chang, 2016).

2.2 Fitting a simple model

As mentioned above in the introduction, post-hoc contrasts can be used on a Linear Model or essentially any extension to it. To model the Zurich data set, we start fitting

a simple Linear Mixed Effects Model. We model the effect of species composition (`sp.comp`) and harvest time (`time.fac`) as fixed categorical variables. `sp.comp` has 16 levels, while `time.fac` has eight levels. Note that we do not allow `sp.comp` and `time.fac` to interact. We thus assume that the effect of harvest time is the same for all species compositions. We are going to relax this assumption in Section 4. Because each replicate of any species composition (i.e. plot) is repeatedly measured, we also introduce a random effect for plots (i.e. `plot`). The response variable is the dry biomass per m^2 (`bio.m2`). Note that the predictor species richness is not explicitly included in the model as this would make the model overparametrised. Appendix B discusses this issue in depth. The Mixed Model is fitted using the add-on package *lme4* (Bates et al., 2016).

```
require(lme4)
mod.simple <- lmer(bio.m2 ~ time.fac + sp.comp + (1 | plot), data = d.zh)
```

After fitting this model, we can look at the summary information to get an overall view of the results.

```
summary(mod.simple)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: bio.m2 ~ time.fac + sp.comp + (1 | plot)
   Data: d.zh

REML criterion at convergence: 8068

Scaled residuals:
   Min     1Q Median     3Q    Max
-2.545 -0.646  0.010  0.585  4.835

Random effects:
 Groups   Name        Variance Std.Dev.
 plot     (Intercept)     1.77   1.33
 Residual             24244.86 155.71
Number of obs: 640, groups:  plot, 80

Fixed effects:
                 Estimate Std. Error t value
(Intercept)        419.87      29.52   14.22
time.fac'05 June   305.27      24.62   12.40
time.fac'05 Aug    199.37      24.62    8.10
time.fac'06 June   316.40      24.62   12.85
time.fac'06 Aug    243.79      24.62    9.90
time.fac'07 June   394.22      24.62   16.01
time.fac'07 Aug     96.95      24.62    3.94
time.fac'08 June   390.26      24.62   15.85
sp.compAn          -85.74      34.83   -2.46
sp.compAr            4.74      34.83    0.14
sp.compF           -94.62      34.83   -2.72
sp.compH            55.01      34.83    1.58
sp.compAlAn        100.98      34.83    2.90
sp.compAlAr         30.89      34.83    0.89
sp.compAlF          33.63      34.83    0.97
sp.compAlH         140.53      34.83    4.04
sp.compAnAr        -41.65      34.83   -1.20
sp.compAnF         -95.63      34.83   -2.75
sp.compAnH          28.73      34.83    0.83
```

```
sp.compArF         -22.24      34.83   -0.64
sp.compArH         121.89      34.83    3.50
sp.compFH           47.69      34.83    1.37
sp.compMix          90.99      34.83    2.61
```

The summary indicates that the variability explained by the random effect `plot` is considerably smaller than the residual variance (standard deviations are resp. 1.3 and 155.7). Regarding the fixed effects, we see that there are important differences among harvests, but also among species compositions (by default **R** uses treatment contrasts matrices (Venables and Ripley, 2013)). Therefore, the intercept in this summary output refers to the first alphabetical level of the species composition and harvest time factors. In other words, the value of the parameter `(Intercept)` (i.e. 419.87) refers to the estimated yield of the monoculture `Al` in August 2004. Any other parameter of the species composition factor indicates the difference from this reference level. To make an example, `sp.compAn` quantifies the difference between the monoculture `Al` and the monoculture `An` (at any time point, as in model the effect of time and species composition is additive).

By looking at this summary output, we can only make a limited number of claims about levels of the `sp.comp` factor. For example, we can cannot say whether the most productive monoculture (i.e. `H`) significantly differs from the least productive monoculture `F`. To answer this kind of question we can use post-hoc contrasts.

Before starting to use this technique we must make sure that there is evidence for differences among levels of the species composition factor. To do that, we formally test this hypothesis.

```
mod.simple.2 <- update(mod.simple, . ~ . - sp.comp)
anova(mod.simple, mod.simple.2)
```

```
Data: d.zh
Models:
mod.simple.2: bio.m2 ~ time.fac + (1 | plot)
mod.simple: bio.m2 ~ time.fac + sp.comp + (1 | plot)
             Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
mod.simple.2 10 8364 8409  -4172     8344
mod.simple   25 8304 8416  -4127     8254  89.9     15      1e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The testing procedure confirms what we observed in Figure 1, there are substantial differences among species compositions. We can thus proceed and use post-hoc contrasts[1].

2.3 Testing biological hypotheses with post-hoc contrasts

In this subsection we are going to test several biological hypotheses commonly encountered in the analysis of diversity-function studies.

[1] If there had been no evidence for an effect of species composition, there would be nothing to test, so the analysis would be finished.

Hypothesis 1: Comparing two monocultures

We start by fitting the simplest post-hoc test. We compare whether the productivity of species `Al` differs from that of `An`. Figure 1 indicates that `Al` may be, on average, more productive than `An`. To test this hypothesis we are going to use the add-on package *multcomp* (Hothorn et al., 2016). The workhorse function of this package is `glht()`. This function needs to be fed with a vector of contrasts. The vector, which is as long as there are groups in the species composition factor (i.e 16), must contain a non-zero entry for the groups that we want to compare. In this case, we are simply comparing two levels and therefore these entries are set to -1 and 1 respectively[2].

```
vec.1 <- c(1, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
names(vec.1) <- levels(d.zh$sp.comp)
vec.1
```

```
  Al   An   Ar    F    H AlAn AlAr  AlF  AlH AnAr  AnF  AnH  ArF  ArH   FH
   1   -1    0    0    0    0    0    0    0    0    0    0    0    0    0
 Mix
   0
```

The names given to the contrasts vector make it clear that we are simply comparing two levels of the species composition factor. We can now fit the contrast, save it into an object named `pht.1` (Post Hoc Test 1) and look at the summary information.

```
require(multcomp)
pht.1 <- glht(model = mod.simple, linfct = mcp(sp.comp = vec.1))
summary(pht.1)
```

```

 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: lmer(formula = bio.m2 ~ time.fac + sp.comp + (1 | plot), data = d.zh)

Linear Hypotheses:
       Estimate Std. Error z value Pr(>|z|)
1 == 0     85.7       34.8    2.46    0.014 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

This output tells us that the monoculture `An` produces on average 85.74 g/m^2 less than the monoculture `Al`. Not surprisingly, this result is in full agreement with the summary output of `model.simple` shown on page 4 (see row `sp.compAn` in the fixed effects). The summary information confirms that the productivity of these two species can be considered statistically different.

Note that here we took the difference `Al - An`. As `Al` is more productive than `An` we obtained a positive value. We may want to take the opposite contrast by simple changing the sign of the non-zero entries of the contrasts vectors. Obviously, the

[2]The sum of the contrast coefficients must be zero.

obtained results are identical up to the sign of the estimated difference.

The test carried out here, is testing the biological hypothesis that these two monocultures have the same productivity. Expressed mathematically:

$$H0_1 : \quad \beta_{Al} = \beta_{An}$$

It is important to note that in this framework, there is no redundancy. This is the only contrast that can be used to test this biological hypothesis.

Note that the `glht()` function is very flexible and accepts several syntaxes (type `?glht` for more details). "Symbolic description" is another syntax available. As the next example shows, this latter is extremely intuitive to use.

```
pht.1.bis <- glht(model = mod.simple, linfct = mcp(sp.comp = c("Al - An = 0")))
```

This procedure is fully equivalent to the one used previously to produce the object `pht.1` and, indeed, their results are identical (not shown). Although, the syntax used to produce `glht.1.bis` is more intuitive, it is less convenient to test complex analysis. For that reason, it is not further used.

Hypothesis 2: Overyielding

Probably the most common question addressed in diversity-function analyses is whether the full-species mixture is more productive than the monocultures (i.e. overyielding). To test this hypothesis, the relative abundances of all species in the mixture are used to estimate weights. To start simply, we assume that all species are equally abundant in the full-species mixture. We will show how to relax this assumption in Section 3. Before testing this hypothesis we must a create a tailored graph that visualises the data in a well-suited way.

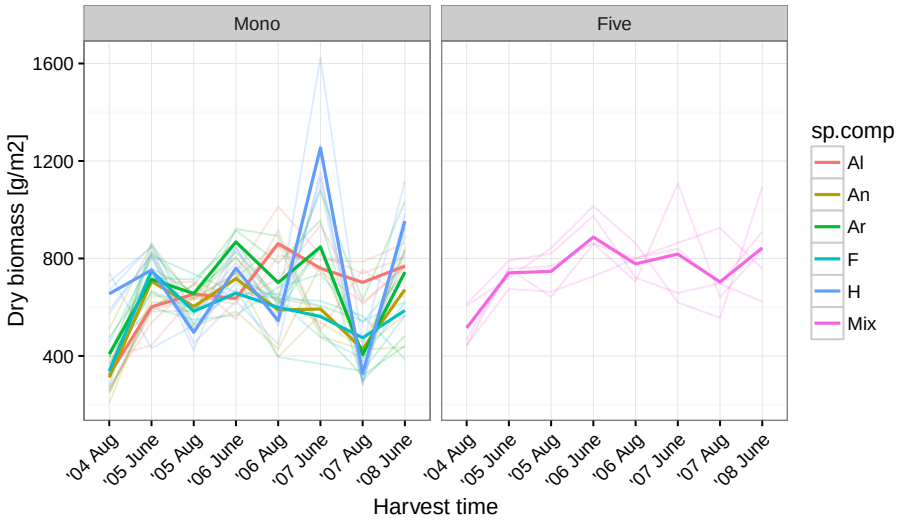


Figure 2: Biomass of the five monocultures and the full-mixture composition plotted against time. Thick lines represent species composition averages (see legend), while thin transparent lines represent individual plots.

This graph indicates that the full-species mixture may, indeed, be more productive than the average monoculture. Roughly speaking, we are testing whether there is a difference

between the parameters (i.e. ≈ mean values) associated with the five monocultures and the full-species mixture. The six compositions involved in this test appear at the first five places (monocultures) and at the last place (full-species mixture) of the species composition levels.

```
vec.2 <- c(-1/5, -1/5, -1/5, -1/5, -1/5, 0,0,0,0,0,0,0,0,0,0, 1)
pht.2 <- glht(mod.simple, linfct = mcp(sp.comp = vec.2))
summary(pht.2)


 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: lmer(formula = bio.m2 ~ time.fac + sp.comp + (1 | plot), data = d.zh)

Linear Hypotheses:
       Estimate Std. Error z value Pr(>|z|)
1 == 0      115         27    4.27    2e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

The coefficient used, respectively $-1/5$ and 1 indicate that the average of the five monocultures is compared with the full-species mixture. In general, when comparing two groups of levels (here the monocultures and the full-species mixture), the sum of the coefficients in each group must be either 1 or -1.

The full-species mixture yields about 115 grams per m^2 more than the averaged monocultures. The low p-value associated with this test indicates that there is thus strong evidence for overyielding. To better interpret these results we may want to estimate a 95% confidence interval for the estimated overyielding. To do that, we can make use of the `confint()` method function[3].

```
confint(pht.2)


 Simultaneous Confidence Intervals

Multiple Comparisons of Means: User-defined Contrasts


Fit: lmer(formula = bio.m2 ~ time.fac + sp.comp + (1 | plot), data = d.zh)

Quantile = 2
95% family-wise confidence level


Linear Hypotheses:
       Estimate lwr     upr
```

[3]Other useful methods for objects created by the `glht()` function can be found by typing `methods(class = "glht")`.

```
1 == 0 115.114   62.239 167.988
```

We notice here that the confidence interval for the overyielding effect is fairly broad. We may report that result as "There is statistical evidence for overyielding". However, whether this results is practically or biologically relevant depends on the context.

Hypothesis 3: Transgressive overyielding
At this stage, we may want to further test whether transgressive overyielding is present. In other words, whether the full-species mixture performs better than the most productive monoculture (i.e. `H`). In this case, we are testing a one-sided hypothesis. This is statistically more powerful than a two-sided test. Note the use of `alternative = "greater"` in the function call below.

```
pht.3 <- glht(mod.simple,
              linfct = mcp(sp.comp = c(0,0,0,0, -1, 0,0,0,0,0,0,0,0,0,0, 1)),
              alternative = "greater")
summary(pht.3)


 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: lmer(formula = bio.m2 ~ time.fac + sp.comp + (1 | plot), data = d.zh)

Linear Hypotheses:
       Estimate Std. Error z value Pr(>z)
1 <= 0     36.0       34.8    1.03   0.15
(Adjusted p values reported -- single-step method)
```

Although the full-species mixture yields more than the best monoculture, there no statistical evidence, at the 5% level, that transgressive overyielding is present. The hypothesis tested here is one-sided, mathematically this can be written as:

$$H0_3: \quad \beta_H < \beta_{Mix}$$

Hypothesis 4: Simultaneously testing several hypotheses
Up to this point we have been testing single hypotheses, Nevertheless, it is usual practice to test for several hypotheses in each analysis of diversity-function studies. The `glht()` function can simultaneously test several hypotheses. To do that, we must feed the function with a matrix of contrasts instead of a single vector. In practice, each row of that matrix represent a contrast (i.e. a biological question). We will test whether *i)* species `H` is statistically distinguishable from species `F`, *ii)* there is evidence that the yield of species `Al` and `An` differs from species `Ar` and `H`, and finally *iii)* whether the most productive species (`H`) is distinguishable from all other four monocultures. In this example, this matrix of contrasts needed has three rows and 16 columns (i.e. the number of levels present in the species composition factor). Rows and columns are named so that contrasts can be easily identified.

```
M.4 <- rbind("H vs F"            = c(0,0,0, -1, 1, rep(0, times = 11)),
             "Al & An vs Ar & H" = c(1/2, 1/2, -1/2, 0, -1/2, rep(0, times = 11)),
             "H vs others"       = c(-1/4, -1/4, -1/4, -1/4, 1, rep(0, times = 11)))
colnames(M.4) <- levels(d.zh$sp.comp)
##
M.4
```

```
                     Al    An    Ar     F    H AlAn AlAr AlF AlH AnAr AnF
H vs F             0.00  0.00  0.00 -1.00  1.0    0    0   0   0    0   0
Al & An vs Ar & H  0.50  0.50 -0.50  0.00 -0.5    0    0   0   0    0   0
H vs others       -0.25 -0.25 -0.25 -0.25  1.0    0    0   0   0    0   0
                  AnH ArF ArH FH Mix
H vs F              0   0   0  0   0
Al & An vs Ar & H   0   0   0  0   0
H vs others         0   0   0  0   0
```

```
pht.4 <- glht(mod.simple, linfct = mcp(sp.comp = M.4))
summary(pht.4)
```

```
 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: lmer(formula = bio.m2 ~ time.fac + sp.comp + (1 | plot), data = d.zh)

Linear Hypotheses:
                       Estimate Std. Error z value Pr(>|z|)
H vs F == 0               149.6       34.8    4.30   <0.001 ***
Al & An vs Ar & H == 0    -72.7       24.6   -2.95   0.0084 **
H vs others == 0           98.9       27.5    3.59   <0.001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

All three biological hypotheses are statistically significant. It is important to note that, as the last line of the summary output indicates, the reported p-values are corrected for multiple testing[4]. This enables the user to automatically solve the issue of p-value correction while testing several hypotheses simultaneously. A large variety of testing corrections are available and implemented in the `glht()` function (see argument `test` of the `summary.glht()` method function). Default settings are recommended.

Hypothesis 5: Breaking down hypotheses

An interesting feature of the approach presented here is that we can break down biological hypotheses. For example, we may want to know whether all two-species mixtures yield more than the corresponding monocultures.

[4]P-values are corrected for number of tests carried out and for their joint multivariate distribution (i.e. their correlation) (Hothorn et al., 2016).

```
M.5 <- rbind("Al & An vs AlAn" = c(-1/2, -1/2, 0,0,0, 1, rep(0, times = 10)),
             "Al & Ar vs AlAr" = c(-1/2, 0, -1/2, 0,0,0, 1, rep(0, times = 9)),
             "Al & F vs AlF"   = c(-1/2, 0,0, -1/2, 0,0,0, 1, rep(0, times = 8)),
             "Al & H vs AlH"   = c(-1/2, 0,0,0, -1/2, 0,0,0, 1, rep(0, times = 7)),
             "An & Ar vs AnAr" = c(0, -1/2, -1/2, 0,0,0,0,0,0, 1, rep(0, times = 6)),
             "An & F vs AnF"   = c(0, -1/2, 0, -1/2, 0,0,0,0,0,0, 1, rep(0, times = 5)),
             "An & H vs AnH"   = c(0, -1/2, 0,0, -1/2, 0,0,0,0,0,0, 1, rep(0, times = 4)),
             "Ar & F vs ArF"   = c(0,0, -1/2, -1/2, 0,0,0,0,0,0,0,0, 1, rep(0, times = 3)),
             "Ar & H vs ArH"   = c(0,0, -1/2, 0, -1/2, 0,0,0,0,0,0,0,0, 1, rep(0, times = 2)),
             "F & H vs FH"     = c(0,0,0, -1/2, -1/2, 0,0,0,0,0,0,0,0,0, 1, 0))
colnames(M.5) <- levels(d.zh$sp.comp)
```

```
M.5
```

	Al	An	Ar	F	H	AlAn	AlAr	AlF	AlH	AnAr	AnF	AnH
Al & An vs AlAn	-0.5	-0.5	0.0	0.0	0.0	1	0	0	0	0	0	0
Al & Ar vs AlAr	-0.5	0.0	-0.5	0.0	0.0	0	1	0	0	0	0	0
Al & F vs AlF	-0.5	0.0	0.0	-0.5	0.0	0	0	1	0	0	0	0
Al & H vs AlH	-0.5	0.0	0.0	0.0	-0.5	0	0	0	1	0	0	0
An & Ar vs AnAr	0.0	-0.5	-0.5	0.0	0.0	0	0	0	0	1	0	0
An & F vs AnF	0.0	-0.5	0.0	-0.5	0.0	0	0	0	0	0	1	0
An & H vs AnH	0.0	-0.5	0.0	0.0	-0.5	0	0	0	0	0	0	1
Ar & F vs ArF	0.0	0.0	-0.5	-0.5	0.0	0	0	0	0	0	0	0
Ar & H vs ArH	0.0	0.0	-0.5	0.0	-0.5	0	0	0	0	0	0	0
F & H vs FH	0.0	0.0	0.0	-0.5	-0.5	0	0	0	0	0	0	0

	ArF	ArH	FH	Mix
Al & An vs AlAn	0	0	0	0
Al & Ar vs AlAr	0	0	0	0
Al & F vs AlF	0	0	0	0
Al & H vs AlH	0	0	0	0
An & Ar vs AnAr	0	0	0	0
An & F vs AnF	0	0	0	0
An & H vs AnH	0	0	0	0
Ar & F vs ArF	1	0	0	0
Ar & H vs ArH	0	1	0	0
F & H vs FH	0	0	1	0

```
pht.5 <- glht(mod.simple, linfct = mcp(sp.comp = M.5))
summary(pht.5)


 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: lmer(formula = bio.m2 ~ time.fac + sp.comp + (1 | plot), data = d.zh)

Linear Hypotheses:
                     Estimate Std. Error z value Pr(>|z|)
Al & An vs AlAn == 0   143.85      30.16    4.77   <1e-04 ***
Al & Ar vs AlAr == 0    28.52      30.16    0.95   0.9822
Al & F vs AlF == 0      80.94      30.16    2.68   0.0692 .
Al & H vs AlH == 0     113.03      30.16    3.75   0.0018 **
```

```
An & Ar vs AnAr == 0    -1.15      30.16   -0.04   1.0000
An & F vs AnF == 0      -5.46      30.16   -0.18   1.0000
An & H vs AnH == 0      44.10      30.16    1.46   0.7703
Ar & F vs ArF == 0      22.70      30.16    0.75   0.9969
Ar & H vs ArH == 0      92.02      30.16    3.05   0.0224 *
F & H vs FH == 0        67.49      30.16    2.24   0.2188
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Although only a few contrasts are significant, we note that almost all estimated contrasts are positive. We are thus led to believe that, in general, pairwise mixtures are more productive than the corresponding monocultures. The output presented here is relatively long and hard to read as a lot of information is presented. Fortunately, fitted contrasts can be visualised graphically with very little effort. This can be very useful to report the results in scientific publications. In addition, the results are visualised in terms of confidence intervals which is often preferred as these carry more information that p-values (Wasserstein and Lazar, 2016). Note that plot margins are modified to make the contrasts names fit. In addition, the range of the x-axis is modified so that we obtain a symmetric graph centred at zero (i.e. no difference).

```
par(mar = c(5.1, 8.1, 4.1, 2.1))  ## to make names fit
plot(pht.5, xlim = c(-220, 220))
```

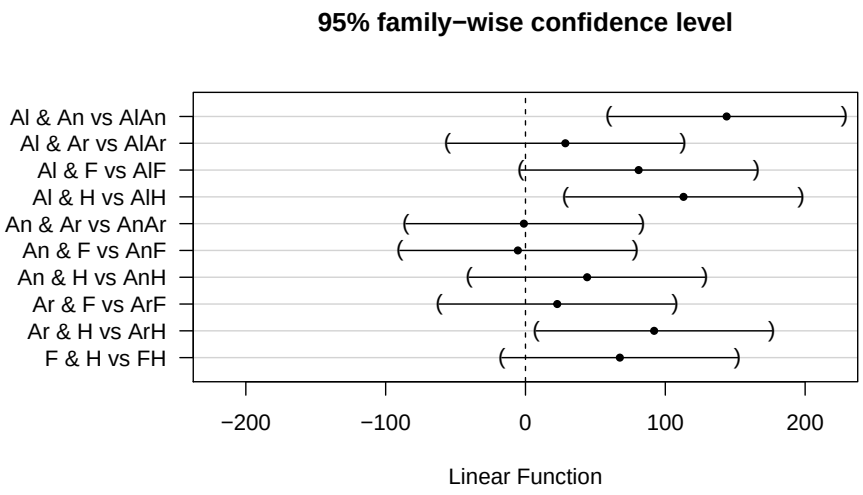


Figure 3: Fitted contrasts for all pairs of monocultures against the corresponding two-species mixtures. Estimates are shown along with confidence intervals.

Figure 3 clearly indicates that on average two-species mixtures are more productive than their corresponding monocultures. This graph also indicates that only three of these contrasts are significant after multiple testing is taken into account. Interestingly, the four two-species mixtures containing the species `Al` seem to show higher increase in biomass compared to the monoculture than mixtures with other species. Note that the width of these confidence intervals is identical as the design is fully balanced and the same number of observations is used to compute each contrast.

At this stage, we may want to formally test whether, in general, two-species mixtures are more productive than the corresponding monocultures. In other words, we may want to draw a general conclusion from all the ten contrasts fitted here. To do that, we can simply collapse the matrix of contrasts used here (i.e. `M.5`) into a single vector of contrasts. Subsequently, we need to make sure that in each group (i.e. monocultures and two-species mixtures) the vector of coefficients sums up to either 1 or -1. We therefore divide the summed coefficients by ten (i.e. as there were ten contrasts in the matrix before collapsing it).

```
## 1. collapsing the matrix into a vector
vec.5 <- apply(M.5, MARGIN = 2, function(x) sum(x)/10)
##
## 2. attach this vector the the M.5 matrix
M.5.extended <- rbind(M.5, vec.5)
##
## 3. rename the last row of the newly obtained matrix
rownames(M.5.extended)[nrow(M.5.extended)] <- "general hypothesis"
```

```
pht.5.extended <- glht(mod.simple, linfct = mcp(sp.comp = M.5.extended))
##
par(mar = c(5.1, 8.1, 4.1, 2.1))  ## to make names fit
plot(pht.5.extended, xlim = c(-220, 220))
```

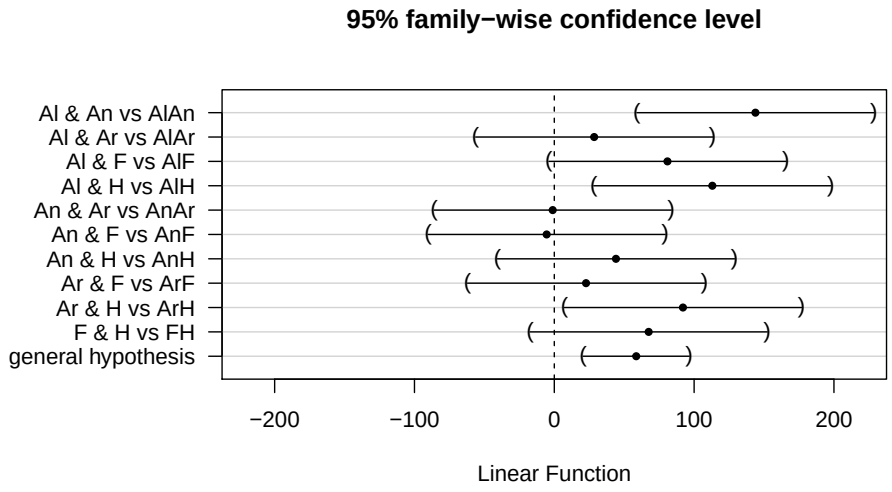


Figure 4: Fitted contrasts for all pairs of monocultures against the corresponding two-species mixtures. The general hypothesis is shown in the last row. Estimates are shown along with confidence intervals.

The results of the updated contrasts presented graphically in Figure 4 confirm that overyielding is present when two-species compositions are considered. Indeed, the confidence interval of the last row, the general hypothesis, does not include zero. This visualisation is very powerful and intuitive. Reporting these results as a table or in written text is far less efficient at conveying information. Note that the confidence interval of the last row is narrower compared to all others as it is based on more observations.

Hypothesis 6: Combining contrasts into new contrasts
In the previous example we have seen that we can break down hypotheses into their single components and then easily test for the general hypothesis. Here we present another example where a general hypothesis is tested by modifying the matrix of contrasts used in the previous example. As seen in Figure 3, we may think that species `Al` seems to contribute to a greater overyielding than the other species. We may thus want to test whether the gain obtained by putting `Al` with any other species is larger than the gain obtained in all other two-species mixtures not containing `Al`. A useful application of this type of contrasts is to test the effect of specific functional groups (e.g. Leguminosae).

To fit these contrasts we use the matrix of contrasts used in the previous example (i.e. `M.5`) where we tested for the gain of two-species mixtures over monocultures. In this matrix the first four rows concern two-species mixtures containing `Al`, while the last six rows concern all other species. We are therefore going to contrast the average gain of the first four rows and the the average gain with the last six rows.

```
## 1. extract rows that contain "Al" and collapse them
M.6.Al <- M.5[1:4, ]
vec.6.Al <- apply(M.6.Al, 2, sum) / 4
##
## 2. extract rows that do not contain "Al" and collapse them
M.6.others <- M.5[5:10, ]
vec.6.others <- apply(M.6.others, 2, sum) / 6
##
## 3. create the new matrix with these two vectors
M.6 <- rbind("Al & X vs AlX" = vec.6.Al,
             "X & Y vs XY"   = vec.6.others)
```

Note that when collapsing the matrices, we made sure that the sum of the coefficients in each group is 1/-1. This was achieved by dividing the coefficients by the number of contrasts involved (respectively four and six). We can now fit a contrast to test this hypothesis formally.

```
pht.6 <- glht(mod.simple, linfct = mcp(sp.comp = M.6))
summary(pht.6)


 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: lmer(formula = bio.m2 ~ time.fac + sp.comp + (1 | plot), data = d.zh)

Linear Hypotheses:
                   Estimate Std. Error z value Pr(>|z|)
Al & X vs AlX == 0     91.6       18.5    4.96  1.4e-06 ***
X & Y vs XY == 0       36.6       15.9    2.30    0.041 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

This output tells us that the two-species mixtures that contain `Al` gain, on average, 91.6 g/m^2 more than the corresponding averaged monocultures. It also tells us that this gain is smaller for the other two-species mixtures that do not contain `Al`. At this stage, we would like to test whether the gain observed for `Al` is statistically larger than the gain of the other two-species mixtures. To do that we can simply take the difference of the two vectors and run a formal test again.

```
## 1. take the difference
vec.6.diff <- vec.6.Al - vec.6.others
##
## 2. add this new vector to M.6 matrix
M.6.extended <- rbind(M.6, "Difference in gain" = vec.6.diff)
##
## 3. fit the model formally and look at inference
pht.6.extended <- glht(mod.simple, linfct = mcp(sp.comp = M.6.extended))
summary(pht.6.extended)


 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: lmer(formula = bio.m2 ~ time.fac + sp.comp + (1 | plot), data = d.zh)

Linear Hypotheses:
                        Estimate Std. Error z value Pr(>|z|)
Al & X vs AlX == 0          91.6       18.5    4.96   <0.001 ***
X & Y vs XY == 0            36.6       15.9    2.30    0.054 .
Difference in gain == 0     55.0       21.0    2.61    0.024 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

On average, the gain obtained by putting the species `Al` with any other species is in a two-species mixture is about 55 g/m^2 larger than the gain obtained by putting together any other species in a pairwise manner. This difference is statistically significant at the 5% level.

Hypothesis 7: Fitting all-pairwise comparisons (*)

(*) Starred paragraphs, subsections and sections indicate a higher technical content or that these parts of the document are not fundamental to the approach presented here. These parts are presented for the sake of completeness. However, the reader interested in a quick read can skip them.

The arsenal of tools made available by the `glht()` function is extremely wide. To make an example, we can also run Tukey Honest Significant Difference tests. This test compares all pairs of levels of a factor while accounting for multiple testing. We are applying this technique to the categorical variable harvest time.

```
pht.7 <- glht(mod.simple, linfct = mcp(time.fac = "Tukey"))
summary(pht.7)
```

```
 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lmer(formula = bio.m2 ~ time.fac + sp.comp + (1 | plot), data = d.zh)

Linear Hypotheses:
                         Estimate Std. Error z value Pr(>|z|)
'05 June - '04 Aug == 0    305.27      24.62   12.40   <0.001 ***
'05 Aug - '04 Aug == 0     199.37      24.62    8.10   <0.001 ***
'06 June - '04 Aug == 0    316.40      24.62   12.85   <0.001 ***
'06 Aug - '04 Aug == 0     243.79      24.62    9.90   <0.001 ***
'07 June - '04 Aug == 0    394.22      24.62   16.01   <0.001 ***
'07 Aug - '04 Aug == 0      96.95      24.62    3.94   0.0022 **
'08 June - '04 Aug == 0    390.26      24.62   15.85   <0.001 ***
'05 Aug - '05 June == 0   -105.90      24.62   -4.30   <0.001 ***
'06 June - '05 June == 0    11.13      24.62    0.45   0.9998
'06 Aug - '05 June == 0    -61.48      24.62   -2.50   0.1965
'07 June - '05 June == 0    88.95      24.62    3.61   0.0074 **
'07 Aug - '05 June == 0   -208.32      24.62   -8.46   <0.001 ***
'08 June - '05 June == 0    84.99      24.62    3.45   0.0134 *
'06 June - '05 Aug == 0    117.03      24.62    4.75   <0.001 ***
'06 Aug - '05 Aug == 0      44.43      24.62    1.80   0.6171
'07 June - '05 Aug == 0    194.85      24.62    7.91   <0.001 ***
'07 Aug - '05 Aug == 0    -102.42      24.62   -4.16   <0.001 ***
'08 June - '05 Aug == 0    190.89      24.62    7.75   <0.001 ***
'06 Aug - '06 June == 0    -72.61      24.62   -2.95   0.0634 .
'07 June - '06 June == 0    77.82      24.62    3.16   0.0336 *
'07 Aug - '06 June == 0   -219.45      24.62   -8.91   <0.001 ***
'08 June - '06 June == 0    73.86      24.62    3.00   0.0546 .
'07 June - '06 Aug == 0    150.43      24.62    6.11   <0.001 ***
'07 Aug - '06 Aug == 0    -146.85      24.62   -5.96   <0.001 ***
'08 June - '06 Aug == 0    146.46      24.62    5.95   <0.001 ***
'07 Aug - '07 June == 0   -297.27      24.62  -12.07   <0.001 ***
'08 June - '07 June == 0    -3.96      24.62   -0.16   1.0000
'08 June - '07 Aug == 0    293.31      24.62   11.91   <0.001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

There is strong evidence that almost all harvests have to be considered different from each other. This highlights the importance of taking this variable into account when testing biological hypotheses such as those tested here, so that its confounding effect can be removed.

3 Using relative abundances to estimate post-hoc contrasts

3.1 Fitting contrasts with relative abundances

Above we computed overyielding while making the implicit assumption that all species are equally abundant in the mixtures. This may not be true, and as a consequence, the results can be misleading. Indeed, it may be that the best performing species is also the dominant species in a given composition. This is known with "selection effect" (Tilman et al., 1997). Therefore, the yield of a mixture can be larger than the average of the two monocultures simply due to the selection effect.

Many measures to estimate overyielding take into account the actual abundance of each species in the assemblages, so that we can better estimate the strength of selection effects and niche complementary (Tilman et al., 1997). This latter refers to the fact that species differing in their ecological properties, can make better use of available resources due to resources partitioning or positive interactions. To cite a few examples, the method proposed by Loreau and Hector (2001), and the method proposed by Loreau (1998).

Hypothesis 8: Taking abundances into account
With the approach presented here, it is straightforward to account for relative abundance in the calculations. Although sorting was carried out in this experiment, for the sake of brevity we will make up differing cover percentages for the species. Let's assume that we want to test whether the mixture `AlAn` is more productive than the two corresponding monocultures. The species `Al` covers on average 35% of the plot surface among all five replicates. We can thus fit the following contrasts.

```
pht.8 <- glht(mod.simple,
              linfct = mcp(sp.comp = c(-0.35, -0.65, 0,0,0, 1, rep(0, times = 10))))
summary(pht.8)


 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: lmer(formula = bio.m2 ~ time.fac + sp.comp + (1 | plot), data = d.zh)

Linear Hypotheses:
       Estimate Std. Error z value Pr(>|z|)
1 == 0    156.7       30.6    5.12  3.1e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

We can thus safely conclude that the two species put together produce more biomass than the weighted average of the two monocultures. This implies that the greater

production is not solely explained by selection effect, but also by niche complementarity.

3.2 Analysing population-level data: A note of caution (***)

If the biomass of each plot is sorted among species (sorting), the following analysis would naturally work at population-level. However, there may be some important statistical caveats when analysing population-level data.

Let's assume that we want to model the Zurich data set at population level. Obviously, the new model must include a predictor species identity (i.e. `sp.id`). It is important to note that every population in each plot is measured repeatedly over time. We must therefore take into account the fact that all eight measurements of, say population `Al` in plot 6, cannot be considered independent. We could therefore include an additional random effect as follows. Note that the response variable is now `bio.pop.m2`, the biomass of each population.

```
mod.sorting <- lmer(bio.pop.m2 ~ time.fac + sp.comp + sp.id +
                      (1 | plot) + (1 | plot:species),
                    data = d.zh.pop.level)
```

In the previously fitted `mod.simple` the random effect `(1 | plot)` takes into account the fact that each plot is measured several times. This random effect can be interpreted as the quality of that plot. Indeed, if a plot is in a favourite position (e.g. more nutrients than other plots), this effect will be observed in all eight measurements as this plot has consistently higher yields. The fact that a random effect plot is present implies that all the observations of one particular plot are positively correlated. This is perfectly sensible. This correlation can be estimated with the intraclass correlation. With `mod.simple` this is estimated to close to zero as the variance of the random effect plot is substantially smaller than the residual variance.

However, when data are sorted to population level, it can be problematic to consider all observations from one particular plot to be positively correlated. Indeed, the actual correlation among observations of one plot is the sum of two processes. On one side, the plot quality introduces a positive correlation. On the other hand, populations may be competing for space within each plot. To make an example, let's look at those plots that have all five species. Assume that in these plots species `Al` covers on average 10% of the surface. If in one of these plots, say plot number 9, this species is more abundant than its average (e.g. covers 30%), this implies that other species will cover less than their average. This implies that observations of the same plot, but different species, are negatively correlated. This can be inspected visually by looking at the residuals. This type of data is often referred as compositional data. The strength of these two processes determines the real intra-plot correlation structure.

Modelling this kind of correlation structure is somewhat feasible with the add-on package *nlme* (i.e. the predecessor of *lme4*; Pinheiro et al. (2016). In practice, the user should specify a "compound symmetry" correlation structure[5].

[5]This is not implemented in the package used here (i.e. `lme4`). However, the *flexLambda* branch of the lme4 project allows that (Bates et al., 2014).

Finally, it has to be noted that this model contains the fixed effect species identity. This means that the interpretation of the species compositions effects is to be interpreted as conditional. In other words, we look at the effect of the composition, for example, `AlAn` when both species effects have been removed. Put simply, the species composition estimates change their meaning. If our aim is to compare the productivity of different species composition, this model parametrisation is inconvenient. We thus discourage the use of the approach presented in this publication with population-level data.

4 Using post-hoc contrasts with complex models (**)

So far we have been fitting post-hoc contrasts on a relatively simple model (i.e. `mod.simple`). In reality, we may want to fit contrasts with more complex models. Up to this point, we assumed that the effect of harvest was the same for all species compositions. We may want to allow each composition to react differently from harvest by including a interaction in the model.

```
mod.complex <- update(mod.simple, . ~ . + sp.comp:time.fac)
anova(mod.simple, mod.complex)
```

```
Data: d.zh
Models:
mod.simple: bio.m2 ~ time.fac + sp.comp + (1 | plot)
mod.complex: bio.m2 ~ time.fac + sp.comp + (1 | plot) + time.fac:sp.comp
             Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
mod.simple   25 8304 8416  -4127     8254
mod.complex 130 8108 8688  -3924     7848   406    105     <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is thus very strong evidence that species are affected differently by harvests. This is in agreement with Figure 1 on page 3. The model fitted here is closer to reality, but also more complex to interpret. Answering an apparently simple question such as "which one is the most productive species?" becomes more involved. The answer may in fact be different at each time point.

Hypothesis 9: Fitting contrasts with complex models (**)
Post-hoc contrasts are very useful with this kind of model too. In order to illustrate their power, we give a simple example. Post-hoc tests compare groups of levels of a given factor. In other words, they can only be used to compare within a given categorical variable. In order to be able to use this technique with a model with an interaction, the model must be re-parametrised. In particular, we must create a new predictor which is the interaction between time of harvest and species composition. We thus re-parametrise `mod.complex` to obtain a fully equivalent model.

```
## 1. create the new factor
d.zh$sp.comp.time.fac <- interaction(d.zh$sp.comp, d.zh$time.fac)
nlevels(d.zh$sp.comp.time.fac)
```

```
[1] 128
```

```
head(levels(d.zh$sp.comp.time.fac))
```

```
[1] "Al.'04 Aug"   "An.'04 Aug"   "Ar.'04 Aug"   "F.'04 Aug"
[5] "H.'04 Aug"    "AlAn.'04 Aug"
```

```
##
## 2. fit the equivalent model
mod.complex.bis <- lmer(bio.m2 ~ sp.comp.time.fac + (1 | plot), data = d.zh)
```

This model has a parameter for each species composition by time point combination. Thus, we can test very specific hypotheses, such as "is overyielding present at all time points?". The newly created factor has thus $16 * 8 = 128$ levels. It it obvious that in this situation writing contrasts by hand may become error-prone. Fortunately, we can use functions such as `match()` or `grep()`. Suppose we are interested in testing whether the full-species mixture is more productive than the average of any species composition containing the most productive species `H`. We could test this hypothesis as follows. For the sake of brevity we are printing the value of all vectors created here below. The provided **R**-code can be used to inspect this objects.

```
## 1. create an indicator vector for "Mix"
Ind.Mix <- grep(pattern = "Mix", x =  levels(d.zh$sp.comp.time.fac))
length(Ind.Mix)
```

```
[1] 8
```

```
## "Mix" was measured 8 times.
##
## 2. create an indicator vector for "H"
Ind.H <- grep(pattern = "H", x =  levels(d.zh$sp.comp.time.fac))
length(Ind.H)
```

```
[1] 40
```

```
## "H" is present as monoculture and in four two-species comp,
## each is measured 8 times. In total is present 40 times.
##
## 3. create an empty vector of the right lenght (i.e. 8*16)
vec.9 <- rep(0, times = nlevels(d.zh$sp.comp.time.fac))
##
## 4. put the coefficients at the right palces and divided them ensure 1/-1 sum
vec.9[Ind.H] <- -1/40
vec.9[Ind.Mix] <- 1/8
```

At this point, we can fit a contrast as usual with the re-parametrised model. Note that we are doing one-sided testing.

```
pht.9 <- glht(mod.complex.bis, linfct = mcp(sp.comp.time.fac = vec.9),
              alternative = "greater")
summary(pht.9)
```

```

 Simultaneous Tests for General Linear Hypotheses
```

```
Multiple Comparisons of Means: User-defined Contrasts


Fit: lmer(formula = bio.m2 ~ sp.comp.time.fac + (1 | plot), data = d.zh)

Linear Hypotheses:
       Estimate Std. Error z value Pr(>z)
1 <= 0     12.2       27.0    0.45   0.33
(Adjusted p values reported -- single-step method)
```

The full-species mixture is more productive than the average of `H, HAl, HAn, HAr, HF`, but this is not statistically significant. It is important to note that by using "regular expressions" (i.e. the function `grep()`) we were able to create a complex contrast vector very quickly and with very low probability of making errors.

Hypothesis 10: Estimating growth with contrasts (**)
Another type of very useful contrast that can be fitted with this model is one comparing the growth of different species compositions. We note that the first yield was the lowest for almost all species composition. We may thus be interested in estimating how much species `H` and `F` increase their yields between the first harvest (i.e. August 2004) and the second one (i.e. June 2005). Finally, we may want to test whether this increase is statistically different in the two groups. Again we make use of regular expressions to facilitate the creation of contrast vectors.

We visualise this hypothesis with an appropriate graph that shows only the two species (panels) and the two time points of interest.

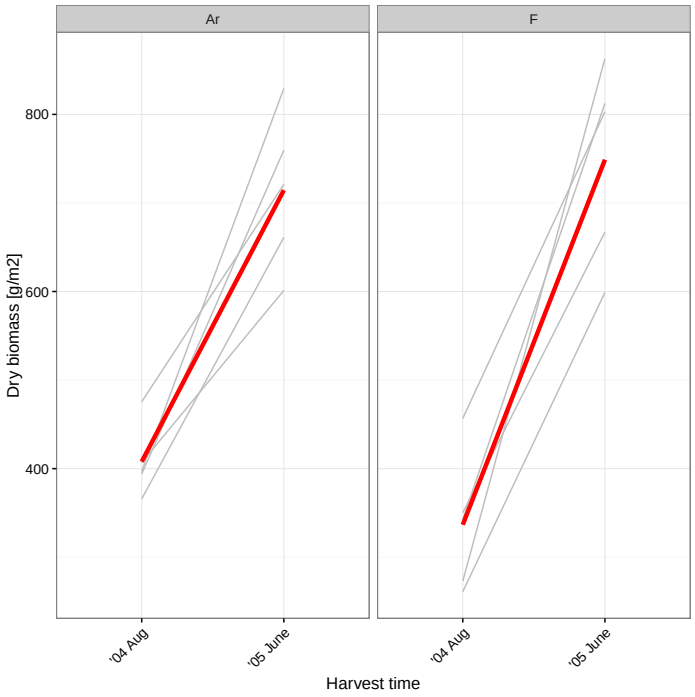


Figure 5: Biomass of Ar and H at the first two time points. Gray thin lines represent plots, while red thicker lines represent average group within each species composition (i.e. panel).

Figure 5 indicates that species `F` grows faster between the first and second harvest. We formally test that.

```
## 1. create vector for "F" (growth between time point 1 and 2)
vec.10.F <- match(table = "F.'05 June", x = levels(d.zh$sp.comp.time.fac),
                  nomatch = 0) -
  match(table = "F.'04 Aug", x = levels(d.zh$sp.comp.time.fac), nomatch = 0)
##
## 2. same for "Ar"
vec.10.Ar <- match(table = "Ar.'05 June", x = levels(d.zh$sp.comp.time.fac),
                   nomatch = 0) -
  match(table = "Ar.'04 Aug", x = levels(d.zh$sp.comp.time.fac), nomatch = 0)
##
## 3. create vector of their difference (growth in "F" - growth in "Ar")
vec.10.diff <- vec.10.F - vec.10.Ar
##
## 4. create the matrix of contrasts
M.10 <- rbind("F growth 1-2"  = vec.10.F,
              "Ar growth 1-2" = vec.10.Ar,
              "diff growth"   = vec.10.diff)
```

```
pht.10 <- glht(mod.complex.bis, linfct = mcp(sp.comp.time.fac = M.10))
summary(pht.10)


 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: lmer(formula = bio.m2 ~ sp.comp.time.fac + (1 | plot), data = d.zh)

Linear Hypotheses:
                   Estimate Std. Error z value Pr(>|z|)
F growth 1-2 == 0     412.2       76.3    5.41  < 1e-04 ***
Ar growth 1-2 == 0    306.6       76.3    4.02  0.00017 ***
diff growth == 0      105.6      107.8    0.98  0.57562
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Both monocultures increase their yields between the first and the second harvest. The increase of monoculture `F` is significantly larger than the one observed for `Ar`.

5 Testing the effect of species richness (***)

Although species richness is not explicitly included in the two models used to test hypotheses (`mod.simple` and `mod.complex.bis`), this predictor is implicitly present as species composition. Indeed, the factor species richness is fully nested in species composition. In other words, the levels of the factor species richness (i.e. 1, 2 and 5)

are simply groups of levels of species composition.

Because of that, it is possible to test for the effect of species richness event though this factor is not implicitly included in the model. It is important to note that species richness cannot be included in the model containing species composition as a fixed effect, without making the model overparametrised[6].The consequence is that the regression coefficients are meaningless. This model cannot be used to compare groups.

If we were truly interested in testing the effect on species richness rather than comparing species composition group, we should fit a model where species richness is present as a fixed effect, while species composition is present as a random effect (see appendix B). Nevertheless, to show the great flexibility of the approach introduced with this publication, we inspect the effect of species richness with a model that does not contain it explicitly.

There are two ways to do this. The first method, which we recommend, uses modern graphical tools. Essentially, we plot the estimated mean values of every species compositions (i.e. 16) against species richness. To get these estimates we make predictions from the fitted model. The second method uses post-hoc contrasts to make formal statistical inference. Note that we use the simple model here, and we use the *lattice* package to visualise data (Sarkar, 2016).

```
## 1. create a new data frame to make predictions
d.fit <- data.frame(sp.comp = levels(d.zh$sp.comp),
                    time.fac = levels(d.zh$time.fac)[1])
##
## 2. make predictions (with no random effects)
fit.sp.rich <- predict(mod.simple, newdata = d.fit, re.form = ~ 0)
##
## 3. plot the obtained predictions against species richness
require(lattice)
d.fit$sp.rich <- c(rep(1, times = 5), rep(2, times = 10), 5)
xyplot(fit.sp.rich ~ sp.rich, data = d.fit, type = c("p", "a", "g"),
       scales = list(y = list(cex = 1.2),
                     x = list(cex = 1.5)),
       ylab = list(label = "Dry biomass [g/m2]", cex = 1.5),
       xlab = list(label = "Species richness", cex = 1.5))
```

[6]This term means that the model contains more parameters than it can be estimated from the data. Appendix B covers this issue in depth.

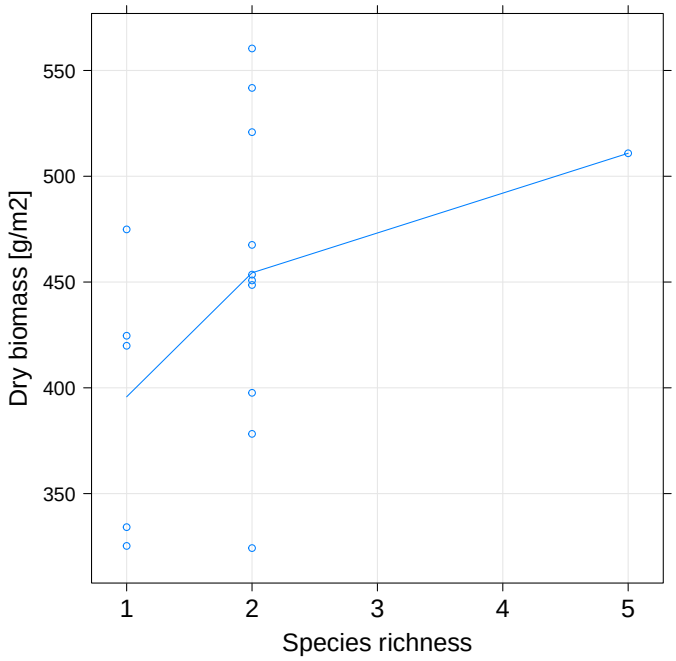


Figure 6: Fitted values for all 16 species compositions plotted against species richness.

This graph clearly indicates that species richness has a positive effect on productivity. In addition, this effect appears to be non-linear. Note that each observation on this graph represents a particular species composition.

Hypothesis 11: Testing species richness (***)
In addition to the graphical evidence, we may want to test the effect of species richness formally. Note that we are using one-sided tests.

```
M.11 <- rbind("SR.1 < SR.2" = c(rep(-1/5, times = 5), rep(1/10, times = 10), 0),
              "SR.2 < SR.3" = c(rep(0, times = 5), rep(-1/10, times = 10), 1))
pht.11 <- glht(mod.simple, linfct = mcp(sp.comp = M.11), alternative = "greater")
summary(pht.11)


 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: lmer(formula = bio.m2 ~ time.fac + sp.comp + (1 | plot), data = d.zh)

Linear Hypotheses:
                  Estimate Std. Error z value  Pr(>z)
SR.1 < SR.2 <= 0     58.6       13.5    4.34 1.4e-05 ***
SR.2 < SR.3 <= 0     56.5       25.8    2.19   0.029 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Not surprisingly, there is strong evidence that species richness positively influences productivity. Note that here we performed two tests as there are three levels of species richness. In this case both test were significant. However, this way of proceeding may be inadequate in other real situations where several levels of species richness are present.

6 Using post-hoc contrasts on any type of data

Possibly one of the greatest advantages of the approach presented here for diversity-function studies is its tremendous flexibility. Indeed, here we can analyse essentially any kind of data very easily. The routines to perform post-hoc contrasts with data other than normal (Gaussian) are readily implemented and work in the same way.

As an example, let's assume that in a forest diversity-function study we were to analyse the number of trees that can be harvested in each plot ten years after planting. Or alternatively, we may want to model the "invasibility" of plots in a grass study, and thus model the number of invading species in each plot. Both these response variables are likely to be analysed appropriately with a Generalised Linear (Mixed) Model with Poisson family. Other kinds of data may best analysed using the binomial family. Fitting post-hoc contrasts with a Generalised Linear (Mixed) Model works in the same way, so there are no additional steps to perform. Therefore, very different ecosystem functions can be analysed in the same framework.

The function to fit post-hoc contrasts used here, `glht()` can also be used with survival models. This could also be very beneficial in this context. Imagine that we recorded each tree for survival. However, trees were not recorded at discrete and shared time points. Possibly the best way to analyse this data is to use a survival model. This introduces another important extension that we can use with these models, censoring[7]. So, we may have censored observations and include this information into our model and use it when fitting post-hoc contrasts.

Analysing response variables of different types is fundamental to avoid getting a one-dimensional picture of ecosystem functioning. The methods used to analyse our data must not hinder our comprehension of the whole picture simply because they are not flexible enough.

7 How to profit from the classical statistical framework

One of the real advantages of the approach presented in this publication is the fact that we can profit from using any tool implemented in Linear Models and their extensions. To make a relevant example here, we may want to account for the different surface area that plots have. Indeed, for logistical reasons the surface harvested at the last time points were smaller than the previous ones (resp. 0.09 m^2 and 0.25 m^2). It follows that

[7]For example, a tree may not be surveyed any more or "leave" the study as an elephant stood on it (Hector et al., 2011).

these observations are less accurate[8]. This information can simply by included in the model through the use of weights[9]. For simplicity we add weights to the model that does not contain the interaction between species composition and harvest time.

```
mod.simple.weights <- update(mod.simple, weights = surface)
```

We may want to compare the estimates obtained when the harvest surface is taken into account (not shown). Some estimates do change considerably. We now compute a contrast that was previously fitted without weights and compare the results. In particular, we test again for transgressive overyielding.

```
pht.3.weights <- glht(mod.simple.weights,
                      linfct = mcp(sp.comp = c(0,0,0,0, -1, 0,0,0,0,0,0,0,0,0,0, 1)),
                      alternative = "greater")
##
summary(pht.3) ## transgressive overyielding no weights

 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: lmer(formula = bio.m2 ~ time.fac + sp.comp + (1 | plot), data = d.zh)

Linear Hypotheses:
       Estimate Std. Error z value Pr(>z)
1 <= 0     36.0       34.8    1.03   0.15
(Adjusted p values reported -- single-step method)
summary(pht.3.weights) ## transgressive overyielding with weights

 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: lmer(formula = bio.m2 ~ time.fac + sp.comp + (1 | plot), data = d.zh,
    weights = surface)

Linear Hypotheses:
       Estimate Std. Error z value Pr(>z)
1 <= 0     65.3       31.7    2.06   0.02 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Interestingly, by taking the harvest surface into account, we can show that transgressive overyielding is present. This is not surprising as the model with weights is more

[8]This is confirmed graphically by comparing boxplots of the absolute residuals of the two groups of surface area.

[9]Here we assume that the precision of the observations is proportional to the harvested surface.

powerful as it accounts for the "quality" of the observations.

8 Final remarks

Transparency and reproducibility
This dynamic document was produced with the add-on package *knitr* (Xie, 2016). The entire **R**-code of this document is attached. To enable full reproducibility of these commands, we used the package *checkpoint* (Corporation, 2016). All information regarding package versions is on page 29.

Other software or packages
The approach presented in this publication does not rely on any particular software or package. Nevertheless, we do believe that the choices made here (i.e. **R**, *lme4* and *multcomp*) are particularly advantageous.

Completeness of the examples
For the sake of brevity the examples used here are not shown in their complete form. An analysis using the approach presented here should perform the following steps:

1. display the data
2. fit the model
3. check whether the model assumptions are fulfilled (e.g. normality of the errors)[10]
4. test whether the predictor/s of interest (here `sp.comp`) is/are significant
5. display all the hypotheses to be tested with an appropriate graph.
6. test all the hypotheses regarding one particular predictor in one single call of the `glht()` function[11]
7. report the results appropriately[12]

Comparison with the results of Vojtech et al
The data analysed here comes from a published study (Vojtech et al., 2008). Although there are some slight differences in the data analysed and questions addressed, we can safely claim that the results shown here highlight that the approach presented in this publication is statistically more powerful. Put simply, this approach is more likely to find effects if there are any.

The original authors could not highlight a clear effect of overyielding. In this analysis the greater productivity of the full-species mixture over monocultures was clear and,

[10] In this case, the assumptions of the complex model were fulfilled. However, for the simple model, there was obviously some structure left as we (wrongly) assumed an additive effect of species composition and time.

[11] This is essential to make sure that the multiple testing adjustment works properly.

[12] It should be made clear what hypotheses were defined before seeing the data (a-priori hypotheses) and which ones after that (a-posteriori hypotheses).

in addition to that, we have even be able to show that transgressive overyielding is present. The reason for the superiority of this approach over classical metrics such as Loreau's $\overline{D}$ is that it is a conditional approach, while classical metrics use a marginal approach. In other words, before measuring, say, overyielding, the effect of all other predictors is removed. In this study, we simply controlled for the effect of harvest time and plot. If other predictors were available, for example soil moisture or soil fertility, we could have estimated overyielding even better (i.e. more precisely). Classical metrics are said to be marginal as they ignore all this relevant information. The conditional approach was formally introduced by Connolly and Bell (Bell et al., 2009; Connolly et al., 2013).

Advantages of the approach presented here

The advantages of the approach presented here are fully discussed in the main publication text. Below we list some of the points that have been highlighted here.

- the method is very flexible in terms of hypotheses tested
- the method is very versatile in terms of data analysed
- the last two points (flexibility and versatility) imply that this method has a strong unifying property
- the method is faster and less error-prone than classical metrics
- the results are in the same scale as the response variable analysed
- the graphical approach makes it very easy to understand and interpret the results
- the method is not-field specific, but a mainstream statistical technique. This implies:
  - its theory is solid
  - the path to statistical inference (p-value and confidence intervals) is clear
  - several modern implementations exist
  - it is easy to communicate results among research fields
  - it is easy to get advice and consultancy
  - it is easy to find literature[13]

[13]For example, the package *multcomp* is accompanied by several documents and a book (Hothorn et al., 2016; Bretz et al., 2016).

Reproducibility

Checkpoint date was set to 2016-05-01.

```
sessionInfo()

R version 3.3.2 (2016-10-31)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Debian GNU/Linux 8 (jessie)

locale:
 [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
 [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
 [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] lattice_0.20-34 multcomp_1.4-6  TH.data_1.0-7   MASS_7.3-45
 [5] survival_2.40-1 mvtnorm_1.0-5   lme4_1.1-12     Matrix_1.2-7.1
 [9] ggplot2_2.1.0   knitr_1.14

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.7      magrittr_1.5     splines_3.3.2    munsell_0.4.3
 [5] colorspace_1.2-7 minqa_1.2.4      stringr_1.1.0    highr_0.6
 [9] plyr_1.8.4       tools_3.3.2      grid_3.3.2       gtable_0.2.0
[13] nlme_3.1-128     digest_0.6.10    reshape2_1.4.1   nloptr_1.0.4
[17] formatR_1.4      codetools_0.2-15 evaluate_0.10    labeling_0.3
[21] sandwich_2.3-4   stringi_1.1.2    scales_0.4.0     zoo_1.7-13
```

References

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting Linear Mixed-Effects Models using lme4. *arXiv preprint arXiv:1406.5823*, 2014.

D. Bates, M. Maechler, B. Bolker, and S. Walker. *lme4: Linear Mixed-Effects Models using 'Eigen' and S4*, 2016. URL `https://CRAN.R-project.org/package=lme4`. R package version 1.1-12.

T. Bell, A. K. Lilley, A. Hector, B. Schmid, L. King, and J. A. Newman. A linear model method for biodiversity-ecosystem functioning experiments. *The American Naturalist*, 174 (6):836–849, 2009.

F. Bretz, T. Hothorn, and P. Westfall. *Multiple comparisons using R*. CRC Press, 2016.

J. Connolly, T. Bell, T. Bolger, C. Brophy, T. Carnus, J. A. Finn, L. Kirwan, F. Isbell, J. Levine, A. Lüscher, et al. An improved model to predict the effects of changing biodiversity levels on ecosystem function. *Journal of Ecology*, 101(2):344–355, 2013.

M. Corporation. *checkpoint: Install Packages from Snapshots on the Checkpoint Server for Reproducibility*, 2016. URL `https://CRAN.R-project.org/package=checkpoint`. R package version 0.3.16.

A. Hector, C. Philipson, P. Saner, J. Chamagne, D. Dzulkifli, M. O'Brien, J. L. Snaddon, P. Ulok, M. Weilenmann, G. Reynolds, et al. The Sabah Biodiversity Experiment: A long-term test of the role of tree diversity in restoring tropical forest structure and functioning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1582):3303–3315, 2011.

T. Hothorn, F. Bretz, and P. Westfall. *multcomp: Simultaneous Inference in General Parametric Models*, 2016. URL `https://CRAN.R-project.org/package=multcomp`. R package version 1.4-6.

M. Loreau. Separating sampling and other effects in biodiversity experiments. *Oikos*, 82(3): 600–602, 1998.

M. Loreau and A. Hector. Partitioning selection and complementarity in biodiversity experiments. *Nature*, 412(6842):72–76, 2001.

J. Pinheiro, D. Bates, and R-core. *nlme: Linear and Nonlinear Mixed Effects Models*, 2016. URL `https://CRAN.R-project.org/package=nlme`. R package version 3.1-128.

D. Sarkar. *lattice: Trellis Graphics for R*, 2016. URL `https://CRAN.R-project.org/package=lattice`. R package version 0.20-34.

D. Tilman, C. L. Lehman, and K. T. Thomson. Plant diversity and ecosystem productivity: Theoretical considerations. *Proceedings of the National Academy of Sciences*, 94(5):1857–1861, 1997.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-PLUS*. Springer Science & Business Media, 2013.

E. Vojtech, M. Loreau, S. Yachi, E. M. Spehn, and A. Hector. Light partitioning in experimental grass communities. *Oikos*, 117(9):1351–1361, 2008.

R. L. Wasserstein and N. A. Lazar. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 2016.

H. Wickham and W. Chang. *ggplot2: An Implementation of the Grammar of Graphics*, 2016. URL `https://CRAN.R-project.org/package=ggplot2`. R package version 2.1.0.

Y. Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2016. URL `https://CRAN.R-project.org/package=knitr`. R package version 1.14.

Appendix B: Model parametrisation

Chapter 2

Contents

1 Introduction

This appendix discusses the different model parametrisations that can potentially be used when analysing biodiversity function experiments. One peculiarity of these analyses is the omnipresence of two fundamental variables, species richness and species composition. These two predictors are tightly interconnected biologically and, as we will see, also from the modelling point of view.

Biodiversity experiments, as the name implies, aim to understand the effects of diversity on ecosystem functioning. When analysing these experiments, is therefore crucial to be able to interpret biologically the results obtained. For this reason, we must fit models whose estimated coefficients can unambiguously be interpreted. As we will see, this may turn out to be harder than commonly thought. Nevertheless, solutions exist and biodiversity function experiments can benefit from using them.

This document is divided in five main parts. After this brief introduction, in Section 2 we discuss the concepts of model contrasts and model identifiability. In other words, we discuss what makes a model overdetermined and what its consequences are. Subsequently, in Section 3, we discuss various methods to model the potential non-linear effect of biodiversity. Then, in Section 4 we discuss the four model parametrisations that involve taking the two biodiversity predictors as fixed effects, as random effects or a mixture of the two. Finally, in Section 5 we make some general remarks about this document.

2 Model contrasts and model identifiability

Before discussing the different modelling options available for including the two biodiversity predictors (i.e. species richness and species composition) in a model, we discuss some important aspects of model contrasts. These are not to be confused with post-hoc contrasts that we used in Appendix A. Model contrasts refer to the parametrisation used to fit a Linear Model or any of its extensions (Venables and Ripley, 2013). Two commonly encountered model contrasts are the treatment contrasts and the sum contrasts. The software used here (i.e. **R**) uses treatment contrasts by default. The interpretation of model coefficients depends on the contrasts used and whether the model is identifiable or not. To better understand what this means and implies we introduce an example with real data.

2.1 Coefficients Interpretation for Linear Models

We start this subsection by introducing the data set used. The data comes from a forest experiment carried out in the Czech Republic. As any biodiversity function experiment, it has a fairly complex design[1]. Here we focus on few relevant variables and a subset of the whole data.

[1]More details about this data set can be found in Chamagne et al. (2016). However, the information provided here is sufficient to understand the examples.

The variable that we want to model is the yearly radial growth measured as the width of tree rings (i.e. `ring`). We start by looking at the effect of species identity (i.e. `sp`) on growth.

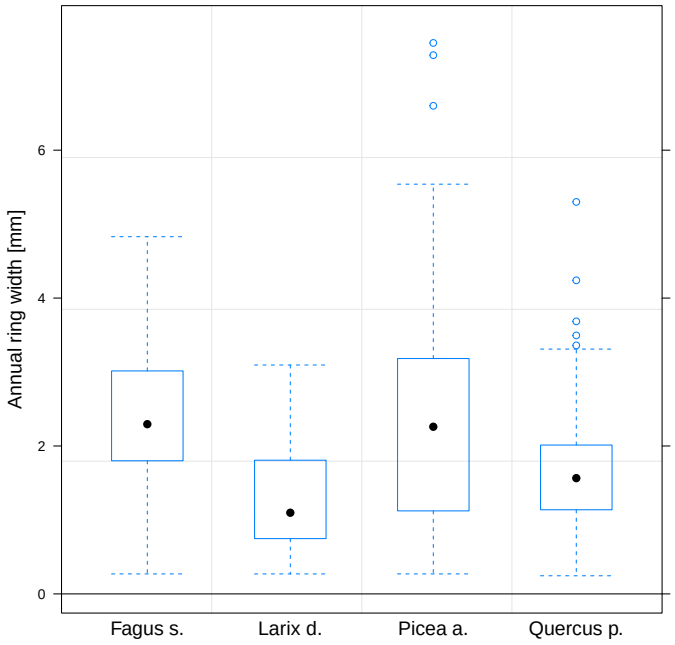


Figure 1: Response variable plotted against species identity.

This graph indicates that species differ in their growth rate, with European larch (i.e. *Larix d.*) being the slowest grower. We now want to model this data within the framework of Linear Models. Ideally, we may want to know how much, on average, rings of each species grow.

```
lm.1 <- lm(ring ~ sp, data = d.ring)
coef(lm.1)
```

```
 (Intercept)   spLarix d.   spPicea a. spQuercus p.
       2.414       -1.147        0.024       -0.754
```

Because treatment contrasts are used for categorical variables, the interpretation is as follows. The first species in alphabetical order, *Fagus sylvatica* (i.e. European beech), is set as a reference. This means that what we read as `(Intercept)` in the model output actually refers to the mean yearly ring increase for European beech[2]. All the parameters for the other species are the difference from the reference level. As an example, the coefficient for European is -1.1. This implies that on average European larch grows -1.1 mm less than European beech (i.e. European larch grows 1.3 mm/year). This parametrisation is inconvenient if we are interested in the group means. By removing the intercept in this model, we can obtain estimates that are easier to interpret (the intercept is represented by `1` in **R** formulas).

[2]If we were to use sum contrasts, then the intercept would refer to the mean value of all four groups of species.

```
lm.2 <- lm(ring ~ sp - 1, data = d.ring)
coef(lm.2)
```

```
  spFagus s.   spLarix d.   spPicea a. spQuercus p.
         2.4          1.3          2.4          1.7
```

By removing the intercept, we obtain a fully equivalent model, where the four estimated parameters are actually the group means. Although for this simple model, the new parametrisation makes it easier to understand what the group averages are, it has some disadvantages. As an example, we may want to test formally if species differ. This implies comparing the current model with one that only contains a grand mean (i.e. no differences among species).

```
lm.3 <- lm(ring ~ 1, data = d.ring)
coef(lm.3)
```

```
(Intercept)
        1.9
```

To then test the hypothesis that species do not differ, we can run an ANOVA. Is important to note that, the two models need to be nested[3] if we want to use this test. Here, the "groups means" model (i.e. `lm.2`) and the "grand mean" model (i.e. `lm.3`) are not nested and cannot be formally compared.

We may think that a way out of this situation would be to fit a model with an intercept and the four species means. In that case, the "grand mean" model would be nested in this latter model. However, this model cannot be fitted. Indeed, we would be estimating five parameters (i.e. an intercept and four group means), from four groups. There is no unique solution to get the four group means based on five parameters. Indeed, we could set the intercept parameter to zero and the four species parameters to their actual mean value, or we could set the intercept to 10 and then the species parameters to their actual value -10. When there is no unique solution to estimate the parameters of a model, as for example `lm.2`, this latter is said to be "overdetermined" (or "non-identifiable") and its design matrix to be "rank-deficient". These concepts will be further discussed at page 7.

On the other hand, the first model that we fitted (i.e. `lm.1`) and the "grand mean" model (i.e. `lm.3`) are nested and can be compared via an ANOVA test.

```
anova(lm.1, lm.3)
```

```
Analysis of Variance Table

Model 1: ring ~ sp
Model 2: ring ~ 1
  Res.Df RSS Df Sum of Sq    F  Pr(>F)
1    389 411
2    392 499 -3     -87.8 27.7 3.1e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[3] Formally speaking, "nested" means that the design matrix of one model is a subset of the columns of the other.

As the output of the `anova` function clearly indicates, there are significant differences among species. Nevertheless, this does not tell us which species are different from the others. To answer this kind of questions we can use post-hoc contrasts seen in Appendix A. Here we showed that species differ, now we may want to test specific hypotheses. As an example, we may want to test which species differ from each other in a pairwise manner. To do that, we can use the `TukeyHSD()` function[4].

```
aov.1 <- aov(lm.1)
TukeyHSD(aov.1)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = lm.1)

$sp
                       diff    lwr   upr p adj
Larix d.-Fagus s.    -1.147 -1.551 -0.74  0.00
Picea a.-Fagus s.     0.024 -0.381  0.43  1.00
Quercus p.-Fagus s. -0.754 -1.108 -0.40  0.00
Picea a.-Larix d.     1.171  0.751  1.59  0.00
Quercus p.-Larix d.   0.393  0.022  0.77  0.03
Quercus p.-Picea a. -0.778 -1.151 -0.40  0.00
```

This output tells us clearly that, to make an example, European larch and European beech do differ in their growth rates. Testing pairwise relationships ii just one option of post-hoc contrasts.

2.2 Model identifiability

We have seen that species differ in their growth rate. In particular, we saw that the yearly ring growth mean value for each species are given by the coefficients of `lm.2`.

```
coef(lm.2)
```

```
  spFagus s.   spLarix d.   spPicea a. spQuercus p.
        2.41         1.27         2.44         1.66
```

We may want to know what the mean growth for conifers and broad-leaved species is. To do that, we can simply take the averages of the estimated coefficients.

```
## 1. broad-leaved species
( mean.broad <- mean(coef(lm.2)[c("spFagus s.", "spQuercus p.")]) )
```

```
[1] 2.04
```

```
##
## 2. conifers
( mean.conif <- mean(coef(lm.2)[c("spLarix d.", "spPicea a.")]) )
```

```
[1] 1.85
```

[4]This function must be fed with an object of `aov` class (`aov` stands for Analysis Of Variance).

On average, the broad-leaved trees grow by about 2.04 mm per year, while the conifers grow by about 1.85 mm per year. Proceeding this way, manually taking averages of estimated coefficients, may seem untidy and hard to apply to complex models. We may want to get these estimates from the model itself. To that, we could naively introduce a new factor with two levels in our model. Because treatment contrasts are used, we would expect the parameter associated with this factor to tell us the difference between the two "families" (i.e. broad-leaved and conifers).

```
levels(d.ring$family)
```

```
[1] "broad-leaved" "conifer"
```

```
lm.4 <- lm(ring ~ sp + family, data = d.ring)
coef(lm.4)
```

```
  (Intercept)    spLarix d.    spPicea a.  spQuercus p. familyconifer
        2.414        -1.147         0.024        -0.754            NA
```

When looking at the estimated coefficients we see that one of them is set to `NA`. The reason for that is the this model is not identifiable. Indeed, we have only four groups and we are attempting to estimate five parameters. The default behaviour of **R** is to set a necessary number of parameters to zero to make the model identifiable[5]. In this case we have four groups, thus, we can estimate at most four parameters. As we attempted to estimate five, we just need to constrain one parameter to be zero to make the model identifiable again. As shown in Appendix A, the difference between broad-leaved species and conifers is best estimated by using post-hoc contrasts. This example was used to introduce the problem of fitting non-identifiable models.

Further down in this document, we will deal with Linear Mixed Effects Models. In this publication we use the function `lmer()` from package *lme4* to fit these models (Bates et al., 2016). We thus briefly look at the behaviour of this function when non-identifiable models are fitted. Here, we add the random effect `site` to our model.

```
require(lme4)
```

```
mod.4 <- lmer(ring ~ sp + family + (1 | site), data = d.ring)
```

```
fixed-effect model matrix is rank deficient so dropping 1 column / coefficient
```

```
fixef(mod.4)  ## fixef() is homologous to coef() for lm objects
```

```
 (Intercept)   spLarix d.   spPicea a. spQuercus p.
        2.40        -1.16         0.02        -0.75
```

In analogy with the function `lm()` some coefficient are set to zero. `lmer()` simply drops these coefficients instead of reporting them as `NA`.

2.3 Collinearity of continuous predictors

To explain the problem of non-identifiability further, we can look at the problem of collinearity. Sometimes some of the predictors considered for modelling are tightly

[5]Note that the summary output reports the constrained parameters to be `NA` to avoid confusion. These parameters are set to be zero, they are not estimated to be zero.

linked and their values are very close. In these cases, we may not be able to include both of them into a model as they are too collinear[6]. The extreme case of collinearity is when two predictors are carrying exactly the same information.

Continuing with our growth example, we can look at effect of tree size on growth. It is well known that, once age is taken into account, larger trees grow faster. So if we were to include two predictors in our model, say diameter in centimetres and diameter in metres, these two would be perfectly collinear. In this case the model would not be identifiable, and **R** would set one of the two coefficients to zero to be able to carry out the estimation procedure.

If we look at the interpretation of the model coefficients we understand that also from a practical point of view, having diameter measured in centimetres and metres in the same model does not make sense. To make an example if the coefficient for diameter in metres was 2.38, we would interpret this as follows. "When the diameter of the tree increases by one unit (i.e. one meter) and all the other predictors are kept at a constant value, then the response variable increases by 2.38". This clearly does not make sense as we cannot increase diameter in metres and not change diameter measured in centimetres.

When dealing with nested factors we face the same situation. Going back to the previous example. We cannot change the family (e.g. go from broad-leaved to conifer), but keep all the other predictors fixed (e.g. remain European beech). This confirms that the classical interpretation of model coefficients in a non-identifiable model breaks down.

The interpretation of correlated variables is not only problematic when the two predictors are perfectly collinear. We may think of tree diameter and tree height. It is natural to expect these two variables to be highly collinear (i.e. taller trees are also larger). So in reality, it would be very unlikely that we can let one variable vary while keeping the other fixed. This implies that the interpretation of model coefficients has to be carried out with care. From the technical point of view we would notice that the standard errors of these two variables are very large, as the model cannot tell the effects apart.

2.4 Rank of the design matrix (**)

Sometimes a model we would like to fit is not identifiable as its design matrix is not of full rank. A rank deficient matrix is a matrix where one or more columns are a linear combination of the other columns. The simplest case of a rank deficient model matrix is when one column is a multiple of another column. This is exactly what we observed with diameter measured in centimetres and in metres. One column is simply the other multiplied (or divided) by 100.

In practical terms, rank-deficiency means that we are trying to estimate too many parameters (i.e. fitting an overdetermined model). The `lm()` function deals with these situations by constraining some parameters to be zero, which makes the design matrix smaller and of full rank.

[6]Collinearity between continuous predictors is usually estimated with the Variance Inflation Factor (VIF).

To better understand the concept of rank-deficient matrix, we look at the design matrix of a model that contains `sp` and `family` (i.e. `lm.4`). For simplicity, we look at a subset of the observations, where each species is present with two observations. The `model.matrix()` function can be used to create a design matrix.

```
d.ex.4  ## toy data set
```

```
  ring         sp       family
1  2.7   Fagus s. broad-leaved
2  2.2   Fagus s. broad-leaved
3  1.1   Larix d.      conifer
4  1.6   Larix d.      conifer
5  2.2   Picea a.      conifer
6  2.9   Picea a.      conifer
7  1.5 Quercus p. broad-leaved
8  1.8 Quercus p. broad-leaved
```

```
( X.lm.4 <- model.matrix(ring ~ sp + family, data = d.ex.4) )
```

```
  (Intercept) spLarix d. spPicea a. spQuercus p. familyconifer
1           1          0          0             0             0
2           1          0          0             0             0
3           1          1          0             0             1
4           1          1          0             0             1
5           1          0          1             0             1
6           1          0          1             0             1
7           1          0          0             1             0
8           1          0          0             1             0
attr(,"assign")
[1] 0 1 1 1 2
attr(,"contrasts")
attr(,"contrasts")$sp
[1] "contr.treatment"

attr(,"contrasts")$family
[1] "contr.treatment"
```

The function `model.matrix()` does not drop columns even if the model is not identifiable. Indeed, all five columns of the design matrix are printed. It is easy to see that, the last column (i.e. `familyconifer`) is actually the sum of the columns that belong to conifer (i.e. `spLarix d.` and `spPicea a.`).

We can compute the rank of this matrix with the `rankMatrix()` function found in *Matrix* package (Bates and Maechler, 2016).

```
require(Matrix)
as.vector(rankMatrix(X.lm.4))  ## as.vector() is used for a clearer printing
```

```
[1] 4
```

```
##
ncol(X.lm.4)
```

```
[1] 5
```

This matrix has five columns, but only rank four. It is thus rank deficient. Whenever the design matrix is rank deficient, additional constraints must be made in order to carry out the estimation process.

2.5 Summary

We noted that special care is needed when dealing with tightly linked predictors. When the model is not identifiable, the default behaviour of **R** is to add constraints to the model so that the estimation can be carried out. In these cases, the classical interpretation of the model coefficients is completely lost. In most cases the best solution out of this problem is to drop one or more predictors in order to make the model identifiable and to be able to interpret the coefficients obtained.

Post-hoc contrasts can then be used to test specific questions regarding the omitted predictors. As an example, we may use this technique to estimate the difference between conifers and broad-leaved species by using a model that only contains the categorical variable species (see Appendix A).

It is worth mentioning that the problem of non-identifiability is shared by Linear Models and their extensions. In particular Generalised Linear Models and Mixed Effects Models are also affected by this issue. Note that changing model contrasts (**R** defaults to treatment contrasts) does not remove the issue of presented here. A non-identifiable model remains such regardless of the model contrasts used.

3 Testing the non-linearity of species diversity effect

When looking at the effect of biodiversity we may want to test whether this can be considered to be linear or not. This enables us to also further discuss the implications of overdetermined models.

The effect of biodiversity may be saturating (Cardinale et al., 2012). This implies that the supposedly positive effect of diversity may be decreasing at higher values. There are several ways to test whether the effect of a predictor can be assumed to be linear or not. In this section we present a few options and highlight their pros and cons. Always, before starting any modelling procedure, we display the relationship with a meaningful graph. Again, we continue with the growth example.

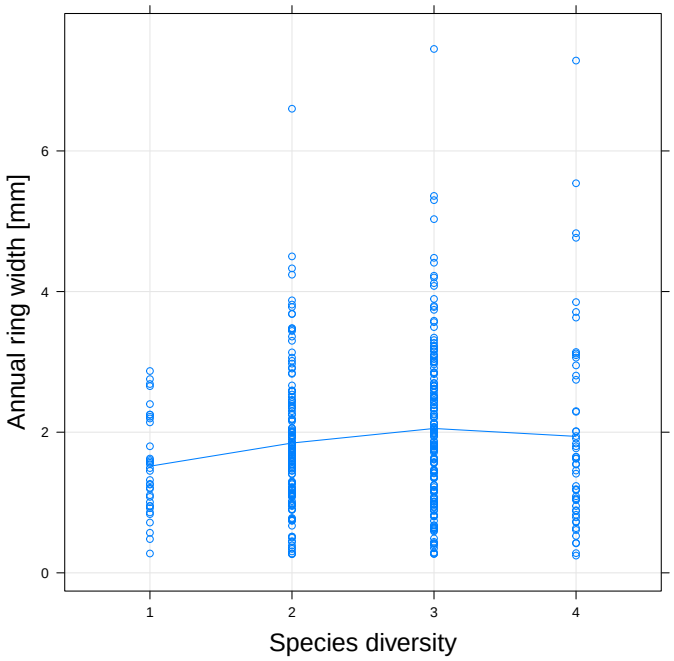


Figure 2: Response variable plotted against species diversity.

This graph indicates that the effect of species diversity may be slightly non-linear. We now look at several options to best model and test this non-linearity.

3.1 Using polynomials to model non-linearity

A simple way to deal with non-linear effects is to use low-degree polynomials. To start with, we may want to use just a quadratic effect. For the sake of brevity, we fit a very simple model where only species richness is used as a predictor.

```
lm.5 <- lm(ring ~ sp.rich + I(sp.rich^2), data = d.ring)
```

We can now test whether the quadratic term is needed by removing it from the model.

```
lm.6 <- update(lm.5, . ~ . - I(sp.rich^2))
anova(lm.6, lm.5)

Analysis of Variance Table

Model 1: ring ~ sp.rich
Model 2: ring ~ sp.rich + I(sp.rich^2)
  Res.Df RSS Df Sum of Sq    F Pr(>F)
1    391 493
2    390 490  1      3.51 2.79  0.096 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is thus little evidence that we need the quadratic term. If we were to use the 5%

level of significance very strictly[7], we would conclude that the quadratic effect is not needed. This does not necessarily imply that the effect of species diversity is linear. Indeed, we may have insufficient statistical power or the effect may be non-linear, but not in a quadratic form. We could allow for greater flexibility by using higher order polynomials. However, these can be hard to interpret biologically and can have technical problems[8]. Ideally, we would like to make no assumptions on the relationship between the response variable and species richness. The intuitive way to do that with a discrete predictor such as `sp.rich` is to use a categorical variable.

3.2 Using categorical variables to model non-linearity

In this example species richness (i.e. `sp.rich`) is a predictor with four discrete values. Therefore, we may want to use the greatest possible flexibility and estimate a parameter for each level of species richness. To do that, we create a new predictor which is a categorical variable modelling species richness.

```
d.ring$sp.rich.fac <- factor(d.ring$sp.rich)
levels(d.ring$sp.rich.fac)
```

```
[1] "1" "2" "3" "4"
```

```
##
lm.7 <- lm(ring ~ sp.rich.fac, data = d.ring)
```

To test whether this additional term in the model is needed we can fit a model where species richness is considered as being a continuous predictor as well as a categorical variable. After we can compare this model with one that only contains a linear effect for species richness.

```
lm.8 <- lm(ring ~ sp.rich + sp.rich.fac, data = d.ring)
anova(lm.6, lm.8)
```

```
Analysis of Variance Table

Model 1: ring ~ sp.rich
Model 2: ring ~ sp.rich + sp.rich.fac
  Res.Df RSS Df Sum of Sq    F Pr(>F)
1    391 493
2    389 490  2      3.74 1.49   0.23
```

There is thus no evidence that the effect of species diversity is non-linear. We may want to remove the the non-linear term and test whether species diversity has an effect at all. It important to note that testing for the effect of `sp.rich` (the continuous predictor) in the model that contains the categorical variable species richness (i.e. `lm.8`) is inappropriate. Indeed, this model is constrained and its design matrix is rank-deficient. To better understand this we can look at the coefficients estimated for species diversity.

[7] We strongly discourage the dichotomous interpretation of p-values as significant or not significant.

[8] Higher order polynomials can be unstable, especially when fitted on discrete data, the function `poly()` creates orthogonal polynomials that solve this problem (Chambers and Hastie, 1991).

```
coef(lm.8)[1:5]
```

```
 (Intercept)      sp.rich sp.rich.fac2 sp.rich.fac3 sp.rich.fac4
        1.38         0.14         0.19         0.25           NA
```

Not unexpectedly, one parameter has to be constrained to make the model identifiable. Indeed, we have four groups (diversity levels) and we tried to model that with five parameters. Note that which parameter is set to zero depends on the order used in the model formula.

```
coef(lm(ring ~ sp.rich.fac + sp.rich, data = d.ring))
```

```
 (Intercept) sp.rich.fac2 sp.rich.fac3 sp.rich.fac4      sp.rich
        1.52         0.33         0.54         0.42           NA
```

If species richness as a continuous variable is declared as the last variable, then this parameter is set to zero. The resulting model is thus simply the model where there is just species richness as a factor (i.e. same as `lm.7`). Note that we are using the conditional sums of squares (i.e. type II SS) for testing, however, this reasoning is identical for the sequential sums of squares (i.e. type I SS).

Model matrix

We may want to look at the design matrix to understand why one parameter is dropped. To do that, we use a very small subset of data again.

```
str(d.ex.8)
```

```
'data.frame': 8 obs. of  3 variables:
 $ ring       : num  2.69 2.22 1.1 1.62 2.25 ...
 $ sp.rich    : int  1 1 2 2 3 3 4 4
 $ sp.rich.fac: Factor w/ 4 levels "1","2","3","4": 1 1 2 2 3 3 4 4
```

```
d.ex.8
```

```
  ring sp.rich sp.rich.fac
1  2.7       1           1
2  2.2       1           1
3  1.1       2           2
4  1.6       2           2
5  2.2       3           3
6  2.9       3           3
7  1.5       4           4
8  1.8       4           4
```

We then look at the design matrix where all columns are kept.

```
model.matrix(ring ~ sp.rich.fac + sp.rich, data = d.ex.8)
```

```
  (Intercept) sp.rich.fac2 sp.rich.fac3 sp.rich.fac4 sp.rich
1           1            0            0            0       1
2           1            0            0            0       1
3           1            1            0            0       2
4           1            1            0            0       2
5           1            0            1            0       3
6           1            0            1            0       3
```

```
7           1           0           0           1       4
8           1           0           0           1       4
attr(,"assign")
[1] 0 1 1 1 2
attr(,"contrasts")
attr(,"contrasts")$sp.rich.fac
[1] "contr.treatment"
```

In this case it trivial to see that we can obtain the last column x_5 (i.e. `sp.rich`) by the following linear combination of the others:

$$x_5 = x_1 + x_2 + 2 * x_3 + 3 * x_4$$

This makes it clear why the matrix is rank deficient and the model not identifiable without additional constraints. it is also clear that the interpretation of this model is problematic, as the effect of `sp.rich` should be interpreted as conditional on the effect the factor `sp.rich.fac`. It is clearly impossible to vary one while keeping the other fixed.

Interpretation of the `sp.rich` regression coefficient

It is important to note that the regression coefficient for `sp.rich` in the constrained model `lm.8` is meaningless. It is not the effect on the response variable when species richness is increased by one. Indeed, due to the constraint applied to the model, the regression line estimated for `sp.rich` is also constrained. In particular, it is forced to go through the mean of the reference group (i.e. `sp.rich = 1`) and the mean of the group constrained to be zero[9] (i.e. `sp.rich = 4`).

We can display this situation graphically. We first plot all the data, we then add the group means and we finally add the regression line obtained from the non-identifiable model (i.e. `lm.8`).

```
## 1. plotting observations with transparency
plot(ring ~ sp.rich, data = d.ring, col = rgb(0.1, 0.1, 0.1, 0.5), cex = 0.5,
     cex.lab = 1.5,
     xlab = "Species Richnes", ylab = "Annual ring width [mm]")
##
## 2. computing groups means and displaying them
fitted.y <- predict(lm.8, newdata = data.frame(sp.rich = 1:4,
                                               sp.rich.fac = factor(1:4)))
points(x = 1:4, y = fitted.y, col = c("red", "blue", "blue", "red"),
       cex = 3, pch = 4)
##
## 3. adding the constrained regression line
abline(a = coef(lm.8)["(Intercept)"], b = coef(lm.8)["sp.rich"],
       col = "red", led = 2)
```

[9] This is the case because treatment contrasts are used. When using other contrasts, other parameters are constrained to be zero. As we will see, changing contrasts does not solve the issue of coefficient interpretation.

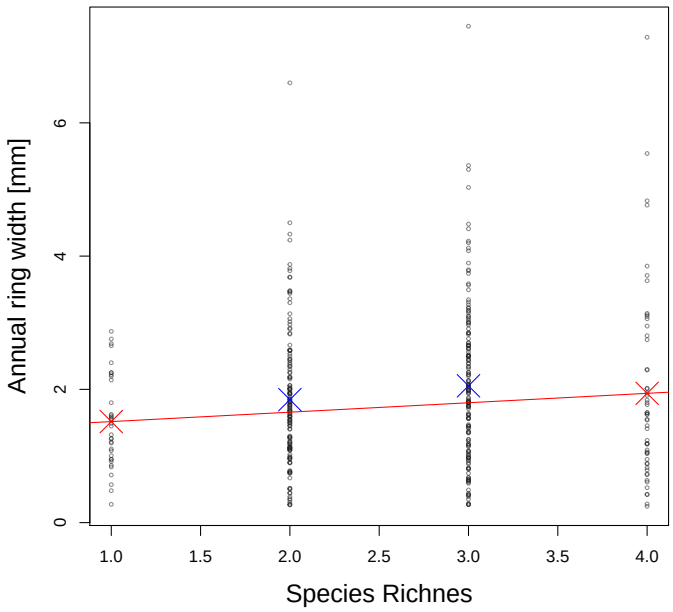


Figure 3: Visualisation of the estimtates of the constrained model lm.8.

This graph shows undoubtedly how the model is constrained. The regression line is forced to go through the first and the last species richness group (1 and 4). The coefficient obtained for `sp.rich.fac2` and `sp.rich.fac3` are simply the vertical distance between the red line and the two blue crosses.

To make this issue clearer, we can fit a model to a subset of the data where only species diversity 1 and 4 are present. This will give us exactly the same regression estimates for species richness as continuous variable (we increased the number of digits printed to make this clearer).

```
d.ring.9 <- subset(d.ring, subset = sp.rich %in% c(1, 4))
lm.9 <- lm(ring ~ sp.rich, data = d.ring.9)
coef(lm.9)
```

```
(Intercept)     sp.rich
    1.37505     0.14138
```

```
coef(lm.8)[1:2]
```

```
(Intercept)     sp.rich
    1.37505     0.14138
```

The output above confirms that the observations used to estimate the regression parameters (intercept and slope) for the continuous predictor species richness come from species richness 1 and 4 solely. As the other observations (groups 2 and 3) are not used to estimate the linear effect of species richness, we may modify them and still obtain the same regression estimates. Here below we show that situation by adding respectively 5 and 7 to the observations in groups 2 and 3.

```
## 1. creating fake data, where we add 5 and 7 to groups 2 and 3
d.ring$fake.ring <- d.ring$ring
```

```
d.ring$fake.ring[d.ring$sp.rich.fac == "2"] <-
  d.ring$fake.ring[d.ring$sp.rich.fac == "2"] + 5
d.ring$fake.ring[d.ring$sp.rich.fac == "3"] <-
  d.ring$fake.ring[d.ring$sp.rich.fac == "3"] + 7
##
## 2. fitting the model with the fake data
lm.8.fake <- lm(fake.ring ~ sp.rich + sp.rich.fac, data = d.ring)
##
## 3. lookimg at regression coeffients
coef(lm.8.fake)[1:2]
```

```
(Intercept)     sp.rich
    1.37505     0.14138
```

```
coef(lm.8)[1:2]
```

```
(Intercept)     sp.rich
    1.37505     0.14138
```

Indeed, the regression coefficients are again the same as for the model fitted on the true data. Let's visualise the data along with the estimated regression line for the fake data set.

```
## 1. plotting observations with transparency
plot(fake.ring ~ sp.rich, data = d.ring, col = rgb(0.1, 0.1, 0.1, 0.5), cex = 0.5,
     cex.lab = 1.5,
     xlab = "Species Richnes", ylab = "Annual ring width [mm]")
##
## 2. compute groups means and display them
fitted.y <- predict(lm.8.fake, newdata = data.frame(sp.rich = 1:4,
                                                    sp.rich.fac = factor(1:4)))
points(x = 1:4, y = fitted.y, col = c("red", "blue", "blue", "red"),
       cex = 3, pch = 4)
##
## 3. add the contrained regression line
abline(a = coef(lm.8.fake)["(Intercept)"], b = coef(lm.8.fake)["sp.rich"],
       col = "red", lwd = 2)
```

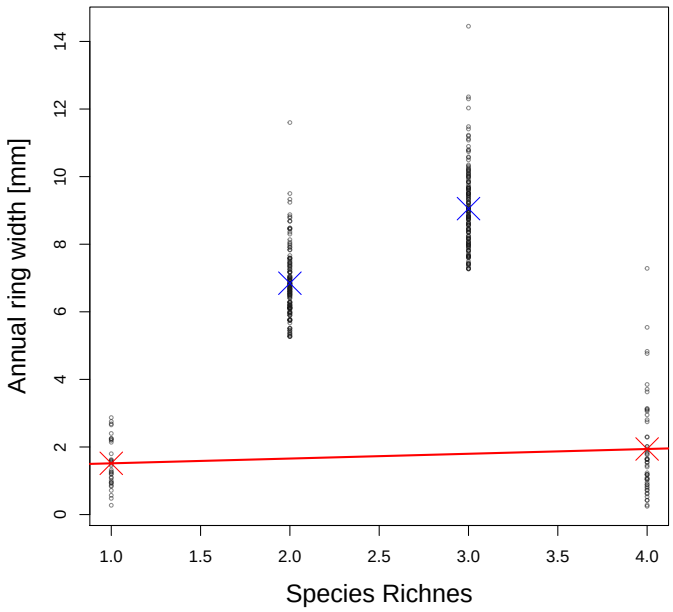


Figure 4: Visualisation of the estimtates of the constrained model lm.8.fake (fitted on fake data).

Figure 4 shows even more clearly the fact that the regression coefficient of a constrained model are in practice meaningless.

Changing reference level

So, the estimated regression coefficient for `sp.rich` of the constrained model `lm.8` refers to a subset of data only. This is undesirable and inappropriate. To better understand this situation we can change the reference level of the factor `sp.rich.fac`. Paradoxically, we can set the level of species richness `3` as reference and get a negative estimate for species diversity. This is because the group mean of species richness `3` is slightly higher than the group mean of group 4 (i.e. the other group constrained).

```
levels(d.ring$sp.rich.fac)
```

```
[1] "1" "2" "3" "4"
```

```
d.ring$sp.rich.fac.BIS <- relevel(d.ring$sp.rich.fac, ref = "3")
levels(d.ring$sp.rich.fac.BIS)
```

```
[1] "3" "1" "2" "4"
```

```
##
lm.10 <- lm(ring ~ sp.rich + sp.rich.fac.BIS, data = d.ring)
coef(lm.10)
```

```
     (Intercept)          sp.rich sp.rich.fac.BIS1 sp.rich.fac.BIS2
         2.39006         -0.11237         -0.76126         -0.31924
sp.rich.fac.BIS4
              NA
```

As expected we get different results for `sp.rich`. We could add a regression line

on Figure 3 that goes through the mean value of group `3` and `4` and this is exactly the `sp.rich` effect estimated on `lm.10`. Obviously there is no reason to set group `3` as reference in this context. This was done to show that the estimated regression coefficients for `sp.rich` in constrained models cannot be interpreted.

Testing

Of course the change of reference level also affects the p-values obtained for the species richness regression coefficient (see summary of the two linear models below). The p-values obtained for the non-constrained levels of the categorical variable `sp.rich.fac` (i.e. `sp.rich.fac2-4`) change as well.

```
summary(lm.8)


Call:
lm(formula = ring ~ sp.rich + sp.rich.fac, data = d.ring)

Residuals:
   Min     1Q Median     3Q    Max
-1.783 -0.761 -0.166  0.604  5.397

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.3750     0.2580    5.33  1.7e-07 ***
sp.rich        0.1414     0.0815    1.74    0.083 .
sp.rich.fac2   0.1883     0.1662    1.13    0.258
sp.rich.fac3   0.2538     0.1489    1.70    0.089 .
sp.rich.fac4       NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.12 on 389 degrees of freedom
Multiple R-squared:  0.0188,Adjusted R-squared:  0.0113
F-statistic: 2.49 on 3 and 389 DF,  p-value: 0.0601
```

```
summary(lm.10)


Call:
lm(formula = ring ~ sp.rich + sp.rich.fac.BIS, data = d.ring)

Residuals:
   Min     1Q Median     3Q    Max
-1.783 -0.761 -0.166  0.604  5.397

Coefficients: (1 not defined because of singularities)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.390      0.580    4.12  4.6e-05 ***
sp.rich            -0.112      0.177   -0.63    0.526
sp.rich.fac.BIS1   -0.761      0.447   -1.70    0.089 .
sp.rich.fac.BIS2   -0.319      0.252   -1.27    0.205
sp.rich.fac.BIS4       NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.12 on 389 degrees of freedom
Multiple R-squared:  0.0188,Adjusted R-squared:  0.0113
```

```
F-statistic: 2.49 on 3 and 389 DF,  p-value: 0.0601
```

These considerations about the p-values of the regression coefficients may seem in contradiction with the fact that we tested whether the non-linear term is needed. This is not the case. The F test used to test non-linearity is appropriate and correct. This latter is testing whether we can better model our data by adding two additional parameters to our model so that the effect of species richness is non-linear. The actual meaning of these two parameters (i.e. which depends on the model parametrisation) is irrelevant to this testing procedure. To prove that we can run the ANOVA test where the categorical variable species richness has a different reference than that one used previously to test for non-linearity.

```
anova(lm.10, lm.6)

Analysis of Variance Table

Model 1: ring ~ sp.rich + sp.rich.fac.BIS
Model 2: ring ~ sp.rich
  Res.Df RSS Df Sum of Sq    F Pr(>F)
1    389 490
2    391 493 -2     -3.74 1.49   0.23
```

This is indeed the same p-value that we obtained on page 11 when we compared the model with only `sp.rich` with the model that contained `sp.rich` and `sp.rich.fac` (i.e. `anova(lm.6, lm.8)`).

Sequential sum of squares (***)

Finally, is worth briefly discussing the results obtained when using sequential sum of squares. This type of testing is implemented in the function `anova()` when it is fed with a single model.

```
anova(lm.8)

Analysis of Variance Table

Response: ring
             Df Sum Sq Mean Sq F value Pr(>F)
sp.rich       1      6    5.66    4.50  0.035 *
sp.rich.fac   2      4    1.87    1.49  0.228
Residuals   389    490    1.26
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The first line of the ANOVA table, which is the test associated with species richness as a continuous variable (i.e. `sp.rich`), tests whether this predictor has an effect when no other predictors are present in the model. Therefore this test is identical to that obtained for model `lm.6` where there is only `sp.rich`.

The second line of the ANOVA table tests whether we need additional parameters to model non-linearity when species richness as a continuous predictor is in the model. This test is again identical to that ones obtained previously (i.e. `anova(lm.6, lm.8)` or `anova(lm.6, lm.10)`).

It is worth noting that the testing procedure used in this document relies on conditional sum of squares (i.e. type II sum of squares), where each term is assessed after controlling for all other terms in the model (Hector et al., 2010b). The reason for that choice is that we strongly rely on the *model-building approach* introduced by Pinheiro and Bates (Pinheiro and Bates, 2006). Indeed, most modern diversity-function experiments are best analysed within the mixed models framework for which this latter modelling approach was developed.

However, this publication does not rely on the use of conditional sum of squares. Nevertheless, it is worth mentioning that sequential sum of squares can be considered suboptimal to deal with non-identifiable models. Indeed, when factors are non-orthogonal, different orders should be used. Unfortunately, this is not the possible when dealing with species richness as continuous and as categorical variable[10]. As a matter of fact, we cannot change order here and include `sp.rich` as last predictor. It would simply be constrained to be zero.

Summary
To reiterate, adding a categorical variable for species richness to a model that contains the same predictor as continuous variable is useful to test for non-linearity. However, this model cannot be used to test the effect of species diversity as a continuous variable. In general, the estimated parameters and the associated p-values (i.e. the output of the `summary` function) of non-identifiable models are meaningless. If there is evidence that the categorical variable is needed (i.e. `sp.rich.fac`), we don't need to carry any other test as we can safely claim that "species richness has an effect, and this is non-linear". If there is no evidence for a non-linear effect, the categorical variable must be removed from the model before testing the effect of species richness as continuous variable.

Using a factor to model non-linearity is convenient as it makes no assumptions on the form of the relationship. However, this comes with a cost of larger number of parameters. If the number of levels of species richness is large, we may be using many parameters. When using a quadratic term, we just use an additional parameter. Nevertheless, most biodiversity function studies have a reasonably small number of diversity levels and this technique can safely be applied.

A real disadvantage of this approach is that it cannot be used with non-discrete predictors. We may want to model the effect of the "realised diversity" (as opposed to the "planned diversity") using the Shannon Index, so that we get an estimate of local diversity that takes relative abundance into account. In this case, the values of this predictor are not discrete and we cannot use a factor. Polynomials can deal with that, but again, make strong assumptions on the form of the relationship between the response variable and species richness. On the other hand, Generalised Additive Models discussed below (GAMs; also known as smoothing techniques) avoid making these assumptions (Wood, 2006).

[10] A simple and quick way to inspect for orthogonality is to look at the variance-covariance matrix of the estimated parameters (i.e. `vcov(lm.8)`). If the two predictors were orthogonal, all their shared entries of that matrix would be zero.

3.3 Smoothing techniques

These modern methods allow us to model non-linear relationships in a very powerful way. They don't make any assumption on the form of the response variable-predictor relationship other than saying that this is a smooth function (Wood, 2006).

In the study analysed here, realised diversity was also computed as the Shannon Index. We display the marginal effect that this predictor has on the response variable ring width.

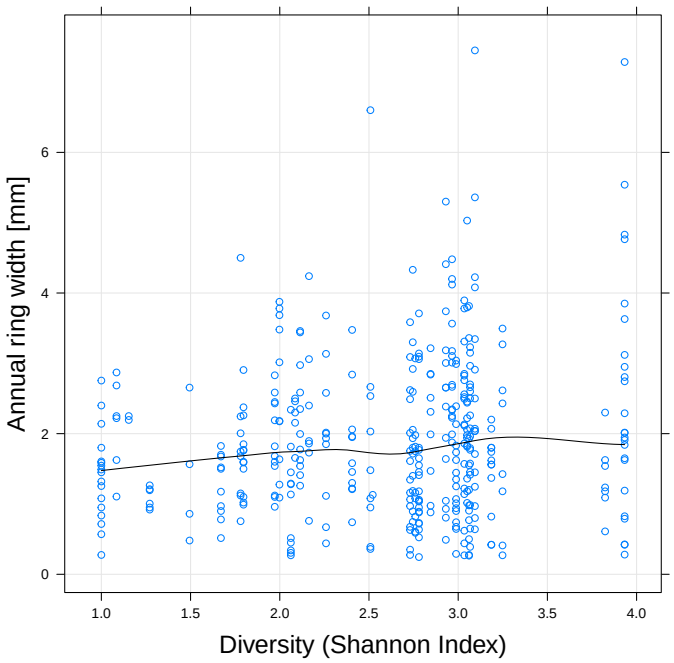


Figure 5: Response variable plotted against diversity measured with the Shannon Index.

We added a smoother to this plot to better characterise the relationship. The diversity measured as Shannon Index does not seem to have a large effect on ring growth. There is no clear evidence for non-linearity and there is some weak evidence for a positive effect. In addition, the variance seems to be increasing.

The smoother added to this plot uses indeed smoothing techniques. Below we use the *mgcv* package to formally fit a model where Shannon diversity is allowed to have a non-linear effect on ring growth (Wood, 2016).

```
require(mgcv)
gam.1 <- gam(ring ~ s(diversity.stand), data = d.ring)
summary(gam.1)


Family: gaussian
Link function: identity

Formula:
ring ~ s(diversity.stand)
```

```
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.9163     0.0563      34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                   edf Ref.df    F p-value
s(diversity.stand)   1      1 9.19  0.0026 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0205   Deviance explained =  2.3%
GCV = 1.2536  Scale est. = 1.2472    n = 393
```

This output confirms the initial impression. The estimated degrees of freedom (i.e. `edf`) of the smooth term for `ShannonIndex` is one (see `Approximate significance for smooth terms:` section of the output). This implies a linear effect of the Shannon index. This predictor appears to have a statistically significant effect[11].

In general, smoothing is very well suited to modelling non-linear relationship and makes very few structural assumptions. Obviously, this technique is not applicable to all situations. As an example, modelling the effect of a discrete predictor with few levels (e.g. `sp.rich`) with a smooth term is unneeded.

3.4 Summary

There are several ways to check for non-linearity of a species diversity effect. Here we discussed the use of polynomials, categorical variables and smoothing techniques. All come with pros and cons. The choice of the best method, depends on the design of the experiment and on the question addressed. Some of them make strong assumptions on the relationship between the response variable and species diversity but are simple to use and intuitive. As an example, the logarithmic transformation of the diversity predictor (not discussed here) assumes a clearly defined relationship, but is easy to use and to interpret (see for example Hector et al. (2010a)). On the other hand, smoothing is very flexible, makes very few assumptions, but requires some additional knowledge. The use of categorical variables and smoothing techniques can be considered a good solution to most situations encountered when modelling the effect of species richness in diversity function studies.

In addition, this section helped to clarify that the parameters of a non-identifiable (overdetermined) model are meaningless.

[11] Because its effect is estimated to be linear, we may prefer to remove it from the smooth terms and use it as a simple linear predictor (i.e. a classical regression predictor).

4 Alternative model parametrisations for biodiversity function experiments

Biodiversity function experiments contain two tightly linked variables, species composition and species richness. When analysing this kind of data the user may consider each of them to be a random or a fixed effect. We will show that this choice should be mainly based on the experimental design and on the research questions addressed and that some combinations are not viable options.

The predictor species richness may be taken as a continuous variable, by assuming a linear effect, or as a categorical variable. This allows for greater flexibility (i.e. the effect of species richness can be non-linear). On the other hand, species composition has no other choice but to be taken as a categorical variable.

It is worth mentioning that the issue addressed here is not technical in is nature, it is biological/practical. Usually, when we want to tell the effect of two predictors apart, we design an experiment where each of them is allowed to vary when the other is kept fixed. In this context, we would design an experiment were all combinations of species diversity and species composition exist and are, ideally, replicated a certain number of times. This is clearly not possible with species composition and species richness as, for example, the species composition `Picea a.` only exists as a monoculture.

In the next subsections we discuss several possible parametrisations. To do this we use a real data set that comes from an experiment on five species of grasses, the Zurich Experiment used in Appendix A. Details of the design can be found in Appendix A or here below as a footnote[12].

As usual practice we introduce this data set with an appropriate graph. We plot the biomass per m^2 against time. We used panels to better highlight differences among species richness levels (i.e. one, two and five).

[12]These five species were grown in monocultures, all ten 2-species compositions and the full mixture of five species. Each one of the 16 species composition was replicated five times (i.e. with five plots). The study lasted for five years (2004-2008) and was carried out in Zurich. The biomass of each of the 80 plots was measured eight times (in June and august of each year). More details can be found in (Vojtech et al., 2008). The analysis presented there is at species level. In other words, the grasses of each plot were sorted at species level and the corresponding biomasses weighted separately. Here we analyse the total biomass of each plot, we are thus analysing at the level of species composition. A detail that is worth mentioning, is that the surface harvested in years 2004-2006 was 0.25 m^2, while later in years 2007 and 2008 only 0.09 m^2 was harvested. To deal with that, we multiplied the measured biomasses so that all observations refer to 1 m^2.

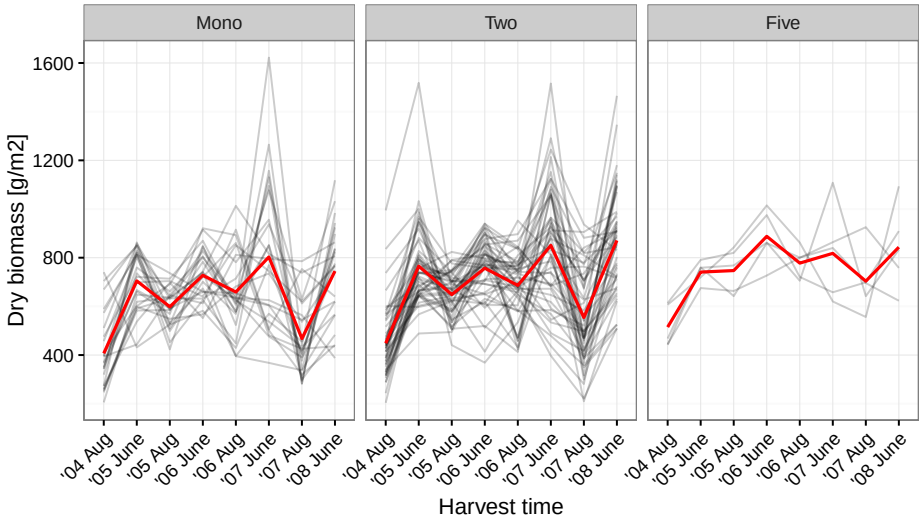


Figure 6: Biomass of the Zurich experiment plotted against time. The eight observations of each plot are connected with a black line. Red lines were used to display the temporal average of each species richness group.

By making use of the reference grid in the background of this graph, we can compare the mean biomasses across panels. It seems that productivity increases with species richness (from left to right). We also see that within species richness levels, there is some, non-negligible, variability among plots. The full mixture shown on the right panel has only five lines (i.e. plots) as each species composition was replicated five times.

In the next subsections we will try to model the biomass of each plot over time with species richness and species composition as predictors. Plots are considered to be random effects as we are not interested in their individual effects. Time is taken to be a categorical variable with eight levels.

There are essentially four possible options that we consider worth discussing here. We may consider i) species richness and species composition to be both fixed effects, ii) both to be random, iii) species richness to be fixed while species composition random and finally iv) just species composition as a fixed effect (as shown in Appendix A).

4.1 Species richness and species composition as fixed effects

Both predictors as factors

Each species composition level can obviously be associated with just one level of species richness. Therefore, species composition is fully nested in species richness. This implies that we cannot fit a model where both predictors are taken as factors. To be more precise, this can be done by adding additional constraints to the model, but the estimated coefficients and associated p-values are meaningless. To illustrate this situation we fit the following model.

```
mod.1 <- lmer(bio.m2 ~ sp.rich.fac + sp.comp + time.fac + (1 | plot), data = d.zh)
```

```
fixed-effect model matrix is rank deficient so dropping 2 columns / coefficients
```

As we expected the model is not identifiable and some parameters of `sp.comp` are set to zero. Nevertheless, we may want to use sequential sums of squares approach to test both predictors[13].

```
anova(mod.1)
```

```
Analysis of Variance Table
            Df   Sum Sq Mean Sq F value
sp.rich.fac  2   674383  337192   13.91
sp.comp     13  2635052  202696    8.36
time.fac     7 10888153 1555450   64.16
```

In this constrained model, species richness as a factor and species composition are not orthogonal[14], therefore different orders should be used. However, this is not feasible because if species composition was entered first, there would be no variance left to explain for species richness as a factor.

It is important to correctly interpret these two F tests shown above (Schmid et al., 2002). The first line of the ANOVA table tests whether species richness as a factor has an effect, when the predictor species composition is not included in the model. The second row of the table tests whether there is an effect of species composition when species richness as a factor is included in the model. In other words, we are testing whether within each level of species richness there is any variation left attributable to species composition.

Species richness as continuous variable

Another way to deal with both predictors as fixed effects, is to take species richness as a continuous variable. Again this model is overdetermined as the design matrix is rank-deficient.

```
mod.2 <- lmer(bio.m2 ~ sp.rich + sp.comp + time.fac + (1 | plot), data = d.zh)
```

```
fixed-effect model matrix is rank deficient so dropping 1 column / coefficient
```

As expected, one parameter is constrained to be zero. It is important to note that, in analogy to what we previously showed, the species richness effect estimated in this model depends exclusively on the first and last level of `sp.comp`. To confirm that we refit a model using the data from these two levels only.

```
nlevels(d.zh$sp.comp)
```

```
[1] 16
```

```
d.zh.subset <- subset(d.zh, subset = sp.comp %in% levels(d.zh$sp.comp)[c(1,
    16)])
mod.2.subset <- update(mod.2, data = d.zh.subset)
fixef(mod.2)["sp.rich"]
```

```
sp.rich
  22.75
```

[13] Note that the `anova()` method function does not report p-values. Type `?pvalues` for more information.

[14] You can confirm that by looking at the variance-covariance matrix of the model parameters with `vcov(mod.1)`.

```
fixef(mod.2.subset)["sp.rich"]
```

```
sp.rich
  22.75
```

Indeed, even if we work on a subset of the data the regression coefficient is the same. This confirms that the regression coefficient for species richness cannot be interpreted as the main effect of species richness. It solely depends on two levels of the species composition factor. To double check that, we can change the reference levels and we would obtain different results.

It is worth mentioning that the issue raised here is independent from the model contrasts used (i.e. treatment contrasts here). To convince ourselves, we can change them to sum contrasts[15]. Below we change the model contrasts and we also change the reference level of the `sp.comp` factor. As expected, the estimated effect of species richness does depend on the reference level. This is clearly inappropriate as the labelling of levels must not influence the results obtained. By changing the level that is constrained to be zero, we can obtained 15 different results.

```
## 1. change contrasts to sum contrasts
options("contrasts")
```

```
$contrasts
        unordered           ordered
"contr.treatment"      "contr.poly"
```

```
options(contrasts = c("contr.sum", "contr.poly"))
##
## 2. refit mod.2 with sum contrasts
mod.2.sum <- update(mod.2)
##
## 3. change the reference level of the "sp.comp" factor
d.zh$sp.comp.BIS <- relevel(d.zh$sp.comp, ref = levels(d.zh$sp.comp)[15])
##
## 4. fit the model with the new reference level (with sum contrasts)
mod.2.sum.ref <- update(mod.2, . ~ . - sp.comp + sp.comp.BIS)
##
## 5. compare the estimated effects of species richness
fixef(mod.2.sum)["sp.rich"]
```

```
sp.rich
    224
```

```
fixef(mod.2.sum.ref)["sp.rich"]
```

```
sp.rich
    818
```

```
## The results for species richness depend on the reference level
## even when sum contrasts are used !!!
##
## 6. set contrasts back to treatment contrasts
```

[15] The **R** function `options` allows to change contrasts for both ordered and unordered factors. Here we just change the contrasts for unordered factors as we are not dealing with ordered factors.

```
options(contrasts = c("contr.treatment", "contr.poly"))
```

This confirms that if we were to include both species composition and species richness as fixed effects the model would not be identifiable. This implies that some parameters are constrained to be zero and that the estimates obtained cannot be interpreted biologically. In other words, adding species richness to a model that already contains species composition as a fixed effect simply leads to a less meaningful re-parametrisation of the model. To show that we can compare these two models with an ANOVA.

```
## 1. fit a model with just species composition
mod.2.NoSr <- update(mod.2, . ~ . - sp.rich)
## 2. compare this latter model with mod.2
anova(mod.2, mod.2.NoSr)
```

```
Data: d.zh
Models:
mod.2: bio.m2 ~ sp.rich + sp.comp + time.fac + (1 | plot)
mod.2.NoSr: bio.m2 ~ sp.comp + time.fac + (1 | plot)
           Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
mod.2      25 8304 8416  -4127     8254
mod.2.NoSr 25 8304 8416  -4127     8254     0      0          1
```

This output clearly indicates that the two models are actually fully equivalent. Indeed, the fit is the same (see the log-likelihood `logLik`) and the number of parameters in the two models is the same (`Df`). However, the one with both predictors is non-identifiable and thus not interpretable.

Finally, we may argue that fitting a model that contains species richness as a continuous variable and species composition can be used for testing with sequential sums of squares. This is not entirely true, for two reasons. First, we cannot change their order in the testing procedure (species richness must come first). However, different orders are actually required as these two predictors are not orthogonal (Schmid et al., 2002; Venables and Ripley, 2013).

Second (***), the p-value associated with species richness does not correspond to the estimated regression coefficient of this model. This latter coefficient is meaningless as it comes from a constrained model. On the other hand, the sequential test for species richness is based on a unconstrained model (as it does not contain species composition). To actually get a meaningful regression coefficient for the model that does not contain species composition, we should refit the model without species composition. To better understand that, we can look at the ANOVA tables of a model that contains both biodiversity predictors and a model that does not contain species composition.

```
## 1. refit a model that does not contain species composition
mod.2.NoSpComp <- update(mod.2, . ~ . - sp.comp)
## 2. comparing the two ANOVA tables
anova(mod.2)
```

```
Analysis of Variance Table
          Df   Sum Sq Mean Sq F value
sp.rich    1   545068  545068   22.48
sp.comp   14  2764367  197455    8.14
time.fac   7 10888153 1555450   64.16
```

```
anova(mod.2.NoSpComp)
```

```
Analysis of Variance Table
         Df   Sum Sq Mean Sq F value
sp.rich   1   238825  238825    9.85
time.fac  7 10888153 1555450   64.16
```

Regardless of the predictors that follow species richness, its sums of squares are supposed to be the same for both ANOVA tables. This is clearly not the case and the reason is that `mod.2` is overdetermined.

Alternative biodiversity indexes

The issue raised above may be bypassed by using alternative biodiversity indexes. Indeed, if we were to use another biodiversity index it may happen that the model becomes identifiable and all parameters can be estimated. As an example the model that contains species composition and the Shannon Index as fixed effects is identifiable. However, this is not a viable solution as the interpretation of the coefficients still remains problematic. Indeed, the classical interpretation of regression coefficient cannot be used. Namely, we cannot increase species richness while keeping species composition fixed.

Fitting two separate models

Another potential solution is to fit and compare two separate models, one that contains species composition only and the other that just contains species richness and then formally compare them.

```
mod.3 <- lmer(bio.m2 ~ sp.rich.fac + time.fac + (1 | plot), data = d.zh)
mod.4 <- lmer(bio.m2 ~ sp.comp + time.fac + (1 | plot), data = d.zh)
anova(mod.3, mod.4)
```

```
Data: d.zh
Models:
mod.3: bio.m2 ~ sp.rich.fac + time.fac + (1 | plot)
mod.4: bio.m2 ~ sp.comp + time.fac + (1 | plot)
      Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
mod.3 12 8356 8410  -4166     8332
mod.4 25 8304 8416  -4127     8254  77.9     13    2.7e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Although the `anova()` function returns p-values these are invalid as the two models are not formally nested (Pinheiro and Bates, 2006). Indeed, the design matrix of the smaller model (i.e. `mod.3`) is not a subset of the design matrix of the larger model (i.e. `mod.4`). In these cases, we can still look at the AIC and BIC to compare the two models (Pinheiro and Bates, 2006). Unfortunately, in this specific case, these two information criteria do not agree on the best model to choose.

This way to proceed has also other disadvantages. As an example, if we were to choose the model with just species richness, we would ignore a design variable (i.e. species composition). Design variables should always be retained in a model (Hurlbert, 1984).

4.2 Species richness and species composition as random effects

If we were interested in the relative contributions of species richness and species composition we may naturally want to take both factors as random and compare their variance components (Hector et al., 2011a). This procedure has the great advantage that both variables can be included in the model without any problem of non-identifiability. As a matter of fact, the estimated parameters are two variance components and not all levels of the two factors. Obviously, this means that we cannot formally compare levels of either of these two factors (e.g. species richness 1 vs. 2). Interestingly, we can also use the estimate of the residual variance to understand whether the estimated effects are relevant or not. In other words, whether the effect of these two factors is of any practical significance. Below we fit this model on the Zurich data and we look at the variance components in the standard deviation scale.

```
mod.5 <- lmer(bio.m2 ~ time.fac + (1 | sp.comp) + (1 | sp.rich) + (1 | plot),
    data = d.zh)
VarCorr(mod.5)
```

```
 Groups   Name        Std.Dev.
 plot     (Intercept)   1.33
 sp.comp  (Intercept)  67.14
 sp.rich  (Intercept)  27.59
 Residual             155.71
```

Here we may conclude that species richness is, roughly speaking, about half as important as species composition (resp. 28 and 67). However, when looking at the residual variance we notice that both these factors are far less important than the residual error. In other words, the practical significance of these differences could be questioned as the unexplained variability is greater than the variability explained by both species richness and species composition together.

Another important point to highlight here, is the fact that these variance components are based on few levels, and may therefore be estimated with low precision. The authors of the software used here (i.e. `lme4`), recommend to have at least about 6 levels (Bates, 2010). To inspect whether we have enough levels to estimate, in a reasonable way, this variance component, we can look at their confidence intervals[16].

```
nlevels(d.zh$sp.rich.fac)  ## very few levels !!!
```

```
[1] 3
```

```
nlevels(d.zh$sp.comp)
```

```
[1] 16
```

```
##
confint(mod.5, parm = 2:4)
```

```
       2.5 % 97.5 %
.sig02    45    104
.sig03     0    104
```

[16] `lme4` by default estimates confidence intervals of variance components with profiling methods. Type `?confint.merMod` for details.

```
.sigma   146    164
```

As expected, the confidence intervals for the variance component of species richness (i.e. `.sig03`) is extremely wide and includes zero. This means that with such a low number of levels, here just three, species richness cannot be taken to be a random effect as the estimated variance is too poorly defined. Note that the residual variance (i.e. `.sigma`) can be estimated well as it is based on many observations. The variance component for species composition (i.e. `.sig02`), which is based on 16 levels and can be estimated with a low precision (estimate: 67; limits: 45, 104).

Application of Gelman's "Bayesian ANOVA"
Hector and colleagues (2011a) proposed to use Bayesian methods to estimate, in a reasonable manner, both variance components[17]. This approach has the great advantage of being very useful to compare the relative contribution of the two biodiversity predictors and, at the same time, their practical significance. Its downside is that it relies on a less commonly used framework (i.e. Bayesian statistics). Nevertheless, this should not be taken as a hindering factor. This approach cannot be used with alternative biodiversity indexes (e.g. Shannon Index) as it only works with discrete variables and not with continuous. Other modern methods can deal with that, nevertheless, their implementation and the interpretation of the different variance components can be difficult (Wood and Scheipl, 2016).

To conclude, although some care must be taken, the method proposed by Hector and colleagues is very useful to compare the relative importance of species richness and species composition with a handy and powerful graphical representation.

4.3 Species composition as random and species richness as fixed

So far we have discussed two possible parametrisations where both biodiversity predictors were modelled both as either fixed or random. A convenient and often used parametrisation is to take species richness as fixed effect (as we are interested in its effect), and species composition as random effect (as this must be considered, as it is part of the design, but is not of primary interest). This parametrisation does not have any identifiability problem. Considering species composition as a random effects can be very useful when its number of levels is very large and they are not of interest (e.g. Hector et al. (2011b)).

This parametrisation allows us to work with continuous biodiversity variables such as the Shannon Index. In addition, we may take species richness as a factor or as a smooth term to test for non-linearity of its effect. From a very rigorous and theoretical point of view, the downside of this approach is that we are assuming all levels of the species composition factor to be independent. In reality, some species compositions are

[17]The authors proposed to estimate the so called "finite population" variance according to the approach developed by Gelman (Gelman, 2005). It has to be noted that, in the strict sense, species richness should rather be estimated with the "superpopulation" variance as not all possible levels are available. This is equivalent to what we presented here with `lmer()` (i.e. `mod.5`).

more alike than others. As an example, the hypothetical species composition $ABCD$ is expected to be more similar to ABC than the monoculture E. From a pragmatical point of view, this argument has little importance[18], especially if the compositions used are truly a random subset of all possible combinations.

To conclude, this parametrisation is convenient as it enables us to draw conclusions about the effect of species richness, about its linearity, but at the same time it includes the fundamental design variable species composition. The relevance of this latter can be assessed by comparing its variance component with the estimated residual variance. It is interesting to note that this parametrisation allows us to model the effects of the two biodiversity predictors with only two parameters (i.e. a slope and a variance component). Finally, the major practical downside of this approach is that no conclusions can be drawn about the single levels of species composition. The next subsection illustrates how this can be done with an alternative parametrisation.

4.4 Species composition as fixed alone

Fitting a diversity-function model with only species composition (as a fixed effect) is the best way to test specific hypotheses regarding different levels of this categorical variable. This is done through the use of post-hoc testing. This way to proceed is not shown here as it is extensively presented in Appendix A. Below we illustrate the reasons not to include other diversity predictors in the model when we aim to compare species compositions.

Adding other diversity predictors to a model that already contains species composition as a fixed effect leads to a model where the coefficients are not meaningful. These models cannot be used to test hypotheses about specific groups of species compositions. To better explain that, we can look at a real example. Below we fit two models and look at their coefficients.

```
mod.6 <- lmer(bio.m2 ~ sp.comp + (1 | plot), data = d.zh)
fixef(mod.6)
```

```
(Intercept)   sp.compAn   sp.compAr    sp.compF    sp.compH sp.compAlAn
      663.2       -85.7         4.7       -94.6        55.0       101.0
sp.compAlAr  sp.compAlF  sp.compAlH sp.compAnAr  sp.compAnF  sp.compAnH
       30.9        33.6       140.5       -41.6       -95.6        28.7
 sp.compArF  sp.compArH   sp.compFH  sp.compMix
      -22.2       121.9        47.7        91.0
```

The output above tells us, to make an example, that the species composition `AlAn` produces 101 g/m^2 more than the reference level (i.e. `Al`). We can double check that by using the a post-hoc test.

```
require(multcomp)
glht(mod.6, linfct = mcp(sp.comp = "AlAn - Al = 0"))


 General Linear Hypotheses
```

[18] Although there are software that allow fitting models with random effects that have specific correlation structures (Rue et al., 2009), this is computationally very challenging.

```
Multiple Comparisons of Means: User-defined Contrasts


Linear Hypotheses:
               Estimate
AlAn - Al == 0      101
```

Indeed, the same results are obtained.

If we were to modify our model and include species richness as a continuous variable, these results would change.

```
mod.6.sp.rich <- update(mod.6, . ~  + sp.rich + .)
fixef(mod.6.sp.rich)[c("(Intercept)", "sp.rich", "sp.compAlAn")]
```

```
(Intercept)     sp.rich sp.compAlAn
        640          23          78
```

We see that the coefficient for species composition `AlAn` is now just 78 (it was 101). This is because we have to interpret this result in a different way. The estimated regression coefficient `AlAn` is not simply the difference from the reference level any more. It is the difference from the reference level when the effect of all other predictors, including `sp.rich`, has been removed. The species composition `AlAn` has richness two, hence one unit higher than the reference level (i.e. the monoculture `Al`). Therefore, to truly see the difference between `Al` and `AlAn` we must also add this effect. In this example, the effect of `sp.rich` is estimated to be 23. This confirms the fact that `AlAn` produces 101 g/m^2 more than the reference level[19]. Indeed, by summing species richness effect and species composition effect we obtain: $78 + 23 = 101$.

Note that this issue of interpretation is present even with identifiable models (`mod.6.sp.rich` was not). To make an example, a model where the Shannon Index was used, would be identifiable. However, the interpretation of the species compositions levels coefficients is affected as shown above.

To conclude, if the aim of the analysis is to compare different levels of the species composition factor, then the model must be fitted with this latter predictor as a fixed effect and no other diversity predictor must enter the model.

4.5 Summary

Below we briefly summarise when each of the four model parametrisations discussed is most useful.

1. both biodiversity predictors as fixed
   to be avoided, it can only be used for testing (but with issues)
2. both as random
   when the relative contribution of the two diversity predictors is to be compared.

[19]Note that the function `glht()` does not work with `mod.6.sp.rich` as it is a non-identifiable model.

3. species richness as fixed, species composition as random
   if the interest is the effect of species richness. Or when there is not enough replication of each species composition to take this latter as a fixed effect.
4. species composition only
   when we are interested in comparing groups of species composition levels as shown in Appendix A.

5 Final Remarks

This supporting information highlighted the importance of fitting models where the estimated parameters are meaningful and can be interpreted. This is fundamental in ecology as well as in other biological fields.

The approach presented in this publication strongly focuses on the biological interpretation of the results. It is thus fundamental to use models whose parameters can be interpreted in a biologically meaningful way.

Reproducibility

Checkpoint date was set to 2016-05-01.

```
sessionInfo()

R version 3.3.2 (2016-10-31)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Debian GNU/Linux 8 (jessie)

locale:
 [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
 [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
 [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] multcomp_1.4-6  TH.data_1.0-7   MASS_7.3-45     survival_2.40-1
 [5] mvtnorm_1.0-5   ggplot2_2.1.0   mgcv_1.8-15     nlme_3.1-128
 [9] lme4_1.1-12     Matrix_1.2-7.1  lattice_0.20-34 knitr_1.14

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.7      magrittr_1.5     splines_3.3.2    munsell_0.4.3
 [5] colorspace_1.2-7 minqa_1.2.4      stringr_1.1.0    highr_0.6
 [9] plyr_1.8.4       tools_3.3.2      grid_3.3.2       gtable_0.2.0
[13] digest_0.6.10    nloptr_1.0.4     reshape2_1.4.1   formatR_1.4
[17] codetools_0.2-15 evaluate_0.10    labeling_0.3     sandwich_2.3-4
[21] stringi_1.1.2    scales_0.4.0     zoo_1.7-13
```

References

D. Bates and M. Maechler. *Matrix: Sparse and Dense Matrix Classes and Methods*, 2016. URL `https://CRAN.R-project.org/package=Matrix`. R package version 1.2-7.1.

D. Bates, M. Maechler, B. Bolker, and S. Walker. *lme4: Linear Mixed-Effects Models using 'Eigen' and S4*, 2016. URL `https://CRAN.R-project.org/package=lme4`. R package version 1.1-12.

D. M. Bates. lme4: Mixed-Effects Modeling with R. *URL http://lme4. R-forge. R-project. org/book*, 2010.

B. J. Cardinale, J. E. Duffy, A. Gonzalez, D. U. Hooper, C. Perrings, P. Venail, A. Narwani, G. M. Mace, D. Tilman, D. A. Wardle, et al. Biodiversity loss and its impact on humanity. *Nature*, 486(7401):59–67, 2012.

J. Chamagne, M. Tanadini, D. Frank, R. Matula, C. Paine, C. D. Philipson, M. Svátek, L. A. Turnbull, D. Volařík, and A. Hector. Forest diversity promotes individual tree growth in central European forest stands. *Journal of Applied Ecology*, 2016.

J. M. Chambers and T. J. Hastie. *Statistical Models in S*. CRC Press, Inc., 1991.

A. Gelman. Analysis of variance why it is more important than ever. *The Annals of Statistics*, 33(1):1–53, 2005.

A. Hector, Y. Hautier, P. Saner, L. Wacker, R. Bagchi, J. Joshi, M. Scherer-Lorenzen, E. Spehn, E. Bazeley-White, M. Weilenmann, et al. General stabilizing effects of plant diversity on grassland productivity through population asynchrony and overyielding. *Ecology*, 91(8):2213–2220, 2010a.

A. Hector, S. Von Felten, and B. Schmid. Analysis of variance with unbalanced data: An update for ecology & evolution. *Journal of Animal Ecology*, 79(2):308–316, 2010b.

A. Hector, T. Bell, Y. Hautier, F. Isbell, M. Kery, P. B. Reich, J. van Ruijven, and B. Schmid. BUGS in the analysis of biodiversity experiments: Species richness and composition are of similar importance for grassland productivity. *PLoS One*, 6(3):e17434, 2011a.

A. Hector, C. Philipson, P. Saner, J. Chamagne, D. Dzulkifli, M. O'Brien, J. L. Snaddon, P. Ulok, M. Weilenmann, G. Reynolds, et al. The Sabah Biodiversity Experiment: A long-term test of the role of tree diversity in restoring tropical forest structure and functioning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1582):3303–3315, 2011b.

S. H. Hurlbert. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, 54(2):187–211, 1984.

J. Pinheiro and D. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer Science & Business Media, 2006.

H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

B. Schmid, A. Hector, M. Huston, P. Inchausti, I. Nijs, P. Leadley, and D. Tilman. *Biodiversity and Ecosystem Functioning: Synthesis and Perspectives. Oxford University Press*, chapter : The design and analysis of biodiversity experiments, pages 61–75. 2002.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-PLUS*. Springer Science & Business Media, 2013.

E. Vojtech, M. Loreau, S. Yachi, E. M. Spehn, and A. Hector. Light partitioning in experimental grass communities. *Oikos*, 117(9):1351–1361, 2008.

S. Wood. *Generalized Additive Models: An introduction with R*. CRC press, 2006.

S. Wood. *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*, 2016. URL `https://CRAN.R-project.org/package=mgcv`. R package version 1.8-15.

S. Wood and F. Scheipl. *gamm4: Generalized Additive Mixed Models using 'mgcv' and 'lme4'*, 2016. URL `https://CRAN.R-project.org/package=gamm4`. R package version 0.2-4.

Chapter 3

A unified model-based framework for the analysis of biodiversity-ecosystem functioning experiments

3.1 Preface

Authors: Matteo Tanadini[1] & Andrew Hector[1]
1. Department of Plant Sciences, University of Oxford, South Parks Road, OX1 3RB.

Target readership: Researchers working in the filed of biodiversity ecosystem functioning experiments.

Context: This chapter illustrates my approach to measure species asynchrony, ecosystem stability and how to interlink the results of analyses ran on different response variables. This chapter represents, along with chapter 3, the core of the thesis.

Article status: This article is in preparation to submit.

3.2 Main text

3.2.1 Biodiversity ecosystem-functioning experiments

Biodiversity-ecosystem functioning experiments are studies that, as their name implies, investigate the relationship between biodiversity and ecosystem functioning (Cardinale et al., 2012). These experiments are carried out with different organisms such as herbaceous species to mimic grasslands, algae to mimic aquatic environments and more recently with trees to mimic forests (Tilman et al., 2014; Gross et al., 2014). In these experiments, several tightly interlinked biological questions are addressed. Possibly the most frequently encountered questions are whether biodiversity: i) Increases ecosystem productivity; and ii) stabilises ecosystems (Cardinale et al., 2012).

To answer the wide variety of questions addressed in each of these studies several separate analyses are carried out independently. These analyses use different approaches that can practically be divided into two groups (described below).

3.2.2 The analysis of biodiversity ecosystem-functioning experiments

Primary analyses

Usually, the first phase of the analysis of a biodiversity-ecosystem functioning experiment, also called a diversity-function experiment, is to test whether biodiversity has an effect on the ecosystem function of interest (e.g. productivity). Most of the time, this is carried out by fitting a Linear Mixed Effects Model (Schmid et al., 2002). We refer to this first part of the analysis of these experiments as the "primary analysis".

To make a practical example of a primary analysis, we can look at a hypothetical grassland study, where the response variable analysed is the dry biomass harvested in each plot. Plots that vary in their species diversity are measured repeatedly over time (e.g. yearly). To analyse these data we could fit the following model.

$$\text{dry.biomass} = \text{biodiversity} + \text{year} + \text{plot} + \text{block} + \text{soil.fertility}$$

Where "biodiversity" is the variable of interest and may represent, for example, the number of species in a given plot. "year", "plot" and "block" are so-called design variables. Finally, "soil.fertility" is a confounding variable for which we want to control for (i.e. a covariate)[1].

Forest studies usually measure the response variable at organism level (as opposed to grassland studies that usually work at community or population level). We may want to test whether biodiversity has an influence on tree growth as seen in annual diameter increase. In this case we could fit the following model.

diameter = biodiversity + species.id + year + tree + plot + forest.dens

Where "biodiversity" is the variable of main interest. The response is measured at tree level, we thus include "species identity" in our model so that differences in growth among species are accounted for. Again, "year", "tree" and "plot" are design variables. The main reason to include design variables is to account for the non-independence of groups of observations. For example, all observations of each one tree cannot be considered independent. Finally, "forest.density" is a confounding variable for which we want to control for.

To summarise, these models fitted in the primary analysis usually contain: i) The variable of main interest (i.e. biodiversity); ii) design variables; and iii) confounding variables. Depending on the design of the experiment and on the questions addressed, a given variable, for example soil.fertility could be considered to be a confounding variable or the variable of main interest. In addition, some variables can belong to more groups simultaneously. For example, year can be considered to be a confounding variable as well as a design variable.

It is important to note that in a primary analysis, when the effect of biodiversity on tree diameter, for example, is analysed, several other variables are considered and included in the model. One reason several variables are included in the model is to take the non-independence of groups of observations (i.e. avoid pseudoreplication) into

[1]Real models are usually more complex, this model is shown for illustrative purposes.

account. Another reason is to remove the confounding effect of those variables that cannot be controlled experimentally and that add noise to the process of interest.

This latter modelling approach is said to be conditional as the effect of the predictors of interests, for example species richness, on the response variable is estimated while removing (i.e. conditioning on) the effect of design and confounding variables. The other analyses carried out in diversity-function studies that we will introduce shortly use a fundamentally different approach.

Secondary analyses

On top of the effects of biodiversity on, say productivity, researchers are also interested in other aspects of ecosystem functioning (Cardinale et al., 2012). To mention a few examples, i) whether biodiversity stabilises ecosystems (stability), ii) whether species respond differently to environmental heterogeneity and to temporal fluctuations (asynchrony) and iii) whether the full mixture of species yields more than the weighted average of the monocultures (overyielding). These questions are answered with an approach that is fundamentally different from the one used in the primary analysis (hereafter, we refer to these analyses as "secondary analyses").

In particular, field-specific methods (metrics) are used rather than mainstream statistical techniques. These metrics were developed independently and use different approaches to answer distinct but closely related questions. For example, an ecosystem that is composed of asynchronous species is very likely a stable ecosystem (Doak et al., 1998). However, the metrics used to quantify ecosystem stability and species asynchrony are based on different frameworks that hinder the sharing of information among these two analyses. Information gleaned in each of the secondary analyses is thus not shared among them, nor is it with the primary analysis. For example, when overyielding is quantified, the information about the effect of soil fertility on biomass production, that was estimated in the primary analysis, is ignored. Additionally, the previously gathered knowledge about the variability among plots or blocks is not used. This approach is said to be marginal, as when these processes (e.g. asynchrony) are quantified, no other factors are contemplated. Put simply, these metrics assume that

nothing else other than the process of interest can influence the response variable.

The rationale of all the metrics used in secondary analyses is to condense a complex ecological process into a single meaningful value. For example, stability is often measured with the coefficient of variation, a metric that estimates the variability (over time or space) of the response variable of an ecosystem and divides it by its mean value ($CV = \sigma/\mu * 100$).

Example 1

To better explain the concept of marginality introduced above, we look at an example. Suppose that we were to quantify species asynchrony over time (i.e. temporal asynchrony). Roughly speaking, we can identify three different situations: Positively asynchronous species (species that react in the same way to temporal fluctuations), no asynchrony (species that react independently) and negatively asynchronous species (species that react in opposite directions to temporal fluctuations). These three situations are illustrated in Figure 3.1.

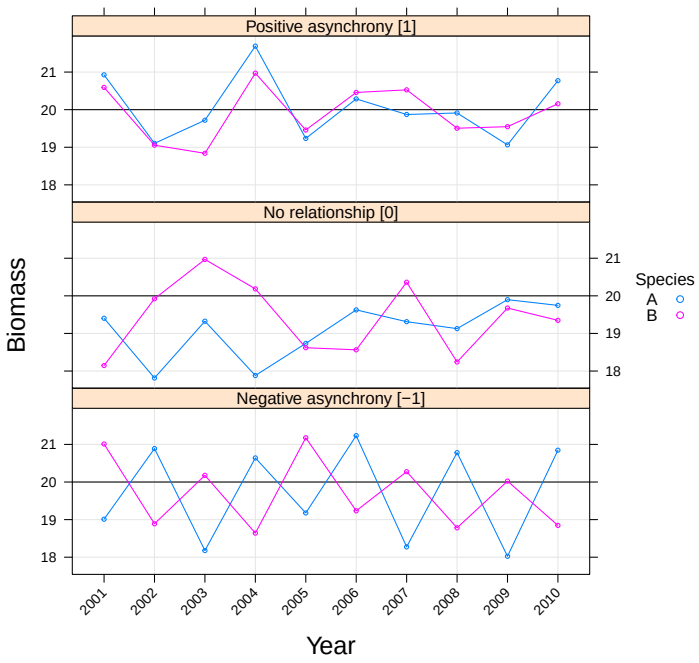


Figure 3.1: Three hypothetical situations of species asynchrony. For the sake of simplicity, only two species are used. The value in the panel strips (-1, 0, 1) indicate the theoretical asynchrony value expected if no noise were added to the data.

When the species biomasses are fluctuating around a stable mean value (here 20) and nothing else influences them, then marginal metrics provide a useful and meaningful estimate of asynchrony. By using the asynchrony metric proposed by Gross and colleagues (2014), currently a popular choice, the estimated values are respectively 0.8 (positive asynchrony; top panel), -0.2 (no relationship; middle panel) and -0.8 (negative asynchrony; bottom panel).

However, the situation, in which species fluctuate around a stable mean, is unlikely to be encountered in most real diversity-function studies. For example, if we were to model tree diameter, than the mean would not be stable, but increasing over time. Figure 3.2 shows the same three situations seen above, with the sole difference, that the mean biomass is not stable.

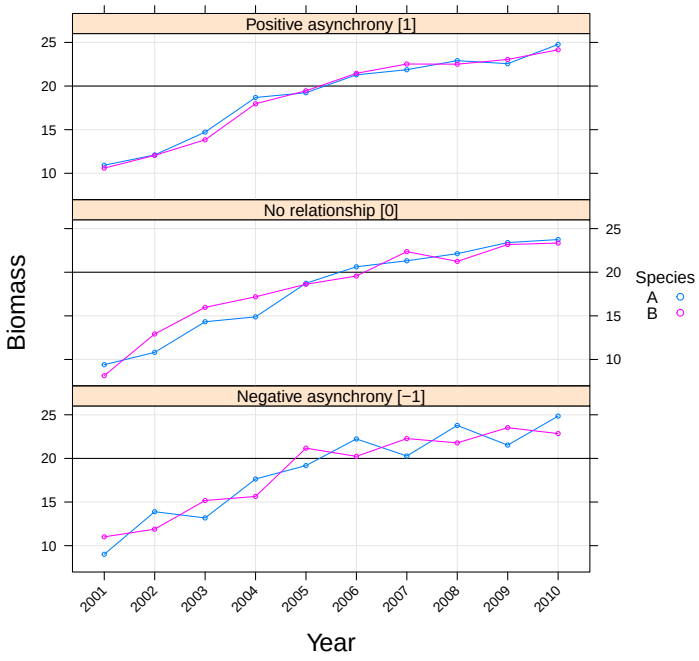


Figure 3.2: Three hypothetical situations of species asynchrony. A trend is present in the data.

By looking carefully across panels, we can still see that, although both species are growing in all three cases, they reveal strikingly different patterns. Unfortunately, any marginal metric would ignore the trend present in the data and estimate species to be strongly positively asynchronous (asynchrony measured according to Gross is higher

than 0.9 in all three cases). The reason is that these metrics do not discriminate between "trend" (or growth) and actual temporal fluctuations.

Paradoxically the trend, which is causing the issue, has been estimated in the primary analysis (as "year" effect), but this valuable information is subsequently ignored in the secondary analyses. In this case, growth is biasing the results as one of the fundamental assumptions of marginal asynchrony metrics, a stable mean, is not fulfilled.

There is other information that was estimated in the primary analysis and that could be used when looking at asynchrony. For example, we estimated the effect of forest density on diameter. The latter effect is considered to be constant over time, so it is not introducing bias when temporal asynchrony is considered. However, the effect of forest density on the response variable introduces noise in the estimation of asynchrony. Ideally, we would like to be able to remove this confounding effect as estimated in the primary analysis. Other secondary analyses suffer from the same issues when marginal metrics are used. For example, stability measured with the coefficient of variation gives also biased results when a trend is present in the data. In particular, ecosystem stability is underestimated as well as the differences in stability among groups (e.g. species diversity groups).

It is important to note that biased results are obtained in other situations where the mean value is not stable. For example, there may be year-to-year fluctuations, with some years where all species perform better and other years where they perform less well. Figure 3.3 illustrates this situation with an hypothetical grassland study ran over ten years.

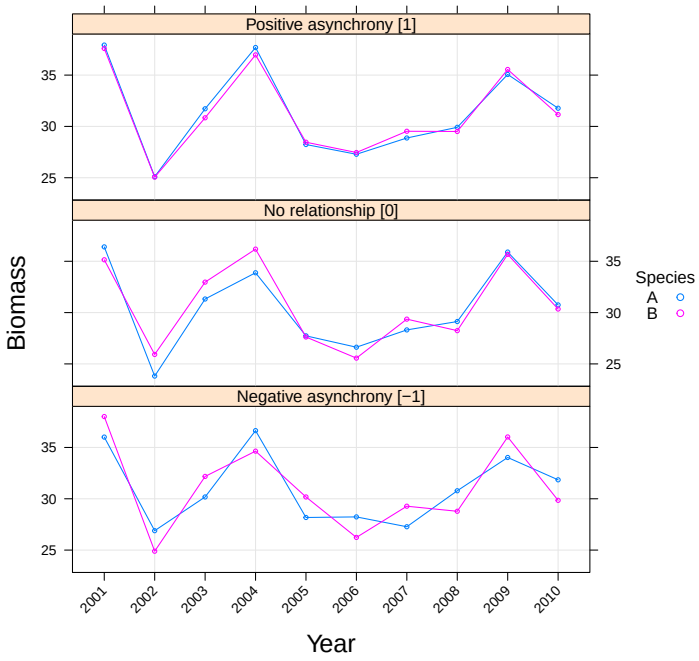


Figure 3.3: Three hypothetical situations of species asynchrony. Shared annual fluctuations are present in the data.

When shared annual fluctuations are present (here 2002 was a bad year for all species), then temporal asynchrony is estimated to be very high (close to 1) in all three cases. However, the species in the bottom panel have opposite behaviours when the shared year effect is removed. Biologically, it would be interesting to tell these two processes of temporal asynchrony apart. One component that estimates the shared responses to annual fluctuations, and another that estimates the magnitude of species-specific deviations from the shared year effect.

Assumptions made by metrics used in secondary analyses

Most metrics used in the secondary analysis of diversity-function experiments make somewhat unrealistic assumptions about the process analysed. Often a stable mean value of the response variable (over time or space) is assumed. In addition, no other factors, that may influence the response variable used to computed these metrics, are considered. Many of these assumptions are likely not to be fulfilled in present-day diversity-function experiments. We offer two possible explanations to explain the mismatch between the metrics used in the secondary analyses and the real situations

encountered in practice. On one hand, the first diversity-function experiments were ran in very controlled laboratory conditions. On the other hand, these metrics are mostly developed in a theoretical framework based on simulations where the effects of confounding or biasing factors is not taken into account. The fact that the secondary metrics lack a formal testing framework is also likely to be a legacy of the simulations framework. In turn, the use of a theoretical framework to develop and evaluate new metrics is likely to have been strongly influenced from decades of theoretical modelling.

Summary

To summarise, the primary analysis of a diversity-function experiment is usually carried out using a linear modelling approach. Then, marginal metrics that assume nowadays unrealistic conditions are used to estimate other important ecosystem properties. This can lead to misleading results. In addition, there is currently no formal connection among all these analyses and thus no sharing of information between and within these two phases (primary and secondary) of the analysis of these studies.

3.2.3 Proposed approach

We propose to put these two phases together into a single framework so that the information gleaned in each part of the analysis is shared and used by all the other parts. In particular, we propose to extend the Linear Mixed Effects Model used in the primary analysis to answer the questions addressed in the secondary analysis. Put simply, the model fitted to the data for the primary analysis is adapted such that processes such as stability or asynchrony can be quantified.

Our approach is extensively described in the appendices accompanying this publication. These studies have often a complex design with plot and blocks measured over time. Therefore, an appropriate visualisation of the data is fundamental. Appendix A focuses on the graphical visualisation of data stemming from diversity-function studies. Appendix B looks at how to quantify spatial and temporal asynchrony. Appendix C focuses on ecosystem stability. Finally Appendix D shows how the analysis run

on different response variables can be interlinked to answer new interesting biological questions.

Temporal and spatial asynchrony

Put simply, to model temporal asynchrony we simply allow species to interact with time when modelling our data. For example, the model previously fitted to diameter would be modified as follow.

$$\text{diameter} = \text{biodiversity} + \text{species.id} + \text{year} + \text{tree} + \text{plot} + \text{forest.dens} + \mathbf{\text{species.id:year}}$$

This last term introduced in the model accounts for the fact that species may react differently to years. Analogously, we can estimate spatial asynchrony by allowing species to interact with space.

$$\text{diameter} = \text{biodiversity} + \text{species.id} + \text{year} + \text{tree} + \text{plot} + \text{forest.dens} + \mathbf{\text{species.id:plot}}$$

Again, here we extend our model so that species are allowed to have different response to an heterogeneous environment. The species-specific spatial preferences estimated here can then be used to quantify spatial asynchrony. It is important to note that the procedure to measure asynchrony proposed here takes into account confounding and biasing effects. In Appendix B, we show in great detail how this approach works and what its advantages are.

Ecosystem stability

Stability can be easily be incorporated in our model-based framework as recently proposed by Carnus and colleagues (2015). In particular, residual variances are estimated separately in each group of interest (e.g. species richness groups). This way to proceed is very convenient as all biasing and confounding effects are quantified and controlled for. In Appendix C, we extend the approach introduced by Carnus so that stability can be measured at several organisational levels (organism, population, community,

ecosystem). We also show how to relax some of the assumptions made by Carnus (in particular, the assumption that the data follow a normal distribution). Our way to measure stability is strongly based on a graphical approach.

Comparing functioning of ecosystems

To model overyielding or compare functioning of different ecosystems in general, we propose to use post-hoc contrasts as proposed by Tanadini and colleagues (in prep.a).

3.2.4 Advantages of our approach over marginal metrics

Interpretation of the results

The most important advantage our proposed approach possibly has over marginal metrics is a facilitated interpretation of the results. Secondary metrics were developed independently and use different approaches that differ in scales and meaning of the obtained values. To make an hypothetical example, we may report the results of a diversity-study as follows: "Overyielding measured with the method proposed by Loreau (1998) is $\overline{D} = 1.7$; species pairwise interactions, that are measured with the method proposed by Connolly and colleagues (2013), are estimated to be $\theta = 0.73$; ecosystem stability, measured in the four species richness levels with the coefficient of variation, is respectively 501, 409, 376 and 251; finally, temporal asynchrony, measured as the mean pairwise correlation of all species, is $\overline{\rho} = -0.12$." This sentence highlights the difficulty in interpreting the results when secondary metrics are used.

In contrast, all the results reported with our approach are on the scale of the response variable and can easily be visualised and interpreted. The ease of interpretation is a direct consequence of using a single model to answer all relevant questions. For example, the same results reported above would become: "Overyielding was quantified as an overproduction of 112 g/m^2 of dry biomass (p-value < 0.01, CI {106, 118}); on average, each two-species composition yielded 23 g/m^2 more biomass than their corresponding monocultures (p-value $= 0.04$, CI {15, 31})..."

With our approach temporal stability is measured as the residual standard deviation

within each group. Standard deviations can be used to compute confidence intervals that help us understanding the variability of the observations within a given group. For example, assume that the full species mixture and all monocultures have the same productivity, say 100 g/m^2 and that the residual standard deviation in the full species mixture is $\hat{\sigma}_{mix} = 20$, while the residual standard deviation in the monocultures is $\hat{\sigma}_{mono} = 40$. By using simple calculations we can affirm that 95% of the observations in the mixture lie, roughly speaking, within -40 and +40 g/m^2 (i.e. $\{-1.96*20, +1.96*20\}$) around the group mean value. Thus, the vast majority of the observations are found to be between 60 and 140 g/m^2. These results can be compared with the stability of the monocultures where 95% of the observations lie within 20 and 180 g/m^2. We can thus report this results as "the variability in monocultures is double to that one observed in the full species mixture (resp. $\hat{\sigma}_{mono} = 40$ and $\hat{\sigma}_{mix} = 20$; p-value < 0.001)".

The standard deviation is on the same scale as the response variable and is indeed what we see when plotting the data (Sarkar, 2008; Venables and Ripley, 2013). For example, take two groups (A and B) that have different variabilities. Group A has a standard deviation of 1 (thus a variance of 1), group B has a standard deviation of 2 (thus a variance of 4). When displayed graphically, the variability in group B is perceived to be twice as large as the variability in group A, not four times larger. This confirms that variability seen on graphs is directly proportional to standard deviations (Limpert and Stahel, 2011; Sarkar, 2008; Venables and Ripley, 2013).

Therefore, the results of stability measured with standard deviations match with their graphical display. In addition, the estimated standard deviations can be compared with any other value that lays on the same scale (i.e. on the scale of the response variable). We can thus compare overyielding and stability. For example, take a situation where the monocultures produce on average 100 g/m^2, overyielding is 112 g/m^2, and the residual standard deviations are those of the previous example ($\hat{\sigma}_{mono} = 40$ and $\hat{\sigma}_{mix} = 20$). This can be displayed graphically as follows.

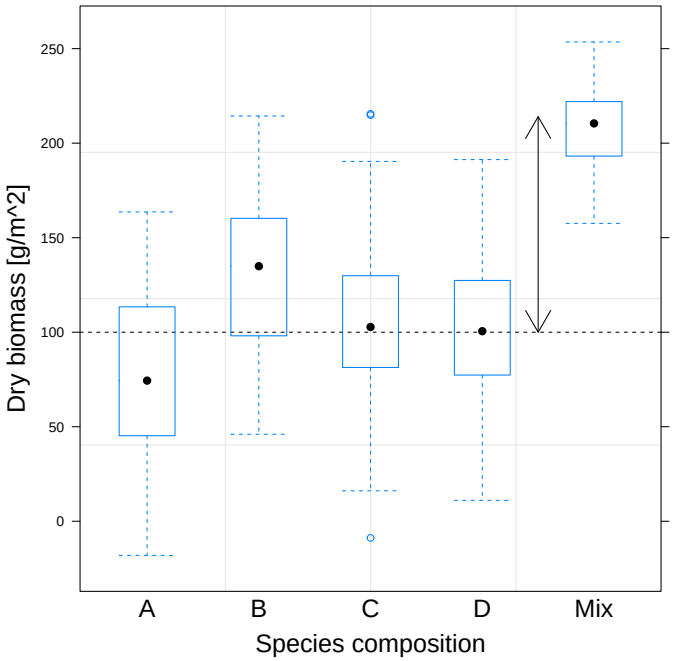


Figure 3.4: Productivity in different species compositions. The horizontal dashed line represents the average of the four monocultures. The vertical arrow quantifies overyielding.

Figure 3.4 shows that there is a direct relationship between what our method estimates and what we actually see when plotting data. Overyielding is estimated to be 112 g/m^2, and this is highlighted by the black arrow between monocultures and the mix (the dashed horizontal line represents the mean monoculture productivity)[2]. The height of the boxes quantifies the variability within each group. Here, stability in the mixture was simulated to be twice as high as in monocultures.

Finally, our method quantifies the evidence for asynchrony by formally testing the strength of the interaction term (e.g. species:time). Our method does not provide a single condensed biological value, as this can be misleading. Very different biological situations can yield the same average pairwise correlation (see Appendix B). Our method, on the other hand, provides tailored responses to specific biological questions. In particular, our methods allows highlighting i) pairwise correlations, ii) finding homogeneous groups or iii) decompose asynchrony in its constitutive elements (see Appendix

[2]Note that for simplicity this overyielding measured here assumes equal abundance in the mixture for all four species. Our approach does not rely on this assumption.

B).

Here we could report our results as: "There was strong evidence that species reacted differently to annual fluctuations (p-value < 0.001)." Then depending on the specific study question addressed we could, continue our reporting of the results with "a subsequent clustering analysis revealed that species E and G group together in a separate cluster, which is clearly distinct from all other species." Alternatively, if we were interested in telling components of temporal asynchrony apart, we could say: "The magnitude of the shared annual fluctuations were found to be consistently larger than the species-specific year deviations ($\hat{\sigma}_{year} = 15$, $\hat{\sigma}_{year:species} = 5$)".

Fewer assumptions are made when estimating stability

The coefficient of variation and other stability metrics (e.g. 1/CV, CV2; Tilman et al. (1998)) make a strong assumption on how the variability of the data increases with the mean (i.e. the "mean-variance relationship"). In particular, CV assumes that the variability of the data (measured as its standard deviation) increases linearly with its mean. In Appendix C we show that there are good theoretical and practical reasons to make this assumption in many real situations. However, there are also several, frequently encountered, circumstances where this assumption breaks down.

In particular, assuming this mean-variance relationship for counts data, proportions or aggregated data (e.g. averages of several measurements) is inappropriate (it leads to an overestimation of the stability of ecosystems associated with large mean values). Our approach does not make any assumption on the mean-variance relationship, but rather estimates it from the data with a simple and intuitive method. Appendix C also highlights the fact that, in real settings, comparing ecosystems for their stability may be harder than commonly thought.

A clear statistical inference path

From the practical point view, is worth noting that our approach is well suited to perform statistical inference (i.e. provide p-values and confidence intervals). It is easy to support claims about differences in stability, for example, with an objective measure of evidence. For most values obtained with marginal metrics, such as the coefficient

of variation, it is not clear how statistical inference should be performed, and the often assumed normal distribution (with or without prior log-transformation) has been shown to be inadequate (Hendricks and Robey, 1936).

Greater flexibility

Our approach allows the user to carry out all secondary analyses with any type of data and not just with continuous response variables. For example, it is possible to use extensions of Linear Mixed Effects Models to analyse data such as counts or proportions. If the data follows a known distribution (e.g. "Poisson" for counts or "Binomial" for proportions), we can carry out the whole analysing using Generalised Linear Mixed Effects Models (Faraway, 2005). If the data do not fit in any of these distributions, we can use other modern methods such as bootstrapping techniques (see Appendix B).

Linking analyses on different response variables

The possibility to use the same approach on very different types of data enables us to compare results across studies that analyse very different ecosystem functions. For example, in Appendix D, we compared the results of an analysis looking at two different response variables, tree diameter and tree survival. It is well known that growth and mortality trade off (fast growing species are expected to show higher mortality, Tuck et al. (2016)). We may want to know whether environmental heterogeneity affects these two response variables. In other words, is a good location to grow also a good location to survive, or is it the trade-off present at spatial level as well? Appendix D addresses several such questions that involve comparing different ecosystem functions.

Most design requirements are dropped

Many design requirements made by marginal metrics can be dropped. For example, when measuring temporal asynchrony, secondary metrics implicitly assume that all species have been measured at the same time points. This may be unfeasible is large-scale or multi-site studies where a complete census may take a long time to be performed. It is fundamental to drop these design requirements in order to address

new questions with the new designs.

3.2.5 Conclusions

The approach presented in this publication enables the user to combine the primary and all secondary analyses into a single unifying framework, where a formal connection between them is present, such that information is shared among all pieces of the analysis. This approach is fundamentally based on the tailored use of modern Linear Mixed Effects Models and their extensions. These methods, which are often used in the primary analysis of diversity-function studies, are well-suited to cope with the complexity of these present-day experiments.

The greatest practical advantage of this approach is an easier biological interpretation of the results. This is due to the fact that all results are on the same scale as the response variable and are easy to visualise. In addition, our approach drops many of the unrealistic assumptions made by most secondary metrics. This method allows to analyse any type of data within the same framework. This implies that the results obtained when analysing different ecosystem functions can be compared among and across studies.

We argue that by using the model-based approach presented here, biologists analysing biodiversity function experiments, will extract better quality and more biological information from their studies. The use of this approach, based on a mainstream statistical technique, has the potential to greatly improve knowledge sharing and communication with other scientific fields. Many other biological fields of research, not just biodiversity function studies, can benefit from using this modern approach.

3.3 Appendices

Appendix A, starting at page 117, introduces the graphical methods used to visualise i) the raw data and ii) to biologically interpret the results of the modelling phase. Appendix B, starting at page 149, shows how asynchrony can be estimated using our approach. Appendix C, starting at page 191, discusses how stability is currently quantified and presents our approach. Finally, Appendix D, starting at page 221, shows how the results of analyses run on different response variables that have been measured in the same study can be interlinked to gather new biological information.

Appendix A: Graphical visualisations

Chapter 3

Contents

1 Introduction

The aim of this appendix is to highlight the importance of an appropriate graphical visualisation for modern ecological experiments such as diversity-functioning studies. Indeed, because of the complexity of their design and of the questions addressed, modern ecological experiments are best visualised with flexible tools that enable us to better understand our data and, as a consequence, to better understand the ecological processes underlying the observed patterns.

This appendix is divided into four sections. After this brief introduction, Section 2 aims to show the power of modern graphical packages. Section 3 introduces an important concept fundamental to the approach presented in this publication. In particular, this section highlights the differences between a marginal and a conditional approach. Finally, Section 4 discusses general aspects of the data visualisation for present ecological experiments.

2 Graphical analysis

This section represents the first step of any statistical analysis: The graphical visualisation of the data. The model that will be fitted to the data is a Linear Mixed Effects Model. We therefore display every potential predictor against the response variable. The examples presented here are based on data sets from two previously published studies. We will use the graphical **R** add-on package *lattice* (Sarkar, 2016).

The "Czech Republic" study
This study was carried out in a experimental forest in the Czech Republic where the species diversity of forest stands was manipulated by assembling trees of different species. In particular, four tree species (*Fagus sylvatica*, *Larix decidua*, *Picea abies* and *Quercus petraea*) were grown together in all possible combinations. There are 15 combinations of these species: Four monocultures, six two-species compositions, four three-species compositions and one full mixture of four species. Each combination was replicated three times, yielding 45 different experimental sites. In each site, six trees of each species were measured for radial growth (557 trees in total). Two cores where taken from each tree to measure annual growth (those were then averaged). For the sake of simplicity we analyse here growth in the period 1991-2000. Further details can be found in Chamagne et al. (2016).

The aim of this analysis is to understand the effects of diversity and species identity on radial growth. The latter was quantified as the width of annual growth rings. In addition, we want to estimate temporal and spatial asynchrony between species to better understand what could modulate the potential effects of diversity (see Appendix B). This experiment was run in semi-natural conditions, where planted trees were left to grow for decades with little human intervention. Therefore, this study can be considered to be half way between experimental and observational (Chamagne et al., 2016). It is thus important to consider possible sources of variation. For that reason, our model will also include stand density, tree age, tree basal area, and other variables. Below, each predictor is visualised against the response variable and discussed.

2.1 Visualising the response variable against the predictors

In this subsection we plot the log-transformed response variable against each potential predictor. This is radial growth which was measured as the yearly ring width. Taking the logarithm of response variables that can take only positive values is standard and recommended practice (Venables and Ripley, 2013; Sarkar, 2008; Limpert and Stahel, 2011). We start with continuous explanatory variables, and then look at categorical predictors[1].

Species diversity

The first explanatory variable presented here is `diversity`. This predictor is a continuous variable measured as the exponential of the Shannon Index. This biodiversity index was measured at site level (there are 45 sites; i.e. three replicates for each of the 15 species compositions). To start with, we use the basic plotting functions available in **R** (R Core Team, 2015). We use a scatterplot to visualise the marginal relationship between the two variables. On top of the scatterplot we add a regression line to better characterise their relationship.

```
plot(log10(ring) ~ diversity, data = d.20)
abline(lm(log10(ring) ~ diversity, data = d.20))
```

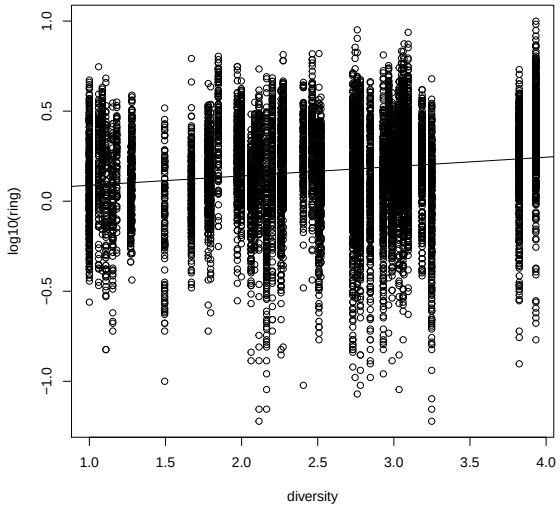


Figure 1: Response variable plotted against stand diversity. Basic R functions are used.

Although somewhat informative, this graph can largely be improved. As an example, we may want check whether different species react differently to the effect of stand diversity or tackle the problem of overplotting (i.e. overlapping observations). Improving this graph is feasible with basic **R** functions, however, this process would be very long and laborious. In the next page we present a similar graph obtained with the `xyplot()` function from *lattice* package.

[1]The order in which we present predictors is arbitrary and solely based on the capabilities of *lattice* that we want to present.

As said above, we may want to inspect the relationship between diversity and annual growth independently for each species. This is achieved with the pipe (i.e. |) symbol shown at line `(a)` of the function call below. This notation specifies that a plot must be created for each level of the categorical variable `species` (i.e. the variable that codes species identity). Note that apart from few exceptions, the syntax of the `xyplot()` function is very similar to basic plotting functions. With the argument `type` at code line `(b)` we specify what has to be drawn. In this case we specify points `"p"` and a regression line `"r"`.

```
require(lattice)
xyplot(log10(ring) ~ diversity | species, ## (a)
       type = c("p", "r"), ## (b)
       data = d.20)
```

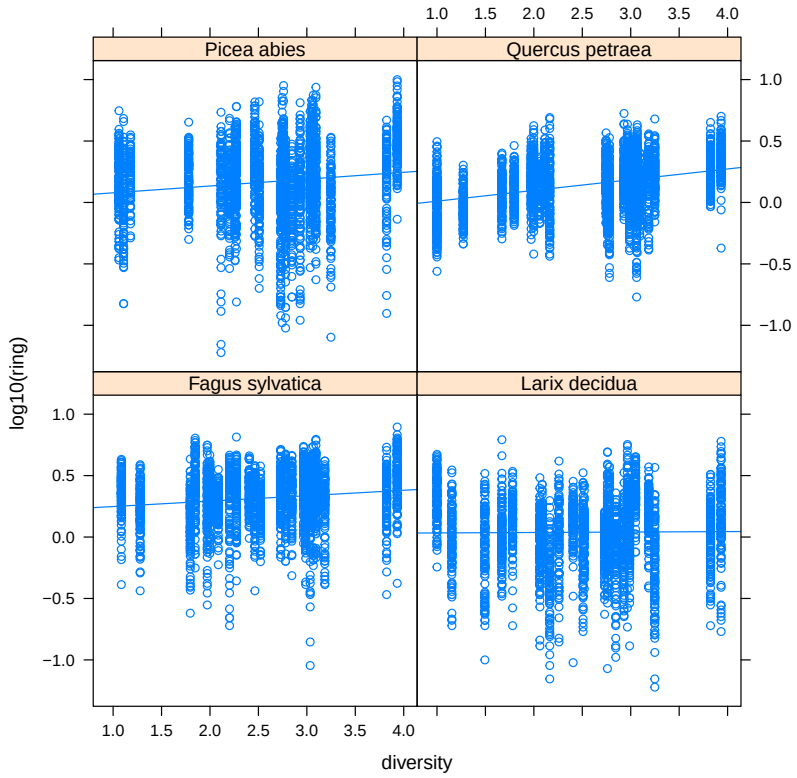


Figure 2: Response variable plotted against species diversity. The function xyplot() is used.

The relationship between response variable and predictor is now presented in four different panels. With this visualisation, we can inspect whether species respond differently to stand diversity. It really seems that species do differ in this respect. However, if the aim is to compare regression lines across panels, we make use of additional tools. In particular, in the next plot we will add a grid (`type = "g"`; line `(a)`) in the background to make comparisons across plots easier. We also change line colour `(b)` to make them stand out more.

Another important aspect to highlight here is that we are about to fit a Linear Mixed Effects Model. As their names says, Linear Models are linear in their coefficients,

nevertheless they can handle non-linear relationships. As an example, we may assume that the effect of diversity is quadratic and thus we could add a squared term to our model. However, in this first phase we are inspecting data, therefore it important avoid making too many assumptions. The obvious choice to characterise the relationship between two variables without making any assumption on its form is to use smoothers. Adding smoothers `(a)` to the plots instead of regression lines will enable us to evaluate whether the relationship can be considered to be linear or not.

```
xyplot(log10(ring) ~ diversity | species, data = d.20,
       type = c("p", "smooth", "g"), ## (a)
       col.line = "black") ## (b)
```

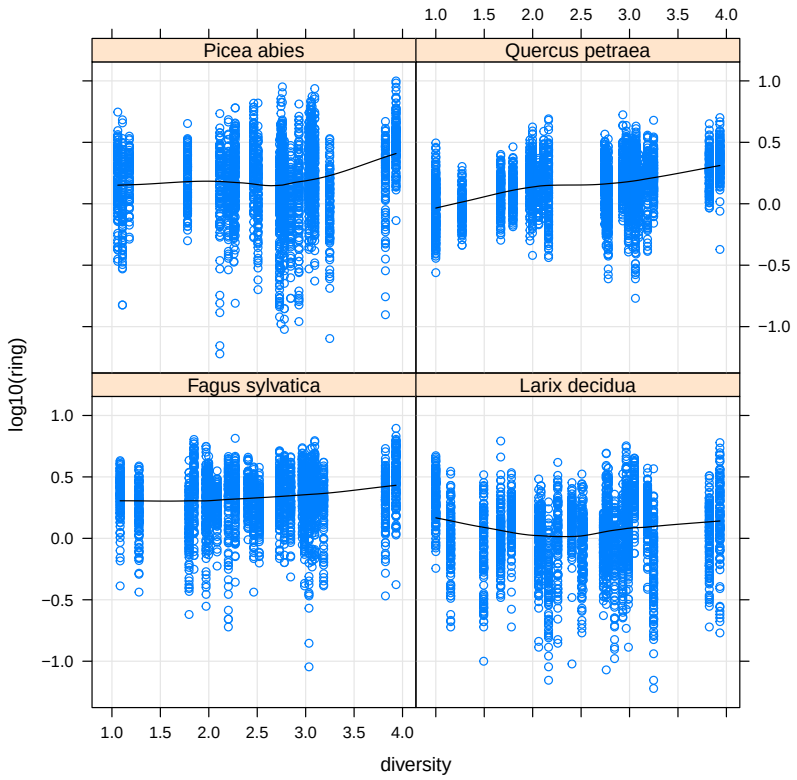


Figure 3: Response variable plotted against species diversity. Non-default arguments of xyplot() are used.

The smoothers shown here are not straight lines, suggesting that the effect of diversity may be non-linear. However, it is important to note that these are marginal plots where we control only for species identity. In other words, the apparently non-linear effect of diversity for the European larch (bottom right panel) may be due to the confounding effect of stand density. Ideally, we would like to add the information about all other predictors on this graph. Unfortunately, this is not possible as the graph would become overly complicated and unreadable.

Fitting Linear Models enables us to estimate conditional relationships. In other words, the can estimate the effect of, for example, species diversity while controlling for all the other predictors. In Appendix B we will show how to graphically display the conditional relationship between the response variable and the predictors for Linear

Models. This is done with the so called "partial residual plots" (Faraway, 2014).

Stand density

The next predictor inspected is `density`. This latter quantifies the density of trees within each site (i.e. each site). We first visualise its effects with the same settings used for the previous predictor.

```
xyplot(log10(ring) ~ density | species, data = d.20,
       type = c("p", "smooth", "g"),
       col.line = "black")
```

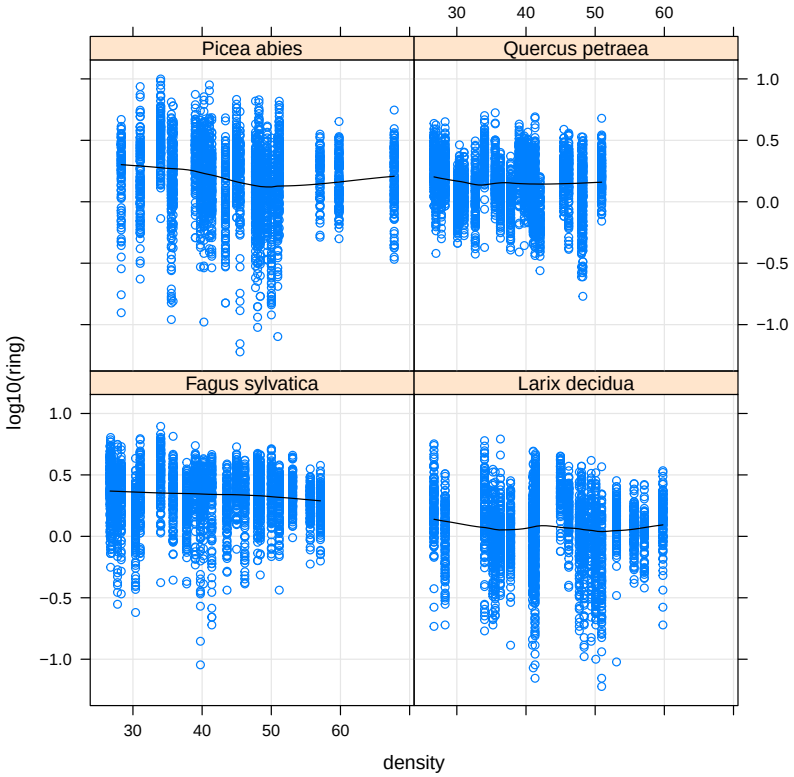


Figure 4: Response variable plotted against stand density.

The range of x-axes shows that the Norway spruce (top left panel) is present in denser areas. To make better use of the space we allow panels to have different x-ranges (`a`).

```
xyplot(log10(ring) ~ density | species, data = d.20,
       scales = list(x = list(relation = "free")), ## (a)
       type = c("p", "smooth", "g"),
       col.line = "black")
```

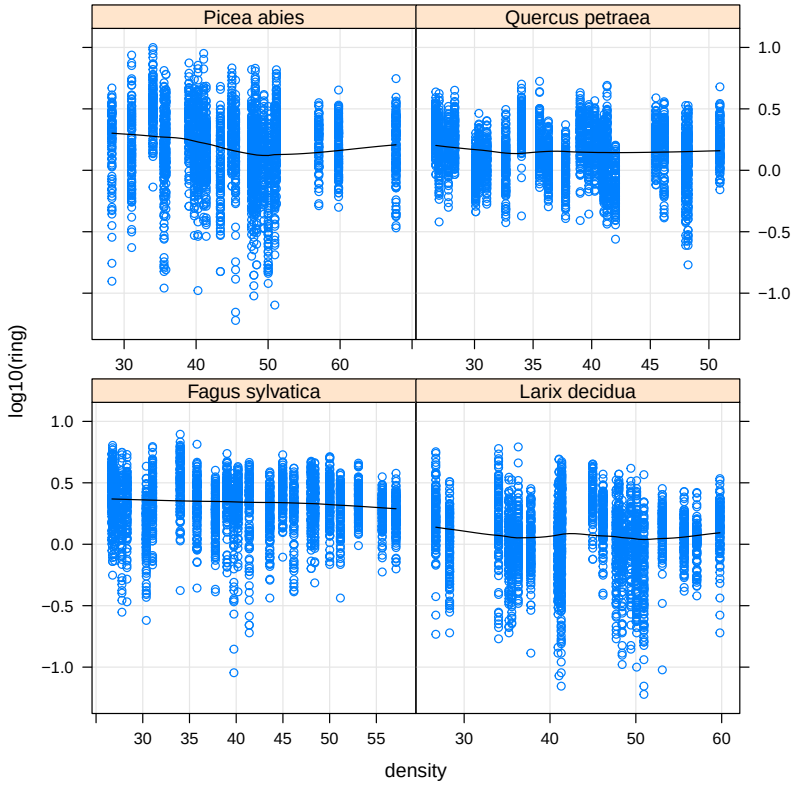


Figure 5: Response variable plotted against stand density. x-axes are allowed to differ among panels.

Not unexpectedly, density seems to have a (weak) negative effect on growth for all four species. Because we modified the x-axes we may also want to improve the appearance of the y-axes. The response variable has been log-transformed here. We may want to use a scale that enable us to understand the actual value of ring width without the need to back-transform values. To do that we can use the package *latticeExtra* (Sarkar and Andrews, 2016). This is achieved providing the untransformed response variable `(a)` and by specifying the scales need for the y-axes `(b)`.

```
require(latticeExtra)
xyplot(ring ~ density | species, data = d.20,
       scales = list(y = list(log = 10)), ## (a)
       yscale.components = yscale.components.log10ticks, ## (b)
       type = c("p", "g", "smooth"),
       col.line = "black")
```

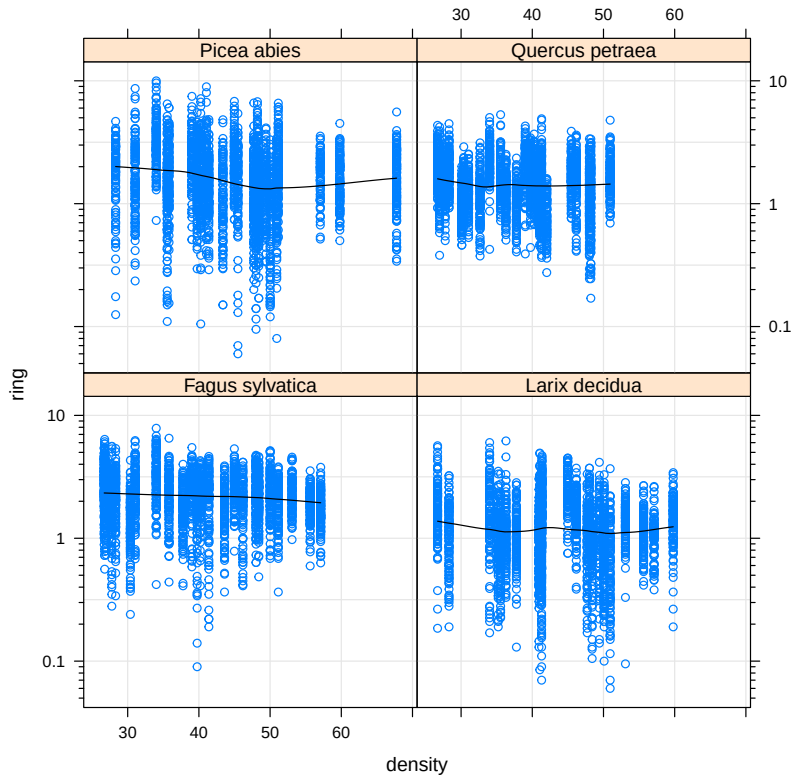


Figure 6: Response variable against stand density. y-axes use logarithmic scales.

From this plot, we can easily grasp the magnitude of ring widths, as the y-axes are in the original scale (i.e. millimetres)[2].

Basal area

Basal area is measured for each tree and is derived from the diameter measured at breast height. This predictor is expressed in centimetres squared. By looking at examples introduced so far it is clear that producing the best graph is an iterative process, where several options are changed at each step, and where we may even decide to revert back to original settings (as we did for the x-axes when looking at the effect of stand density). We may end up with a seemingly complex call by starting from a simple call where each piece is modified at each step (Sarkar, 2008). Rewriting a long function call can be annoying when we want to make a single small change. To overcome this issue modern graphical packages allow to store a function call into an **R**-object ready to be reused. This procedure is illustrated below with the predictor basal area.

```
xy.basal.area <- xyplot(log10(ring) ~ basal.area | species, data = d.20,
                        type = c("p", "g", "smooth"),
                        col.line = "black")
print(xy.basal.area)
```

[2]Note that in this plot we reverted back to use the same ranges for all x-axes as the difference in densities is an important pattern that we must keep in mind when interpreting the results of the statistical analysis.

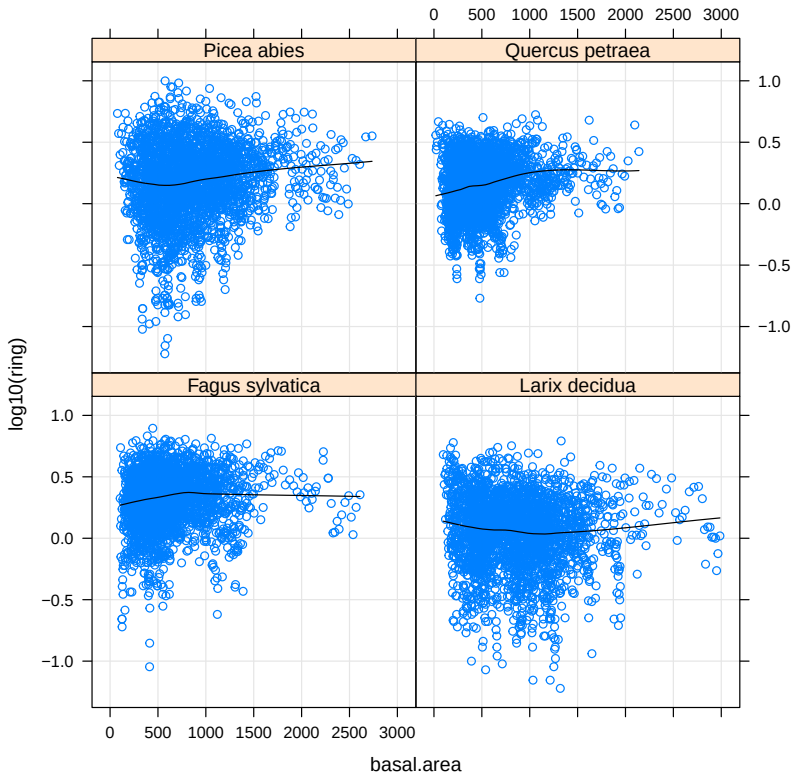


Figure 7: Response variable plotted against basal area.

When the function call is stored in an object the corresponding graph is not produced by default. In order to do that, we use the generic function `print()`. Figure 7 highlights a problem often encountered with large data sets: Overplotting. Many observations overlap making it difficult to draw strong conclusions on the relationship between these two variables. To alleviate this problem we can decrease symbol size. To do that, we can use the function call stored previously and adapt it via the generic function `update()`.

```
update(xy.basal.area, cex = 0.1)
```

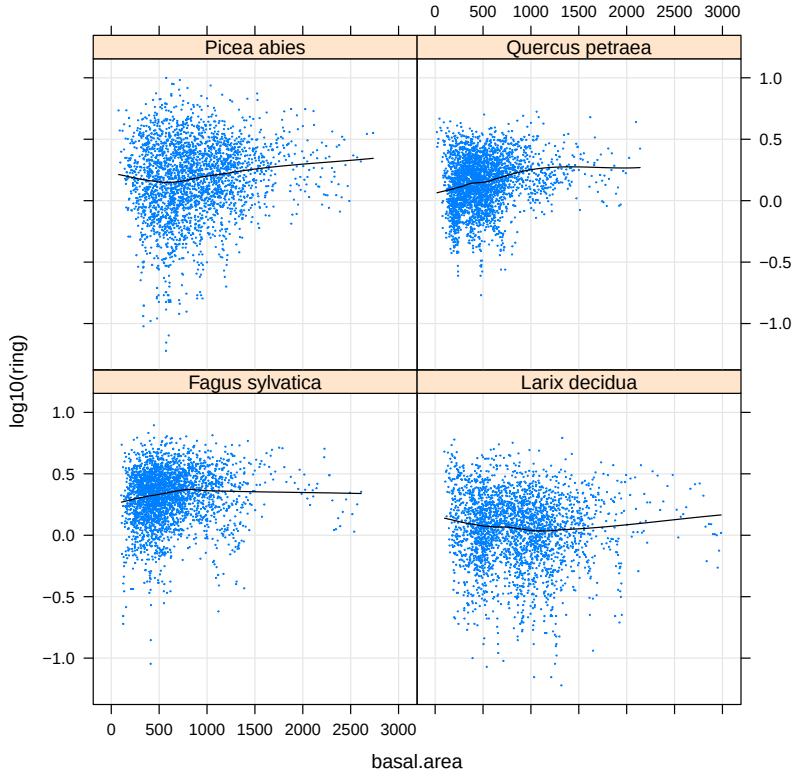


Figure 8: Response variable plotted against basal area. Symbols are drawn as small circles.

The use of `update()` is very handy as we can quickly modify an existing graph. This is another important advantage over basic plotting functions. Although the issue of over-plotting is less severe, it is still present. At page 12 we will see how to use transparency to better deal with it.

Trees age

The age of the trees is included in the model as it has been shown to influence radial growth (Chamagne et al., 2016). Older trees are expected to grow slower.

```
xy.age <- xyplot(log10(ring) ~ age | species, data = d.20,
                 type = c("p", "g", "smooth"),
                 col.line = "black",
                 cex = 0.1)
print(xy.age)
```

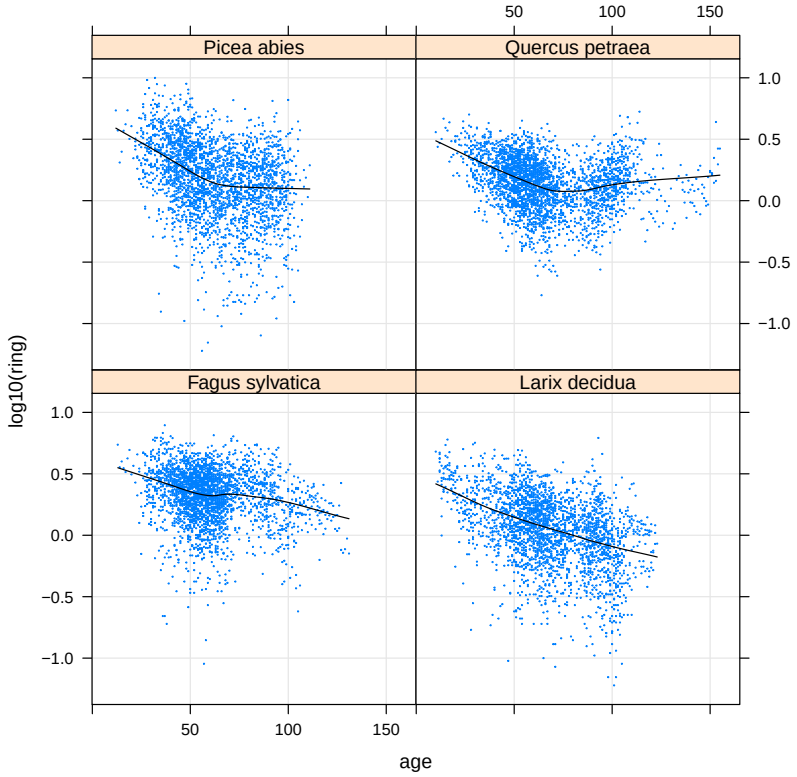


Figure 9: Response variable plotted against tree age.

As expected, there seems to be a negative relationship between radial growth and tree age. However, we observe a unexpected pattern on the panel showing the sessile oak (top right panel). Growth for trees older than 100 years does not seem to decrease, on the contrary growth seems to increase again. To better understand this pattern, we make use of another very powerful graphical technique "superposition" (or grouping). This consists of rendering observations associated with different levels of a grouping variable (e.g. `tree`), with different graphical characteristics. In particular, we connect the observations belonging to the same tree `(a)` with a coloured line `(b)`. We use the argument `subset` to specify that we want to work only with sessile oak `(c)`.

```
xy.age.2 <- xyplot(log10(ring) ~ age | species, data = d.20,
                   groups = tree, ## (a)
                   type = c("l", "g"), ## (b)
                   subset = species == "Quercus petraea") ## (c)
print(xy.age.2)
```

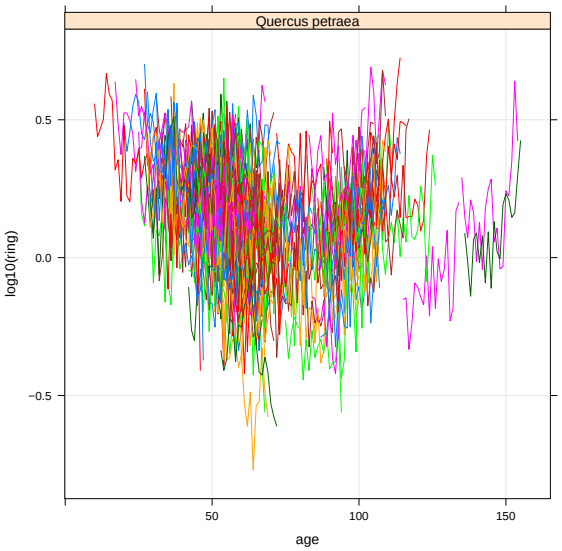


Figure 10: Response variable plotted against tree age for sessile oak. Trees are highlighted with colours.

With this visualisation it is now clearer that the observed pattern is mainly due to a few old trees behaving in a unexpected way. This visualisation is not entirely satisfying yet, as there is a lot of overlapping between lines. The next graph better copes with this issue.

Date (i.e. year of the ring)

The next predictor inspected is `date`. In this context this is a very important predictor as it is used to estimate temporal asynchrony. This latter is a quantity that estimates how species react to time fluctuations. In Figure 10 we used grouping to connect the observations that belong to the same tree. When the number of levels of the categorical variable used to define groups is large as it is the case for trees, using colours may not be fully satisfactory.

This situation is often encountered in the framework of Linear Mixed Effects Models, where we want to highlight random effects (e.g. here trees), but still obtaining a readable graph. A possible solution is to use transparency. Instead of using colours, all groups (i.e. trees) are rendered in black with some transparency. Transparency can be set via the argument `alpha`.

```
xyplot(log10(ring) ~ date| species, data = d.20,
       alpha = 0.2,
       col.line = "black",
       groups = tree, type = c("l", "g"))
```

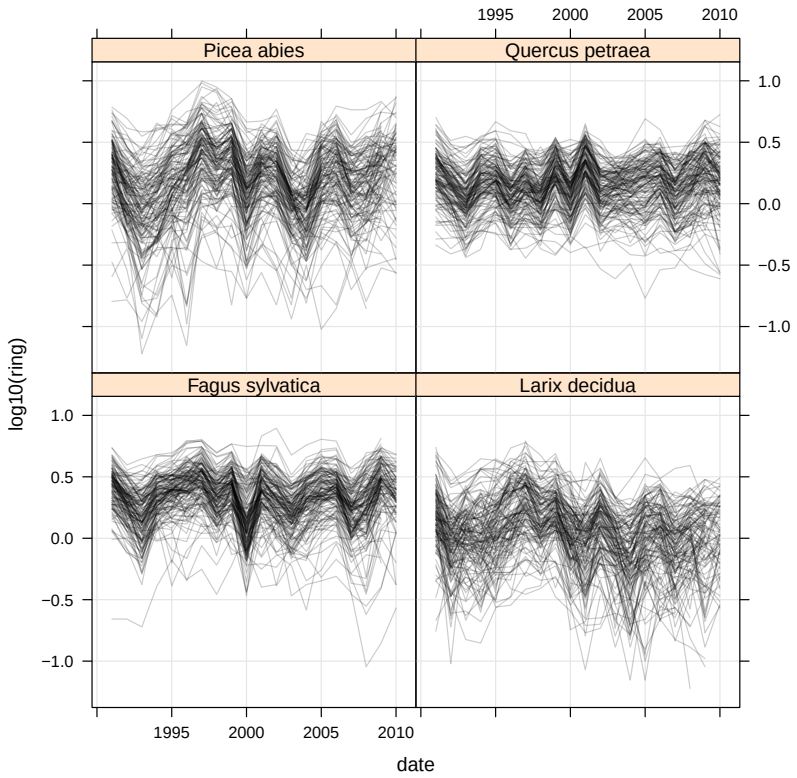


Figure 11: Response variable plotted against date. Trees are highlighted with grouping and transparency is used.

This graph is very helpful, due to the use of transparency, regions with many overlapping lines are rendered darker. This plot shows very clearly that date does not have a linear effect on growth (i.e. it is not just a trend). Indeed, it as a very discontinuous effect. When fitting a model to these data, we will use the maximum of the flexibility by including date as a categorical variable. This step is fundamental to be able to quantify temporal asynchrony. On top of that, we also note that species do not seem to react exactly in the same way to year fluctuations. As an example, year 2000 had lowest radial growth for European beech (bottom left panel), while this was not the case for the European larch (bottom right panel). This implies that an additive effect of years and species may not be enough to properly characterise this data. Very likely, we need to allow species to respond differently to years. Biologically, this is an indication of temporal asynchrony (i.e. species respond differently to annual fluctuations).

This graph can help determine the structure of the random effects of our model. Indeed, we can sense what trees behaviour looks like[3]. In particular, tree lines seem to run roughly parallel separated by a constant vertical shift. This implies that a model with a random intercept for tree may be adequate to model this data. Similar graphs are invaluable also when analysing grassland biodiversity studies(Tanadini and Hector, in prep.).

Species composition

[3] `tree` will be considered to be a random effect as we are not interested in estimating the effect of each tree, but we aim to take this design variable into account and to estimate their variability.

In the previous plots we inspected the effect of the continuous predictors available. We now turn our attention to categorical predictors. For the sake of brevity we only inspect the effect of species composition. This is an important variable in any biodiversity functioning study. We start by displaying this data with a simple boxplot.

```
bwplot(log10(ring) ~ composition, data = d.20)
```

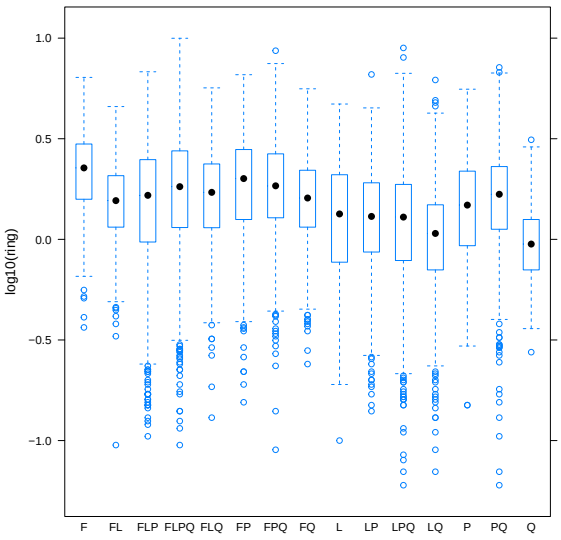


Figure 12: Boxplots of the reponse variable for species compositions.

Although informative this graph does not exploit its full potential. Indeed, the ordering of the groups is alphabetical[4]. However, very often reordering the groups labels is much more informative (Sarkar, 2008). In this context, we may want to order groups by the number of species they contain. This can be easily done with the function `reorder()`. In addition, we may want to separate species richness groups with vertical lines. To achieve that, we introduce here the use of panel functions. We don't discuss them in detail as it would be out of the scope of this publication. We can think of them as functions used to add layers to an existing (empty) graph. Here we first add a grid in the background `(a)` to better compare compositions, we then add vertical lines separating species richness groups `(b)` and we finally plot data `(c)`. Sarkar's lattice book devotes an entire chapter to panel functions (Sarkar, 2008).

```
d.20$composition.ord <- reorder(d.20$composition, d.20$sp.rich)
bwplot(log10(ring) ~ composition.ord, data = d.20,
       panel = function(x, y, ...) {
         panel.grid(...) ## (a)
         panel.abline(v = c(4.5, 10.5, 14.5), lty = 2, lwd = 0.5) ## (b)
         panel.bwplot(x, y,...) ## (c)
         })
```

[4]This is the default behaviour of **R**.

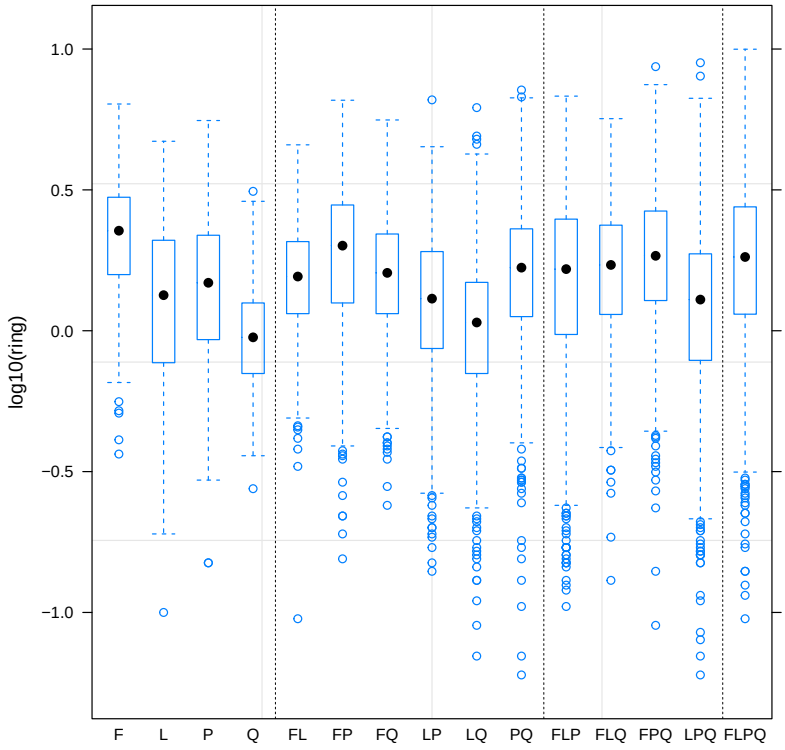


Figure 13: Boxplots of the reponse variable for species compositions. Groups are ordered according to species richness.

This graph conveys substantially more information than the previous one. It indicates that there may be differences between monocoltures (boxplots on the left). As an example the sessile oak monoculture (`"Q"`) appears to grow slower than the European beech (`"F"`). It is not clear whether the differences among non-monoculture compositions can be entirely explained by species identity or not. There seems to be little difference among species richness groups.

This information could hardly be extracted from Figure 12. It is interesting to note that with just a few changes to a graph we have been able to dramatically improve the amount of information that it conveys.

Visualising interactions between continuous predictors (**)

(**) Starred paragraphs, subsections and sections indicate an higher technical content or that these parts of the document are not fundamental to the approach presented here. These parts are presented for the sake of completeness. However, the reader interested in a quick read can skip them.

As stated above, a complete graphical analysis inspects the relationship between the response variable and all predictors. In addition to that, there is some degree of specificity to any analysis, this involves tailored graphs. For example, we may want to test the hypothesis that the effect of species diversity is actually modulated by forest density. We may hypothesise that the effect of diversity is stronger in low-density sites, while this latter is weak or absent in high-density sites. Below we represent one possible

way to inspect this hypothesis in a graphical way. Essentially, we plot the relationship between the response and diversity for different forest densities using panelling. We arbitrarily decided to divide forest density into four groups. To produce this plot we make use of the `equal.count()` function that discretises continuous variables into categorical variables. As the name of the function indicates, the groups obtained are of same size. Note that we specify the layout of the plot to better highlight the effect of increasing density (i.e. all four panels on one row).

```
xyplot(log10(ring) ~ diversity | equal.count(density, 4),
       data = d.20,
       type = c("p", "r"),
       cex = 0.1, col.symbol = "lightgray",
       layout = c(4, 1))
```

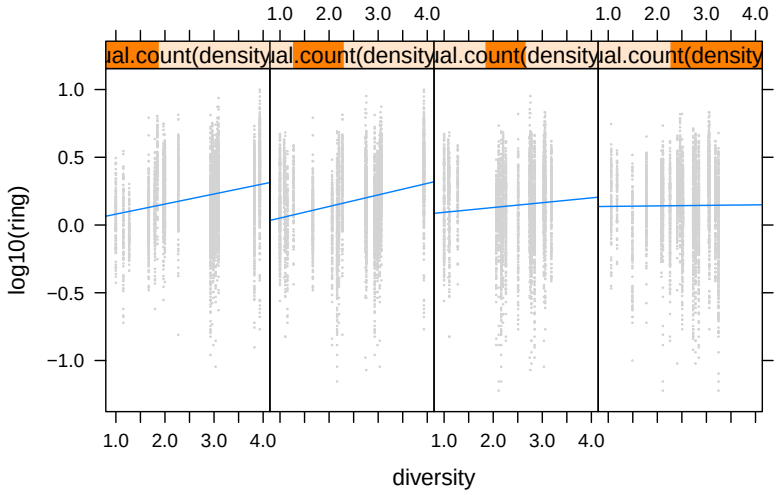


Figure 14: Effect of diversity on tree ring growth for four forest density classes.

Interestingly, it seems that the positive effect of diversity declines at higher densities (left to right). Nevertheless, density was not manipulated experimentally in this study and it is not the main focus here. Therefore, this will not be inspected further. Nevertheless, this graph shows how very specific hypotheses can be easily visualised with modern graphical packages.

2.2 Visualising other aspects of complex ecological experiments

In the previous subsection we have been visualising the marginal effect of each predictor on the response variable. This is fundamental to any analysis, however, there are other aspects of modern experiments that need to be visualised appropriately.

Study design (spatial and temporal structure)
Modern ecological studies often have designs that are not easily described in words.

Powerful graphs can be of great help when trying to understand how a study was actually carried out. A large amount of information can be visualised in a single meaningful graph. To demonstrate that, we introduce a new data set without describing its experimental design with text, but rather with a series of graphs (code not shown).

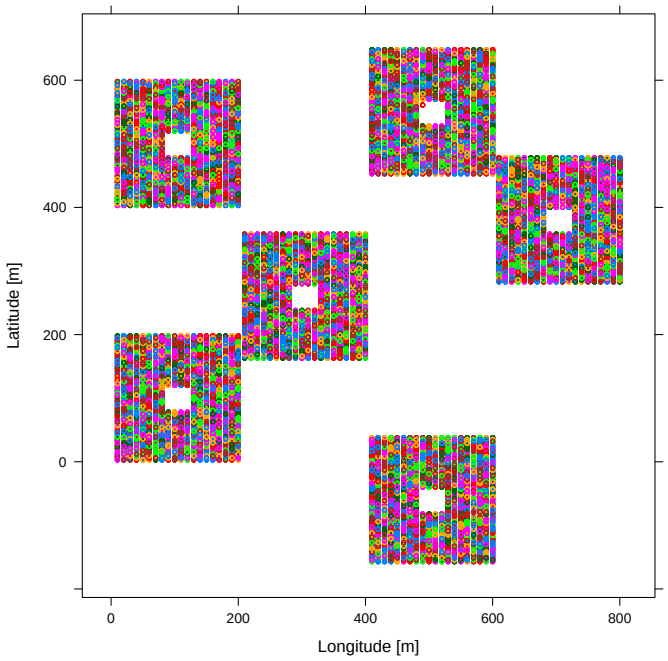


Figure 15: Study site, experimental design.

From Figure 15 we can get a rough idea of how the study was laid out. Based on this Figure is very easy to explain the actual design. This study is a subset of the Sabah Biodiversity Experiment, a large-scale diversity-function experiment run in Malaysia (Hector et al., 2011). In particular, here we look at the subset of plots that were planted with the full-species composition (i.e. 16 species) and that were surveyed more frequently, the so called 'Sabah Intensive Experiment'. As Figure 15 indicates there are six plots (i.e. the six squares). At their centre no seedlings were planted. This data was previously published by other authors (Philipson et al., 2014). We now focus on a single plot to better understand the details of this study.

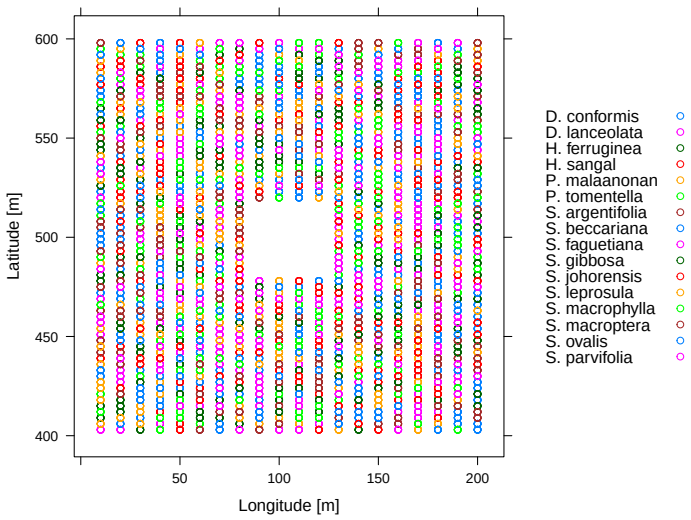


Figure 16: One particular plot of the Sabah Intensive Experiment inspected more closely.

As Figure 16 shows, in each of these plots, 20 (vertical) lines of seedlings were planted. Each line is supposed to contain 66 seedlings (this was not always feasible). So in total each plot contains about 1,200 seedlings. The colours used on Figure 15 represent the sixteen different species (see legend). The two last graphs illustrate the spatial structure of this experiment. It is important to note that these graphs have an isometric aspect ratio (Sarkar, 2008). This latter is the ratio of the physical height and width of a panel. In other words, the the x- and y-axes are forced to have the same scales, which is appropriate for situation where both variables have the same units (here metres)[5]. Going back to our study, there is also a time component to it, that we illustrate in the following graph.

[5]When using *lattice* we can use the isometric aspect ratio with `aspect = "iso"`.

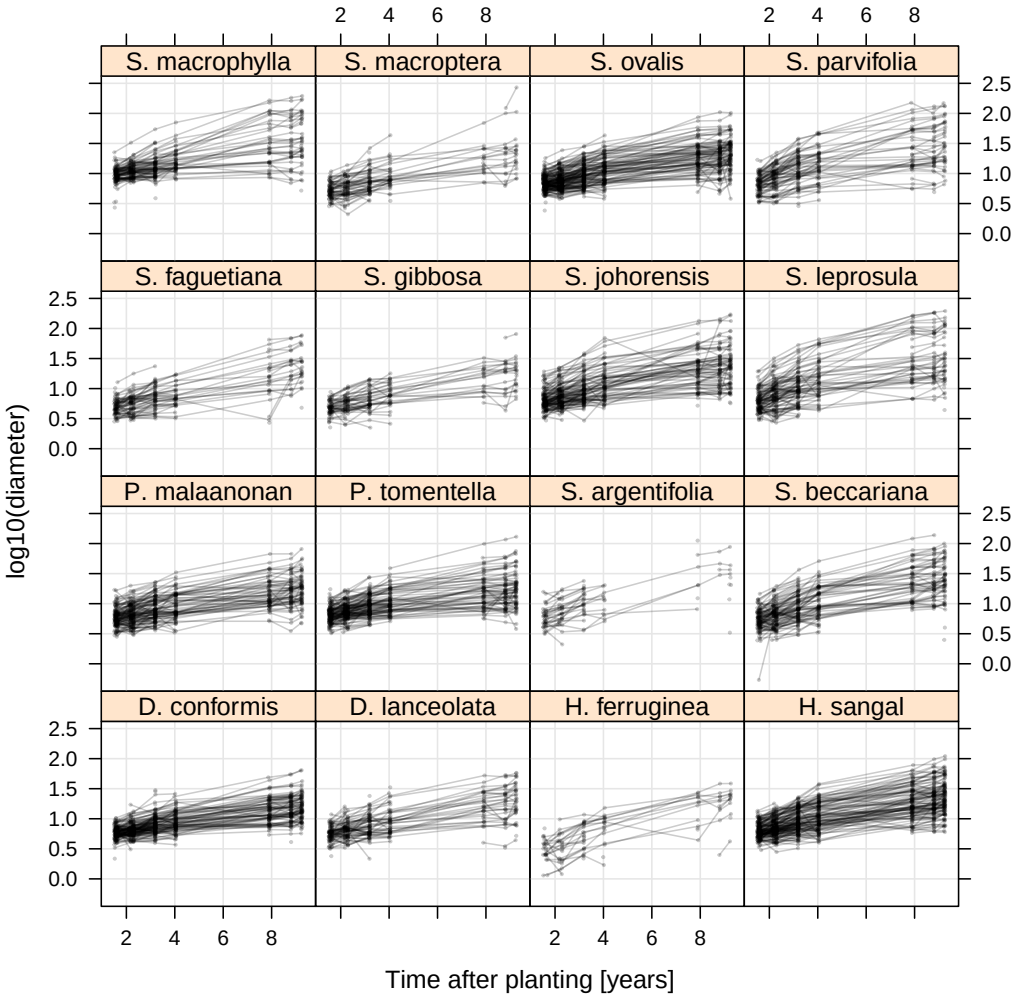


Figure 17: Temporal component of the Sabah Intensive Experiment.

Figure 17 explains the temporal structure of this study. Panels indicate clearly that 16 species were considered. The y-axis indicates that the response variable measured in this study was the diameter (measured at breast height), while the x-axis indicates that seven measurements were taken in a interval of about seven years after planting. In analogy to Figure 11, we used black transparent lines to highlight and group the observations that belong to one particular tree.

These last three graphs are very useful at illustrating a design that is relatively simple, but that would be much more difficult to fully describe in words. On top of the main information (such as six plots), there are a few subtleties that we can grasp from these graphs. As an example, it is trivial to understand that the seven measurements are not equidistant. By looking carefully, we can also understand that the measurements are not carried out at exactly the same seven time points. Indeed, the observations are not perfectly aligned on seven vertical lines, but rather group together into seven pretty compact groups. Given the large number of trees involved in this study, it was indeed impossible to carry out each survey in a single day. This detail is important from the modelling point, especially when looking at temporal asynchrony, as we will see in Appendix B. Finally, it is also clear that some species had more trees inspected

than others (e.g. *S. argentifolia* and *H. sangal*).

Comparing growth patterns

The original publication from this data set focused on comparing the growth rates of the 16 species (Philipson et al., 2014). The analysis was based on the raw measurements. Here, we we want to compare the growth rate of the different species and look for evidence of of spatial asynchrony. In other words, we are interested in modelling species growth and whether species-specific growth is influenced by environmental variability.

The analysis that we present here uses aggregated data. In particular raw measurements are aggregated at population level. Put simply, for each species we computed the mean (log-transformed) diameter at each time point in each plot. As we were interested in spatial asynchrony we computed these means for the six plots separately. Figure 18 below illustrate the aggregated data.

```
xy.growth.1 <- xyplot(m.grow ~ m.time.years | species.short, groups = plot,
                      data = d.sbe.int.agg,
                      xlab = "Mean time since planting [years]",
                      ylab = "Mean log-transformed diameter [cm]",
                      type = c("b","g"),
                      auto.key = list(space = "right", title = "plot",
                       cex.title = 1))
xy.growth.1
```

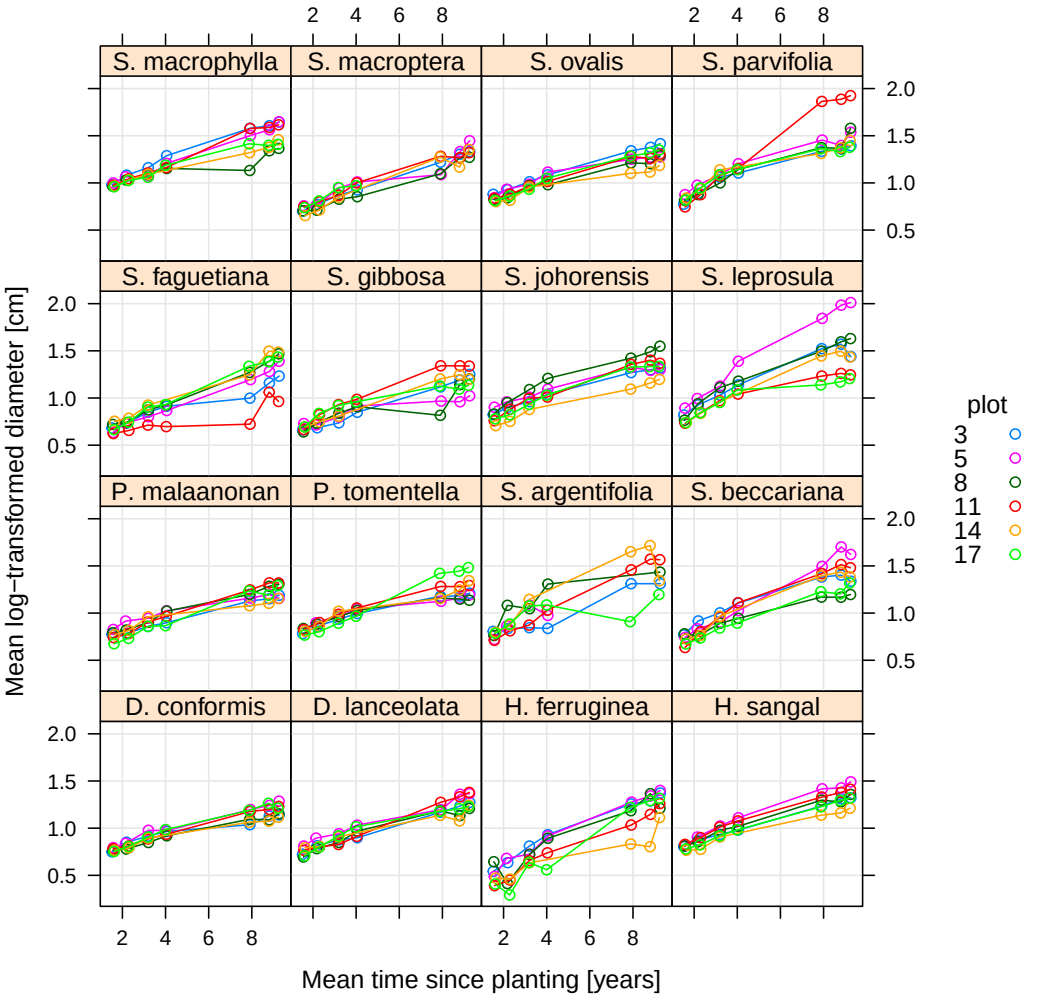


Figure 18: Growth of the 16 species in the Sabah Intensive Experiment. Each panel shows a species, each coloured line refers to a particular plot.

Each coloured line on this graph represents the (log-transformed) averaged diameter in one particular plot, for one particular species at a given time point. To make an example, the top right panel represent the species *S. parvifolia*, the red dots connected through a line in that panel correspond to the plot 11 (see legend). So each of the red dots in this panels is the averaged log-transformed diameters of all trees concerned. This graph is helpful to compare growths, however, it can be drastically improved. Indeed, if we were to put this graph into a publication it would be very difficult to understand the ordering of species in terms of growth. This information could be conveyed in the text, but this would be very inefficient. Here the fastest growing species is *H. ferruginea* (bottom row, third column), the second fastest is *S. leprosula* (last column, second row). So the panel ordering does not help, and in addition, this graph does not highlight differences in growth.

Usually, when graphs are created all the available space is automatically used. This is obviously the best option in a wide variety of situations. However, when comparing slopes (here growth rates) the use of alternative aspect ratios can be very beneficial (Sarkar, 2008). In the next Figure we modify the aspect ratio so that differences

among slopes are better spotted[6]. In addition to that, we order panels by increasing slope, instead of alphabetically.

```
xy.growth.2 <- update(xy.growth.1, aspect = "xy", ## other aspect ratio
                      index.cond = function(x,y) coef(lm(y~x))["x"],
                      ## change order of panels
                      strip = strip.custom(par.strip.text = list(cex = 0.5)))
                      ## to make panel names fit
xy.growth.2
```

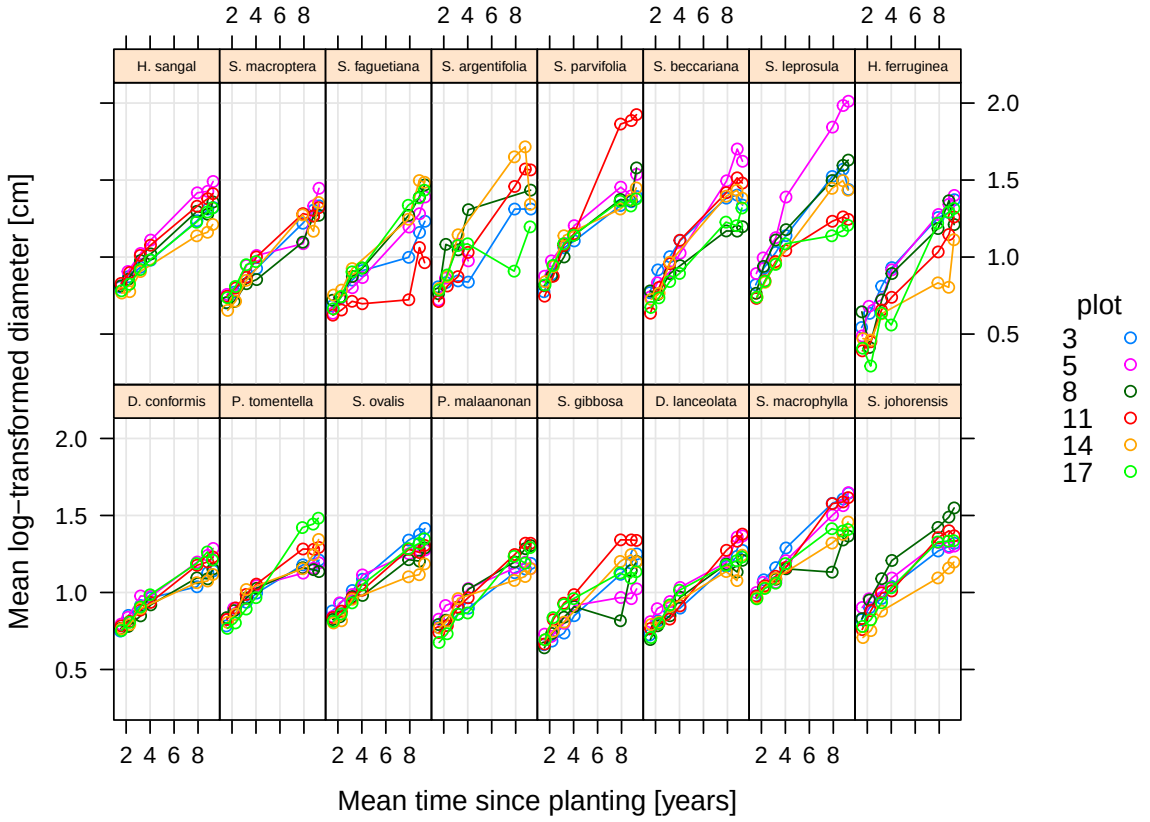


Figure 19: Growth of the 16 species in the Sabah Intensive Experiment. Each panel shows a species, each coloured line refers to a particular plot (see legend). Panels are ordered by increasing slopes (i.e. faster growth rates) from left to right and from bottom to top.

Figure 19 makes it much easier to discuss the differences in growth rates as panels are

[6]In order to do that, we produce a graph that uses the "45 degrees banking rule" (Sarkar, 2008).

ordered and the aspect ratio is adjusted. On top of that, there is new information. The upper right panel (i.e. *H. ferruginea*) has the greatest growth rate, but at the same time has a very small size (i.e. has by far the lowest intercept). As a matter of fact, this species was identified to have an unusual growth and survival pattern by a deeper analysis (Tuck et al., 2016). This pattern was not visible in Figure 18.

The Sabah Intensive Experiment also recorded the survival of each tree at each survey. We can use a similar plot to display mean survival in plots. We ordered panels by mean survival (code not shown).

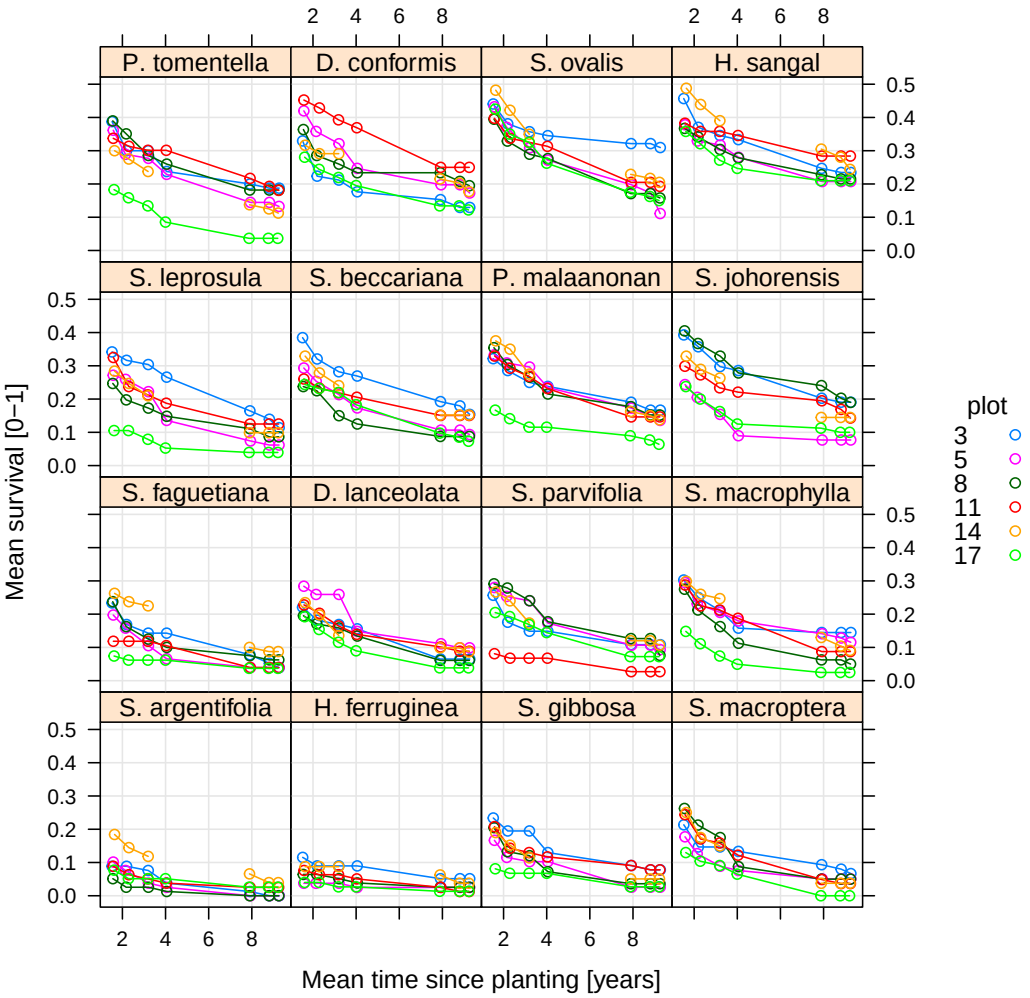


Figure 20: Survival of the 16 species in the Sabah Intensive Experiment. Each panel shows a species, each coloured line refers to a particular plot (see legend).

Figure 20 shows that are large differences in survival among species and that after about nine years since planting less the 30% of the trees were still alive. There are several other interesting aspects that this graph highlights. As an example, we note that plot 14 (yellow) was not surveyed at all time points (broken line). This information can be relevant when modelling the data. On top of this "technical" detail, this graph is very efficient at putting in evidence important biological information. Indeed, if we look at the effect that plots have on species survival, we note clear differences. Plot

17 seems to have overall the worse condition for survival, however, the magnitude of this effect is varying and not consistent among species. For species *S. johorensis* and *S. parvifolia* this is not the worse plot. This information is biologically very relevant as it indicates that spatial asynchrony is very likely to be present.

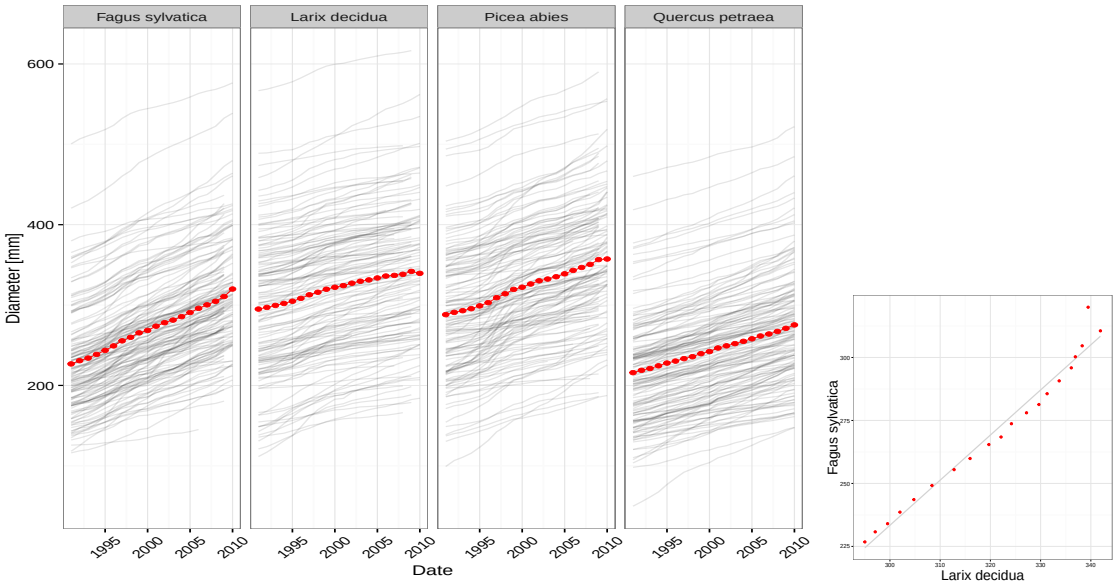
Temporal asynchrony and conditional effects
Given that there is a strong indication for spatial asynchrony, we may want to understand whether temporal asynchrony is present. In other words, whether species are affected differently by time (climatic fluctuations). Classical metrics to quantify asynchrony such as that developed by Gross and colleagues (2014), implicitly assumes that "species biomasses are fluctuating around a stable mean". This is obviously untrue here. Indeed, both traits diameter and survival are not constant over time. So by analysing these traits with these classical metrics we would conclude that all species are strongly synchronous. This conclusion would be wrong and misleading. This artefact is simply due to the fact that the trait analysed is not stable over time.

To improve this situation we may want to remove the trend (e.g. growth) before quantifying temporal asynchrony. This way to proceed would remove the biasing effect of growth on temporal asynchrony. Nevertheless, we may also want to remove the obscuring effect of other variables that add noise to the biological process of interest without biasing it. The next section explains this concept in depth with a real example.

3 Estimating conditional effects

As discussed above, we may want to estimate temporal asynchrony when other obscuring or biasing effects are removed from our data. This way to estimate effects is referred as conditional. Indeed, here we would obtain the effect of temporal asynchrony when we control or condition for the effect of growth and other nuisance factors.

To better illustrate this concept we use again the Czech Republic data set inspected in Section 2. In particular we look at tree diameter over time to quantify temporal asynchrony. To start with, we display the diameter of trees against time. Each species is plotted in a separate panel. The observations of trees are connected by a grey line. On top of the raw diameters, we also display the average annual diameter for each species in red. This quantity is important as it is used to quantify temporal asynchrony. In this section the add-on package *ggplot2* is used to produce figures (Wickham and Chang, 2016).



(a) Tree diameter plotted against date

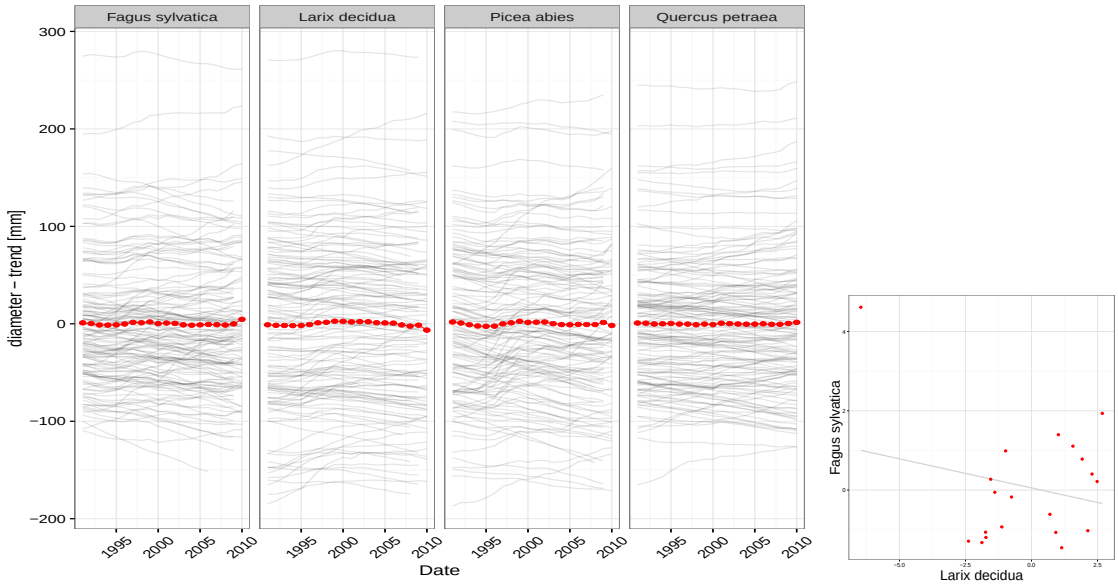
(b) Scatterplot of the mean annual diameter for two species

Figure 21: Raw data used to quantify temporal asynchrony.

Many of the currently used temporal asynchrony measures are derived from the temporal correlation of pairs of species. These pairwise correlations are often averaged to obtain a global measure of asynchrony (Gross et al., 2014). In order to keep our reasoning straightforward we start focusing on the relationship of two species (i.e. European beech and European larch which are shown in the two left panels).

If we were to correlate the raw diameters of the two species, we would get a very high correlation. Indeed, the Pearson correlation coefficient estimated is 0.99. However, this correlation does not represent the common response that these two species have towards years (i.e. temporal asynchrony). This highly positive correlation is mainly driven by the fact that trees are growing. In other words, we are not correlating temporal fluctuations, but tree growth. Applying currently used asynchrony measures to raw data like diameter is obviously of little help.

In order to truly correlate year fluctuations we must remove the effect of tree growth. To do that we remove the trend in each species. For the sake of brevity, the exact procedure of these steps are not shown here. Appendix B describes all steps needed to obtain conditional, and thus meaningful, estimates of temporal asynchrony.



(a) Tree diameter plotted against date when trend is removed

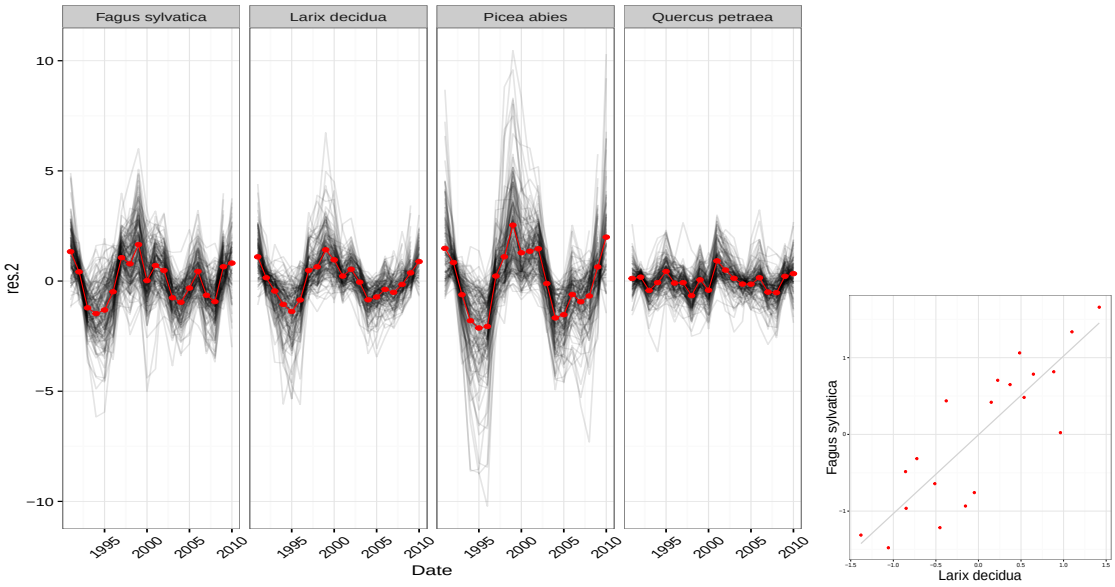
(b) Scatterplot of the mean annual diameter for two species

Figure 22: Year effects when trend is removed.

Once that the growth trend has been removed we can truly look at how species react to annual fluctuations. Unfortunately, there is a large scatter due to tree effects that obscures the interesting pattern that we are looking for. The situation depicted here is similar to the one found in medium-long term grassland studies where the biomass is stable over time. Trees can be seen as plots in this context (i.e. they can be rendered with grouping techniques).

The graph on the right (b) shows the correlation of the two species when the growth trend is removed. The estimated Pearson correlation is negative (-0.22). It is important to note that this correlation is strongly affected by a single observation (i.e. year 2010)[7]. To better compare year fluctuations, we may want to remove the confounding effects of trees. Analogously, in a grassland study we may want to remove the obscuring effect of plots, blocks or sites.

[7]It can be argued that a plain Pearson correlation is not a suitable choice in this situation. When the number of species involved is large, it can be difficult to spot this kind of anomalies. In Appendix D we will show how to deal with these situations.



(a) Tree diameter plotted against date when trend and tree effects are removed

(b) Scatterplot of the mean annual diameter for two species

Figure 23: Year effects when trend and tree effects are removed.

Figure 23 shows the year effects once the obscuring effects of growth (i.e. trend) and trees have been removed. The year fluctuations are now very clear. Note that the range of the y-axis has decreased from about 500 mm (see Figure 21) to just 20 mm. In other words, we are now looking at the interesting pattern, temporal fluctuations, while controlling for other nuisance variables.

The correlation between European beech and European larch is strong and positive (0.85). However, when measuring temporal asynchrony we are interested in the fluctuations around a stable mean. It is clear that we cannot consider the mean to be stable here. As an example all species were negatively affected by year 1994. While the year around 2000 seems to be beneficial for all of them. In other words, some years appear to have a positive effect on all species, while others have a negative effect. It is legitimate to be interested in the deviations from the main year to understand how species react to year when the "main year effect" (i.e. the effect shared by all species is removed).

The supporting information dedicated to asynchronies, Appendix B, shows that after the "main year effect" is removed, European beech and European larch are actually strongly asynchronous (i.e. negatively correlated). This does not come unexpected as it is common practice for foresters to grow these two species together as they are thought to exploit different resources. In addition, when these two species are grown together they show the highest stability of all two-species compositions (not shown). This highlights that biological concepts such as asynchrony and stability are tightly interconnected and their analysis must communicate.

On this graph, we note another interesting pattern that will be inspected in Appendix C, differences in variability. The Norway spruce (third panel from left) seems to be more strongly affected by years, while the Sessile Oak (last panel on the right) shows the lowest variability.

The scope of this section was to introduce the concept of conditional estimates. We showed that that when quantifying asynchrony, we may want to control for the effects of other nuisance variables that may strongly affect the results. In the above example, we sequentially removed these effects before measuring asynchrony. This way to proceed was used for illustrative purposes only. Our approach uses a model-based framework where all ecological processes (e.g. asynchronies and stabilities) of interest are estimated simultaneously in a single step. This way to proceed gives us conditional estimates.

4 Final remarks

Complexity of modern ecological experiments - the power of modern graphical packages
In this document we highlighted the importance of using appropriate ways to visualising data for present-day ecological experiments. These are often complex with several design structures such as plot, block and sites measured over time. Modern graphical packages such as *lattice* or *ggplot2*[8] are very well-suited to properly display this kind of data.

Indeed, panelling can be used to display separately the effect of a treatment (e.g. species identity or species richness), while grouping is very powerful to illustrate observations that belong to the same entity (e.g. plot or tree). Other tools of modern graphical packages are very helpful to convey substantially more and better information in a given graph than simple scatterplots. Here we showed the power of ordering panels or changing aspect ratio. This, revealed unseen patterns that are biologically important.

In addition, the approach presented in this publication is fundamentally based on a graphical visualisation of the results obtained from the modelling phase (see Appendices B, C and D).

Conditional effects
In this document we introduced the concept of conditional effects. This concept is fundamental to the approach presented here and will be further discussed in Appendix B.

Notes on reproducibility
This dynamic document, that mixes **R** code, output and figures with plain text, was created using the add-on package *knitr* (Xie, 2016). All the appendices of this publication are created with this method. Each of the appendices is linked to a **R**-script that reproduces exactly the steps used to produce these documents. This is very convenient for those interested in interacting with the code to rerun our analysis or even modify and improve it.

[8]We used these two package for personal preference. However, the approach presented in this publication does not rely on any particular package or software.

Related to that last point (rerunning the analysis), it worth noting that to create this document we used the add-on package *checkpoint* (Corporation, 2016). This package "creates a local library into which it installs a copy of the packages required by your project as they existed on CRAN on the specified snapshot date. Your **R** session is updated to use only these package" (type `?checkpoint()` for further details). This makes sure that rerunning the provided code yields exactly the same results presented here on any computer, regardless of the packages installed by the user[9].

[9]The `R` version installed is extremely unlikely to affect the results obtained.

Reproducibility

Checkpoint date was set to 2016-05-01.

```
sessionInfo()

R version 3.3.2 (2016-10-31)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Debian GNU/Linux 8 (jessie)

locale:
 [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
 [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
 [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] ggplot2_2.1.0       latticeExtra_0.6-28 RColorBrewer_1.1-2
[4] lattice_0.20-34     knitr_1.14

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.7      digest_0.6.10    plyr_1.8.4       grid_3.3.2
 [5] gtable_0.2.0     formatR_1.4      magrittr_1.5     evaluate_0.10
 [9] scales_0.4.0     highr_0.6        stringi_1.1.2    reshape2_1.4.1
[13] labeling_0.3     tools_3.3.2      stringr_1.1.0    munsell_0.4.3
[17] colorspace_1.2-7
```

References

J. Chamagne, M. Tanadini, D. Frank, R. Matula, C. Paine, C. D. Philipson, M. Svátek, L. A. Turnbull, D. Volařík, and A. Hector. Forest diversity promotes individual tree growth in central European forest stands. *Journal of Applied Ecology*, 2016.

M. Corporation. *checkpoint: Install Packages from Snapshots on the Checkpoint Server for Reproducibility*, 2016. URL `https://CRAN.R-project.org/package=checkpoint`. R package version 0.3.16.

J. J. Faraway. *Linear Models with R*. CRC Press, 2014.

K. Gross, B. J. Cardinale, J. W. Fox, A. Gonzalez, M. Loreau, H. W. Polley, P. B. Reich, and J. Van Ruijven. Species richness and the temporal stability of biomass production: A new analysis of recent biodiversity experiments. *The American Naturalist*, 183(1):1–12, 2014.

A. Hector, C. Philipson, P. Saner, J. Chamagne, D. Dzulkifli, M. O'Brien, J. L. Snaddon, P. Ulok, M. Weilenmann, G. Reynolds, et al. The Sabah Biodiversity Experiment: A long-term test of the role of tree diversity in restoring tropical forest structure and functioning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1582):3303–3315, 2011.

E. Limpert and W. A. Stahel. Problems with using the normal distribution-and ways to improve quality and efficiency of data analysis. *PLoS One*, 6(7):e21403, 2011.

C. D. Philipson, D. H. Dent, M. J. O'Brien, J. Chamagne, D. Dzulkifli, R. Nilus, S. Philips, G. Reynolds, P. Saner, and A. Hector. A trait-based trade-off between growth and mortality: Evidence from 15 tropical tree species using size-specific relative growth rates. *Ecology and Evolution*, 4(18):3675–3688, 2014. ISSN 2045-7758.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org/`.

D. Sarkar. *Lattice: Multivariate data visualization with R*. Springer Science & Business Media, 2008.

D. Sarkar. *lattice: Trellis Graphics for R*, 2016. URL `https://CRAN.R-project.org/package=lattice`. R package version 0.20-34.

D. Sarkar and F. Andrews. *latticeExtra: Extra Graphical Utilities Based on Lattice*, 2016. URL `https://CRAN.R-project.org/package=latticeExtra`. R package version 0.6-28.

M. Tanadini and A. Hector. A modern model-based framework to compare ecosystem functioning. in prep.

S. L. Tuck, M. J. O'Brien, C. D. Philipson, P. Saner, M. Tanadini, D. Dzulkifli, H. C. J. Godfray, E. Godoong, R. Nilus, R. C. Ong, B. Schmid, W. Sinun, J. L. Snaddon, M. Snoep, H. Tangki, J. Tay, P. Ulok, Y. S. Wai, M. Weilenmann, R. Glen, and A. Hector. The value of biodiversity for the functioning of tropical forests: Insurance effects during the first decade of the Sabah Biodiversity Experiment. *Proceedings of the Royal Society B*, 283 (1844):20161451, 2016.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-PLUS*. Springer Science & Business Media, 2013.

H. Wickham and W. Chang. *ggplot2: An Implementation of the Grammar of Graphics*, 2016. URL `https://CRAN.R-project.org/package=ggplot2`. R package version 2.1.0.

Y. Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2016. URL `https://CRAN.R-project.org/package=knitr`. R package version 1.14.

Appendix B: Spatial and temporal asynchrony

Chapter 3

Contents

1 Introduction

The aim of this appendix is to show how to use our approach to model temporal and spatial asynchrony and to show its advantages over currently used metrics. It is usual procedure to analyse diversity-function in several disconnected steps. Often, the first of the analyses, looks at the effects of biodiversity on productivity. This almost always involves fitting some kind of Linear Mixed Effects Model. Afterwards, in a completely disconnected manner, other analyses are carried out (e.g. stability, asynchrony). Our approach differs fundamentally from the path used by classical metrics as it puts together, in a very powerful way, all these analyses. Our approach strongly relies on graphical visualisations.

After this brief introduction, Section 2 illustrates how Linear Mixed Effects Models can be fitted to diversity-function data. This section illustrates several modelling and testing techniques that will be used when modelling temporal asynchrony (Section 3) and spatial asynchrony (Section 4). Later, Section 5 illustrates how to use the method proposed here to quantify stability for traits such as counts or proportions. This section also elaborates on the concept of conditionality introduced in Appendix A. Finally, Section 6 summarises and adds some concluding remarks to this document.

2 Fitting a starting model to quantify biodiversity effects

2.1 Fitting a starting model

In this section we fit a Linear Mixed Effects Model to the *Czech dataset* graphically inspected in Appendix A ("Graphs"). To do that, we use the package *lme4* (Bates et al., 2016). As discussed in the main article, Linear Mixed Effects Models are well suited to analyse modern biodiversity function experiments. Indeed, these experiments virtually always have designs that are best analysed by using both random and fixed effects. As an example, the biomass of a grass plot or the diameter of a tree can be measured repeatedly over time. The non-independence of these grouped measurements must be taken into account. In addition to that, Mixed Effects Models also provide variance estimates that quantify the variability accounted for by each of the random effects[1]. As we will see, these variances can be interpreted biologically and also be used to design more powerful experiments.

After fitting the starting model, we perform inference (compute p-values and confidence intervals) on random and fixed effects to find a more parsimonious model. These steps are important as estimating temporal asynchrony and spatial asynchrony uses the same procedure. Indeed, the procedure presented here to quantify asynchronies uses existing inferential methods. In other words, we are not introducing a new way to test for asynchronies, but rather we introduce a new way to parametrise biodiversity function models that can be used to inspect asynchronies.

[1]Generalised Least Squares (GLS) can also be used to analyse non-independent data, but this method does not provide variance estimates.

Here, we are particularly interested in the effect of diversity and in the difference among species. Therefore, we start by fitting a model where species are allowed to respond differently to all explanatory variables. As a matter of fact, Appendix A indicated that the effect of several predictors may differ among species. In statistical words, we are fitting a model where the predictor `species` interacts with all other terms in the model[2].

```
require(lme4)
mod.0 <- lmer(log10(ring) ~ species * (density.stand + diversity.stand +
                                         date.fac + age + basal.area) +
                (1 | tree) + (1 | composition) + (1 | site),
              data = d.20)
```

The response variable is the log-transformed ring width. The model fitted contains 96 fixed-effect parameters and four variance components (i.e. the three random effects variances and the residual variance). The explanatory variable coding for year (i.e. `date.fac`) is modelled as factor and let to interact with species. This implies that 20 "year" parameters are estimated for each species (i.e. 80 in total). At first sight, the number of parameters estimated here may appear to be high. However, what really matters is the ratio between the number of observations and the number of parameters. In Linear Models the difference between these two gives the number of degrees of freedom. In Mixed Effects Models it is not entirely clear how to calculate degrees of freedom (Pinheiro and Bates, 2006). Nevertheless, the ratio is still indicative, here we are estimating 96 parameters with 10901 observations, their ratio is thus 114:1. Thus, this is undoubtedly large enough, as the number of observations is of two magnitudes larger than the usual thumb rule "10-20 observations per parameter" used in the Linear Model framework (Harrell, 2015).

The next logical step after fitting a model is to look at its summary information. For the sake of brevity this is not printed here.

```
summary(mod.0) ## not printed
```

We rather focus on the variances components estimated for the random effects.

```
VarCorr(mod.0)
```

```
 Groups      Name        Std.Dev.
 tree        (Intercept) 0.1325
 site        (Intercept) 0.0601
 composition (Intercept) 0.0000
 Residual                0.1458
```

Trees make up a considerable part of the variation. Indeed, their standard deviation (i.e. 0.13) is similar to the residual standard deviation (i.e. 0.15). Sites appear to explain less variability, while `composition` seems to be have no influence at all. `composition` is a design variable and ought to be in the model regardless of its contribution. Nevertheless, in this particular case the estimated standard deviation is numerically zero, and for that reason we will remove this term from the model[3]. Thus,

[2] In **R** formulas this is achieved by using the star symbol.

[3] Note that the results of the inferential procedure with or without `composition` as random effect are virtually identical.

we can conclude that the random effect composition is not explaining any additional part of the variability in this model that already contains species richness and species identity. Note that these results do not imply that that species richness is more important than species compositions as the two are modelled differently (one is a fixed effect, the other a random effect) and cannot be compared. Other model parametrisations allow the user to compare the relative effect of these two biodiversity predictors (Hector et al., 2011a; Tanadini and Hector, in prep.). Note that these results are reported in terms of standard deviation, rather than variances, as this has less skewed confidence intervals (Bates, 2010), and as standard deviation is what we see on graphs (Sarkar, 2008).

2.2 Statistical inference procedure when the assumptions are fulfilled

In this subsection we perform statistical inference while assuming that all the model assumptions are fulfilled. Section 5 of this appendix shows how to deal with cases where the distributional assumptions of the model are not fulfilled (e.g. when data does not follow a normal distribution).

It is not entirely clear whether inference should first be performed on the random effects part or on the fixed effects part of a Mixed Model (Pinheiro and Bates, 2006). Nevertheless, it can be argued that the random effects part should be addressed first as it computationally more involved (Martin Maechler personal communication).

The inferential procedure presented here is strongly based on the "model-building approach" introduced by Pinheiro and Bates (Pinheiro and Bates, 2006). As a first step we remove the random effect `composition` and refit the model via the `update()` function.

```
mod.1 <- update(mod.0, . ~ . - (1 | composition))
VarCorr(mod.1)
```

```
 Groups   Name        Std.Dev.
 tree     (Intercept) 0.1325
 site     (Intercept) 0.0601
 Residual             0.1458
```

Not unexpectedly, the standard deviations of the the retained random effects do not change when `composition` is removed. We now continue with the model simplification procedure by looking at the fixed effects.

The model that we fitted contains several interaction terms. It is thus important to respect the principle of marginality when carrying out inference (Venables and Ripley, 2013). In other words, main effects (e.g. density) can only be tested when no other higher terms (e.g. species:density) are present in the model. The function `drop1()` respect this principle. As recommended, nested models are compared using Likelihood Ratio Tests[4].

[4]Note that by default `lmer()` fits models using REstricted Maximum Likelihood (REML). However, REML models that differ in their fixed effects components cannot be compared (Bates, 2010). Therefore, all models are automatically refitted with Maximum Likelihood (ML) prior to comparison.

```
drop1(mod.1, test = "Chisq")

Single term deletions

Model:
log10(ring) ~ species + density.stand + diversity.stand + date.fac +
    age + basal.area + (1 | tree) + (1 | site) + species:density.stand +
    species:diversity.stand + species:date.fac + species:age +
    species:basal.area
                       Df   AIC  LRT Pr(Chi)
<none>                    -9301
species:density.stand   3 -9298    9   0.032 *
species:diversity.stand  3 -9306    1   0.803
species:date.fac        57 -7596 1819 < 2e-16 ***
species:age              3 -9274   33 3.8e-07 ***
species:basal.area       3 -9297   10   0.023 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By looking at the output of the `update()` function we find that there is some evidence that density affects species differently. There is no evidence that the diversity effect is different among species. There is strong evidence of an interaction between species and date as a factor. Indeed, `species:date.fac` shows the strongest effect even though a very large number of parameters (i.e. 57) are tested. This implies that species are asynchronous as they respond differently to years. Note that statistical significance does not imply practical or biological significance (Cox, 1982). In section 3 we will investigate the effect of time on radial growth among species. There is also strong evidence that the effect of `age` differs among species. Finally there is some evidence that tree size (i.e. `basal.area`) effect differs among species.

In this study we consider `density`, `age` and `basal.area` to be control variables. Indeed, our primary focus is to model the effect of diversity when controlling for the effect of other confounding variables (i.e. we are interested in the conditional effect of diversity). We therefore would keep these interactions in the model even if they were not statistically significant. Diversity is of primary interest, thus we remove its interaction with species as there is very little evidence for it. We then compute Likelihood Ratio Tests again.

```
mod.2 <- update(mod.1, . ~ . - species:diversity.stand)
drop1(mod.2, test = "Chisq")

Single term deletions

Model:
log10(ring) ~ species + density.stand + diversity.stand + date.fac +
    age + basal.area + (1 | tree) + (1 | site) + species:density.stand +
    species:date.fac + species:age + species:basal.area
                     Df   AIC  LRT Pr(Chi)
<none>                  -9306
diversity.stand       1 -9305    3   0.081 .
species:density.stand  3 -9302   10   0.019 *
species:date.fac      57 -7600 1820 < 2e-16 ***
```

```
species:age            3 -9279   33 3.4e-07 ***
species:basal.area     3 -9301   11   0.015 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As expected, essentially nothing changed for the other predictors. There is now some weak evidence that diversity may have an overall effect on ring growth. In Section 5 we will see how to further interpret these results.

2.3 Summary

The aim of this section was to present the inferential techniques that can be used to test for temporal and spatial asynchrony. As we will see next, these can be fully analysed within the usual framework of Mixed Effects Models.

Notes
It is important to note that `composition` should be considered a design variable and thus should not be removed from the analysis of biodiversity functions experiments. Very often random effects are necessary to capture the correlation structure of the data and thus must be kept in the model. However, in this semi-experimental context the estimated variance of this term was numerically zero. We thus removed it without any further testing. In general, formal statistical inference should be performed before removing a random effect which has non-zero variance. The recommended way to do that with `lme4` package is to estimate confidence intervals for the variance component[5] (Bates, 2010).

In this publication we use the model-building approach to find meaningful, but parsimonious models. However, our approach does not rely on it. Its main advantage is that enables us to fit models that can be interpreted biologically. This is fundamental to diversity-function studies that aim to understand the ecological effects rather than making good predictions.

3 Temporal asynchrony

3.1 Proposed approach

In Appendix A we illustrated how applying classical asynchrony measures to raw data can lead to misleading results. The main reason is that these metrics assume that species biomasses are fluctuating around a stable mean. In reality, this may often be untrue. For example, plot biomass or tree diameter may be increasing over time. We saw that when a trend is present in the data, these metrics wrongly estimate asynchrony to be strongly positive. In the next example we also show that when no

[5]Likelihood Ratio Tests could be used too, but these tests are not to be trusted as they are testing the parameters on the edge of the parameter space for variances (i.e. zero). This violates one of the assumptions underlying Likelihood Ratio Tests (Held and Sabanés Bové, 2014).

trend is present, assuming that the mean is constant may not be true and how this would in turn affect our estimates of asynchrony. In this study, it is clear that even when the trend is removed, there is still a "year effect" that is shared by all species. Some years are generally good years for all species while other are not. We may want to estimate asynchrony also when the "common year effect" has been removed.

We propose to estimate and remove the effects of confounding variables such as growth before looking at asynchrony. These effects may bias our results, by providing wrong biological answers. On top of that, we also propose to remove the effect of confounding variables that add noise to the signal without biasing it. For example, we saw that `density` may affect our response variable and this adds noise, however, the measured forest density is constant over time and thus does not bias the picture of temporal asynchrony.

Back to our example, we are now inspecting the temporal asynchrony for species. Note that our approach is flexible and can be used with any type of groups (e.g. functional groups). We saw in the previous section that the predictor `date` taken as a categorical variable strongly interacts with species identity. This is a clear sign of temporal asynchrony. However, this test and its associated p-value does not tell us a lot of biological information. This test simply tell us that the effect of the two variables is not additive, but that in at least one year, at least one species behaves differently from the others.

It is therefore important to have a better understanding of how species are responding to years, their relationships and whether the results are biologically relevant or not. The first step is thus to produce a meaningful graph. It is very important to use a visualisation that easily highlights difference among species. We have already seen that panelling can do that. However, superposition is the most powerful way to compare groups when their number is low (Sarkar, 2008).

Here we display the raw observations with different colours for the four species `(a)`. On top of that, we add the year average[6] for each species as a coloured line `(b)` and add a legend `(c)`. To focus on year averages we rendered actual observations smaller and made lines thicker `(d)`.

```
require(lattice)
xyplot(log10(ring) ~ date, data = d.20,
       groups = species, ## (a)
       type = c("a", "p", "g"), ## (b)
       auto.key = TRUE, ## (c)
       cex = 0.1, lwd = 2) ## (d)
```

[6] `type = "a"` stands for average.

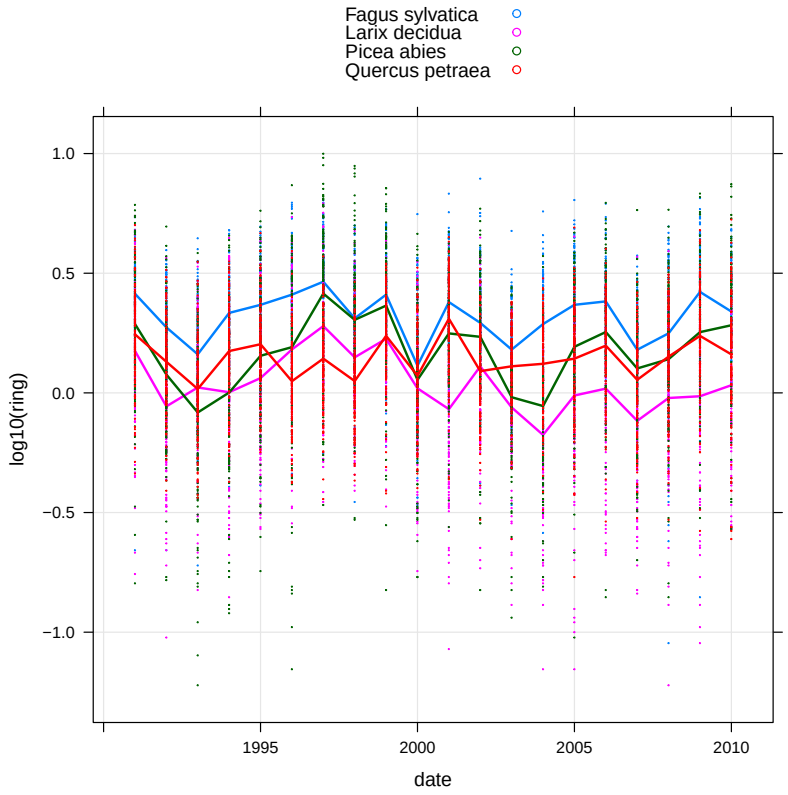


Figure 1: Response variable plotted against date. Species identity is higlighted by colour.

It is obvious that species averages are not running perfectly parallel. Nevertheless, species are not just fluctuating randomly around a stable mean value. Some years are undoubtedly better than others in terms of ring growth. Therefore, the assumption made by all classical asynchrony metrics does not appear to be fulfilled. It may be very interesting to partition the common year effect shared by all species and the species-specific year deviations. In other words, we may want to remove the general year effect to better focus on species differences. Finally, it may interesting to compare the variability shown by "common years effects" with that of "species-specific year deviations".

Nevertheless, there are other biologically interesting patterns to look at. As an example, the steepness of the descending segments can considered to be indicative of species resistance. On the other hand, the steepness of ascending segments can be related to resilience. For example, all species but for European larch have a very steep drop between 1999 and 2000, but then rise quickly again in the next year. These concepts are obviously tightly linked to species variability, which in the context of diversity-function studies is named stability (see Appendix C "Stability"). This is another confirmation that all parts of the analysis of a biodiversity function experiments are tightly linked together and cannot be analysed independently.

Below we illustrate one possible way to quantify temporal asynchrony by using mainstream statistical modelling tools. If there was no temporal asynchrony all species would react in the same way to temporal fluctuations. In that case we would correctly model the data with a model that does not contain the `species:date.fac` interaction

term. Through the use of the function `update()` we can quickly refit a model without that interaction. This model would make sense if the assumptions of additivity between years and species was fulfilled. By looking at the residuals of this model we can better understand how species react to years once we control for the other confounding effects.

```
mod.3 <- update(mod.2, . ~ . - species:date.fac)
```

Model `mod.3` contains `date.fac` which means that we are estimating the "common year effect". Thus, when looking at the residuals we are comparing species for the responses to years while we have removed the "common year effect". On top of that we also control for the other variables present in the model (i.e. we control for other nuisance effects).

Plotting the residuals of a model against each predictor is commonly used to check the structural assumptions[7] of Linear Models. Here, we may display the residuals of the model against diversity to check whether assuming a linear effect for diversity is correct. When plotting raw data we may miss patterns such as non-linear relationships. By looking at the residuals we can check this as we are controlling for the confounding effect of all other predictors. Of course, we could also include the non-linearity into the starting model and test whether this is needed through formal testing. Nevertheless, there are several ways to achieve that (i.e. shall we add a quadratic term or take the log of diversity?) and a statistically significant deviations from linearity may be practically irrelevant (Cox, 1982).

The graph below uses the residuals of the model that contains a "common year effect", but not a species-specific one. Thus, when plotting these residuals against date, we should see no pattern if species were to react in the same way to years. In other words, if species were perfectly synchronous the following graph would show a random scatter of residuals around zero[8].

```
xyplot(resid(mod.3) ~ date | species, data = d.20,
       groups = tree,
       type = c("l", "g"),
       col.line = "black",
       alpha = 0.2,
       abline = 0)
```

[7] These are the assumptions made on the relationship between response variable and predictors. In other words, the assumed model equation.

[8] As usual practice when visualising residuals we add a horizontal line on zero.

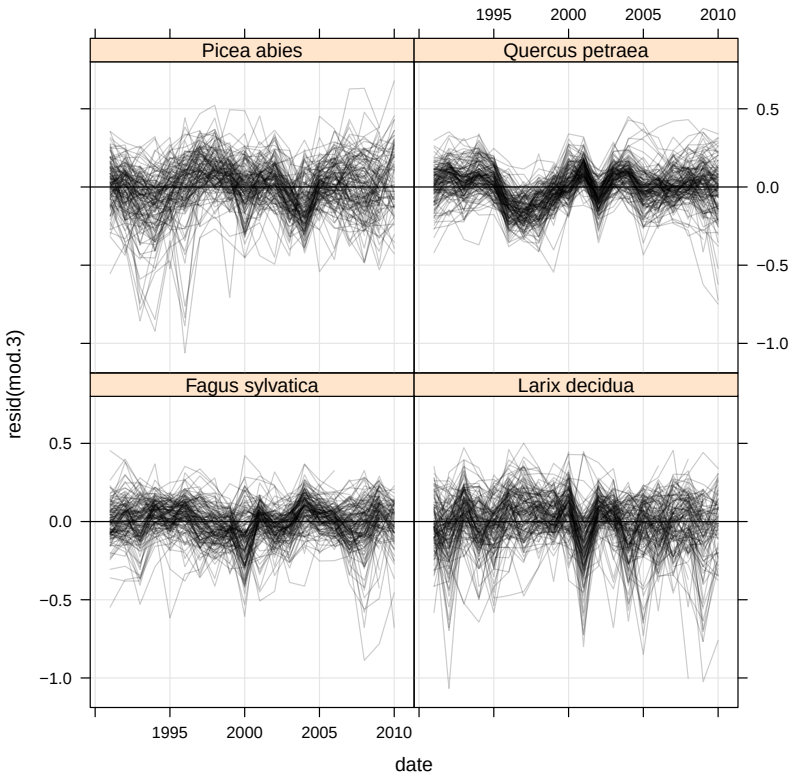


Figure 2: Residuals of a model containing "common years effects" plotted against date. Trees are rendered with lines.

This figure is tightly linked to Figure 1. The main difference between the two is that Figure 1 plots the response variable against date (i.e. shows their marginal relationship), while Figure 2 shows the response variable against date once the effect of all other variables present in the model has been removed (i.e. shows their conditional relationship). To better understand that difference, we note that the range of the y-axis in Figure 2 is smaller than the one in Figure 1 as we are controlling for several predictors. In other words, we can better focus on our specific question.

In this case this model contains date as a categorical variable, thus, by looking at the residuals we are visualising the species-specific deviations from the "common year effect"[9].

The figure above tells us that fitting a model without interaction `species:date.fac` is not enough. Indeed, the residuals are not randomly fluctuating around zero, but show clear patterns in time. This is an alternative way to interpret temporal asynchrony in a graphical way. Note that this graph highlights trees rather than species. Temporal asynchrony is computed using species, therefore, we now use another visualisation to better compare species.

In the next graph we display the residuals of `mod.3` against date while grouping for species. We put emphasis on species means, by adding averages `(a)`, by shrinking the

[9]Of course all considerations on the conditional graph rely on the assumption that we properly estimated the effect of the nuisance variables.

y-axis (b) and by making points smaller and lines thicker. we also add a non-default legend (d).

```
xyplot(resid(mod.3) ~ date.fac, data = d.20,
       groups = species,
       type = c("a", "p", "g"), ## (a)
       ylim = c(-0.2, 0.2), ## (b)
       cex = 0.1, lwd = 2, ## (c)
       col.symbol = "lightgray",
       auto.key = list(columns =  4, lines = TRUE, points = FALSE, cex = 0.75), ## (d)
       abline = 0,
       scales = list(x = list(rot = 45)))
```

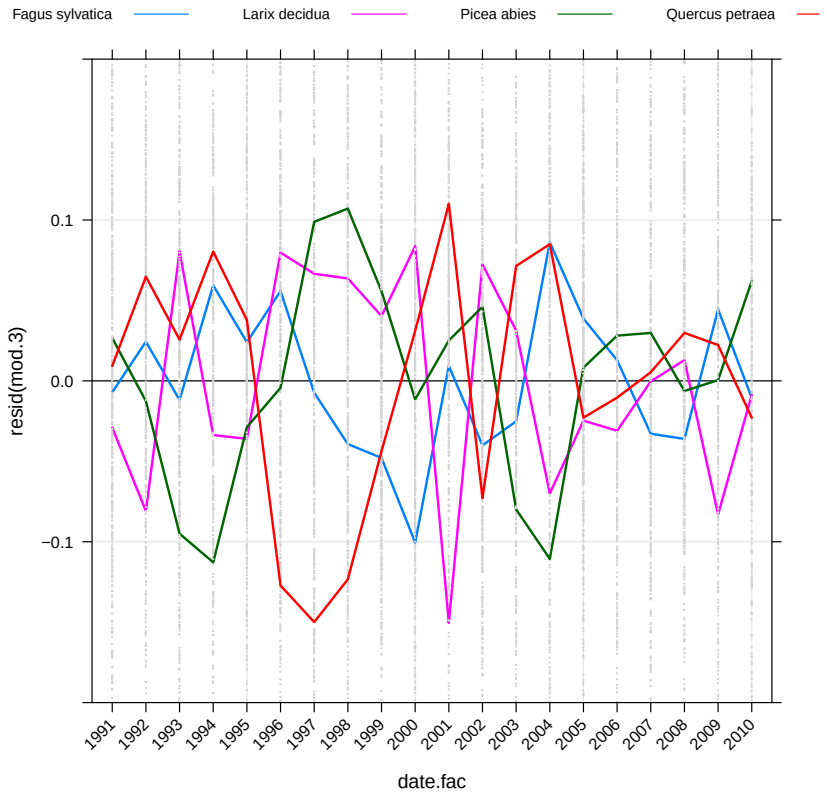


Figure 3: Residuals of a model containing "common years effects" plotted against date.

Once the confounding effects have been removed from the raw data, we note several interesting patterns. Sessile oak and Norway spruce behave in a very different way from each other (compare red and green lines). Positive year deviations for one species appear to be negative for the other species. In other words, the show opposite fluctuations around zero. This is particularly evident in the time period 1996-1999.

Finally, note that the year deviations seen on Figure 3 are actually estimated by the coefficients corresponding to the `species:date.fac` interaction term in the model `mod.2`. However, it may be difficult to interpret these numbers directly from the summary output as they depend on the model contrasts[10]. We may want to estimate post-hoc contrasts to test specific questions. For example, to test the hypothesis that

[10] **R** uses treatment contrasts by default.

in drought year 2003 ring growth was greater in broad-leaved species than in conifers. There is some evidence that this was the case[11]. A tightly linked publication showed how to test these questions in this context (Tanadini and Hector, in prep.).

Quantifying pairwise relationships

So far we have been visualising temporal asynchrony without producing specific graphs for pairwise relationships. A scatterplot matrix is well suited to visualise pairwise relationships (see Appendix A). To be able to use this graphical technique we must aggregate the data. Indeed, we want to compare the species-specific year deviations. In other words, we need the vertical distance of the "species-specific year mean" to the "common year mean" as seen on Figure 3. We can get these quantities from the the residuals of the model without interaction[12]. Through the use of the function `tapply()` we will compute the mean value of the residuals for each year-species combination. These are species-specific year deviations.

```
M.agg.3 <- tapply(X = resid(mod.3),
                  INDEX = list(d.20$date.fac, d.20$species),
                  FUN = mean)
dim(M.agg.3)
```

```
[1] 20  4
```

```
head(M.agg.3)
```

```
     Fagus sylvatica Larix decidua Picea abies Quercus petraea
1991          -0.007        -0.029      0.0266          0.0093
1992           0.024        -0.081     -0.0126          0.0648
1993          -0.012         0.080     -0.0950          0.0256
1994           0.059        -0.034     -0.1128          0.0804
1995           0.024        -0.036     -0.0288          0.0378
1996           0.056         0.080     -0.0043         -0.1272
```

As expected we obtain a matrix of 20 rows (i.e. years from 1991-2010) and four columns (i.e. species). Each entry of this matrix is the species-specific year deviation. We can now produce a scatterplot matrix with the *lattice* function `splom()`. Note that for improved visualisation we removed the axis scales `(a)` and added lines on zero `(b)`.

```
splom(M.agg.3,
      pscales = 0, ## (a)
      abline = list(h = 0, v = 0, col = "gray")) ## (b)
```

[11] This contrast is computed here but omitted from the dynamic document. See provided **R**-code.

[12] As mentioned above we already estimated these values as they coefficients of the interaction term between `species` and `date.fac` (from `mod.2`). Thus, we could also access them from the fitted model.

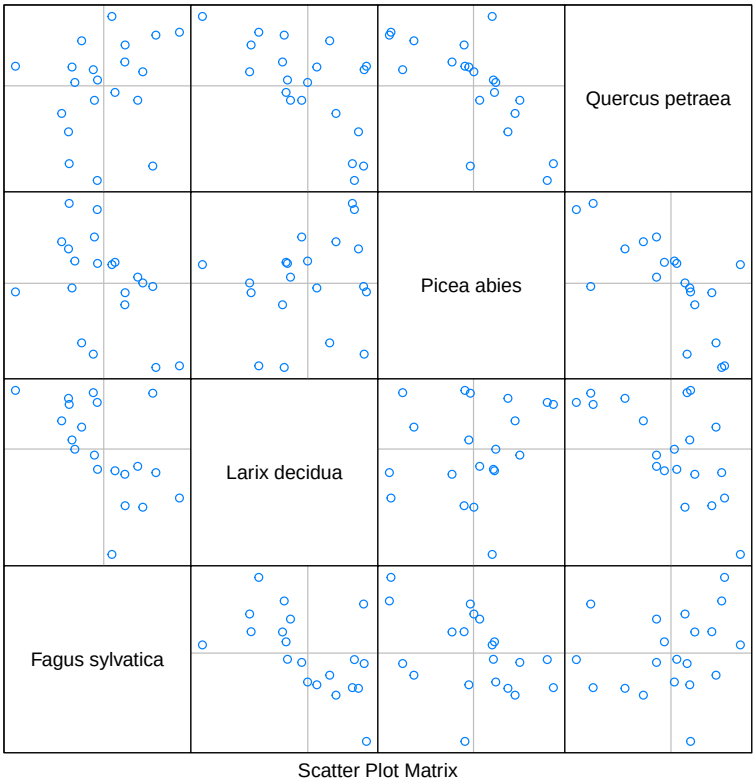


Figure 4: Scatterplot matrix of "species-specific" year deviations.

The suggested negative relationship between Norway spruce and sessile oak that we noticed on Figure 3 is very clear on this graph (i.e. top panel, right column). We also observe that other pairs are negatively correlated. European larch and European beech show a negative relationship too. This is biologically interesting as it confirms what foresters are doing: Planting these two species together as they are believed to exploit different resources. Obviously here we are just looking at the temporal dimension of the ecology of these species. The spatial dimension will give us further information. Although not shown here, these two species grown together show the highest stability. Once again, we make clear that by letting these analyses to communicate, we get a better understanding of the whole picture.

Note that Figure 4 is analogous to a correlation matrix. Indeed, we can compute the corresponding correlation matrix as follows.

```
( M.corr.3 <- cor(M.agg.3) )
```

```
                Fagus sylvatica Larix decidua Picea abies Quercus petraea
Fagus sylvatica            1.00         -0.56       -0.44            0.21
Larix decidua             -0.56          1.00        0.19           -0.66
Picea abies               -0.44          0.19        1.00           -0.70
Quercus petraea            0.21         -0.66       -0.70            1.00
```

The main difference is that to compute a correlation matrix we must make an additional assumption, linearity, which is not required by its graphical homologue, linearity.

Indeed, it is standard procedure to use the the Pearson correlation coefficient[13]. In addition, a graph gives us an hint of practical relevance of the estimated correlation.

3.2 Comparison with classical metrics

Classical metrics to quantify temporal asynchrony are usually applied on raw data, making the unrealistic assumption that species fluctuate around a stable mean. Below we perform the analysis using a widely used measure of asynchrony. This metric simply correlates the mean biomasses (or any other response variable) of all species in a pairwise manner. In practice, we correlate the pairs of biomass vectors measured over time, and then take the average of all pairwise correlations.

```
M.agg.raw <- tapply(X = log10(d.20$ring),
                    INDEX = list(d.20$date.fac, d.20$species),
                    FUN = mean)
( M.corr.raw <- cor(M.agg.raw) )
```

```
                Fagus sylvatica Larix decidua Picea abies Quercus petraea
Fagus sylvatica            1.00         0.534        0.74           0.625
Larix decidua              0.53         1.000        0.70           0.035
Picea abies                0.74         0.696        1.00           0.443
Quercus petraea            0.62         0.035        0.44           1.000
```

Interestingly, when using this metric on raw data we obtain results that speak for a strong positive synchrony between species. These results are mainly driven by the fact that species are not simply fluctuating around a stable mean, even if there is no trend (the response variable `diameter` showed a trend, `ring` does not). On the contrary there are important year-to-year variations in the data.

It can be argued that this is truly temporal asynchrony and that a stable mean is not necessarily required. Indeed, in a hypothetical case that all species react very closely to temporal fluctuations, we would observe this pattern. This is only partially true, and will be discussed further in Subsection 3.3. However, we must keep in mind that in some cases when looking at raw data we confound effects of age with those of time (see Appendix A "Graphs") and that by applying these metrics to raw data we do not control for the confounding effect of all other variables (e.g. tree variability is not removed).

Classical metrics usually try to condense all biological information into one single value. This is achieved by taking the mean value of all pairwise correlations. Below we compute the average correlation for the species-specific deviations[14].

```
require(gdata)
lt.raw <- lowerTriangle(M.corr.raw)
mean(lt.raw)
```

```
[1] 0.51
```

[13]Note that other correlation coefficients (e.g. Spearman) do not rely on the linearity assumption. We are not aware of their use in the analysis of biodiversity function experiments and they are not further discussed.

[14]This procedure is carried out with the `lowerTriangle()` function of *gdata* package (Warnes et al., 2015).

On average we have a correlation of about 0.51. There are a couple of important issues when we try to interpret this value biologically.

- The same average correlation value can result from very different situations. For example, the case that all pairwise correlations are 0.51. However, this case is biologically very different from the one seen above (see Figure 4). This issue is further developed at page 23.
- The average correlation that we can obtain in reality is bounded at $-1/(S-1)$ (where S is the number of species) (Willis, 1959). Thus, it is difficult to compare values across studies that differ in species number. Here we are actually very close to the boundary.
- It is not clear how to test whether this value is significantly different from zero[15].

Gross and colleagues presented a new metric that does not depend on the number of species (Gross et al., 2014). This metric looks at the correlation of each species and the sum of all the others (i.e. $\rho_i = cor(species_i, \sum_{j \neq i} species_j)$)[16]. All these elements are then averaged (i.e. $\bar{\rho} = mean(\rho_i)$).

Although the new averaged metric (i.e. $\bar{\rho}$) is bounded between -1 and 1 regardless of the number of species, its biological and practical interpretation does differ depending on the number of species involved. As an example, we may obtain the value -1 in two studies that have respectively 4 and 16 species[17]. This implies that in the first study there is a negative correlation of $-1/3 \cong -0.33$ between each species and the sum of all the the others[18]. However, this also implies that in the larger study (16 species) all elements are $-1/15 \cong -0.07$. So in the small study we have a strong negative correlation, while in the second we have essentially no correlation between each species and the sum of all the others.

The important thing to note is that the issue of this lower bound comes from averaging pairwise correlations. It is not a mathematical issue, but rather a real issue. Indeed, many species could react in the same way being all equal, but there is no way that all species react differently being all different from each other.

Instead of getting a single value to quantify temporal asynchrony we may want to use a single meaningful graph (i.e. the scatterplot matrix shown on Figure 4). There is no unique solution as slightly different graphs may be needed in different situations. We already saw in Appendix A that the effect of each predictor was best visualised with similar but not identical graphs.

It is clear that when the number of species involved is large, scatterplot matrices become quickly unreadable (i.e. they have too many panels). Below we show another option that is better suited to visualise large correlation matrices that was borrowed from Sarkar's *lattice* book (Sarkar, 2008). The **R**-code used to produce this graph is

[15] Obviously all the pairwise correlations cannot be considered independent as each species is involved in a pairwise manner $S-1$ times.

[16] For temporal asynchrony, $species_i$ and $species_j$ are two time series of identical length measured at the same time points.

[17] We chose to take 4 and 16 species as this is the number of species involved in the two studies presented in this publication (i.e. the Czech republic study and the Sabah Intensive Experiment).

[18] Note that when the average correlation proposed by Gross equals -1, there is only a single solution of the set of equations: All correlations equal -1.

long and omitted[19]. We visualise the correlation matrix corresponding to the "species-specific year deviations".

```
lt.3 <- lowerTriangle(M.corr.3)
mean(lt.3)  ## not a meaningful number to look at

[1] -0.33

##
f.corrPlot(M.corr.3)
```

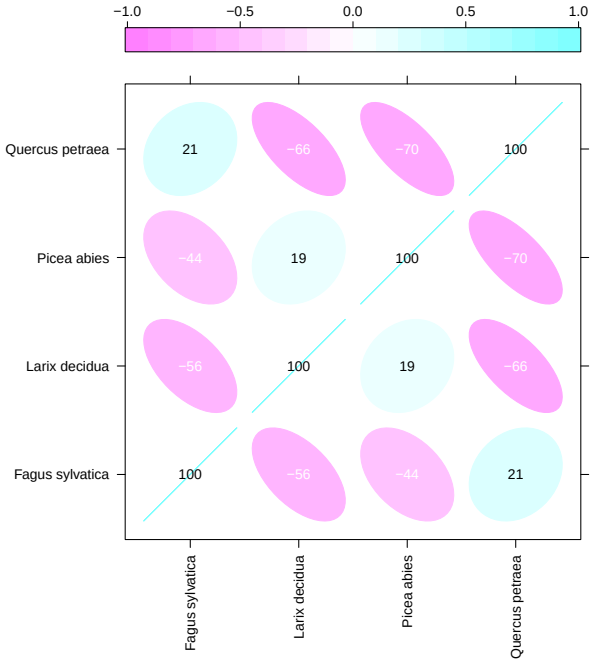


Figure 5: Levelplot used to display pairwise correlations. The shape and colour of the ellipses are based on the estimated correlation.

This type of visualisation can be used with a large number of species. The book presenting this technique used a correlation matrix with 18 columns (Sarkar, 2008). This graph is appealing as it is very intuitive to read. The colour and shape of the ellipses quantifies the sign and strength of the correlation (see legend at the top of fig. 5).

To conclude this subsection, we note that it may be argued that the plots presented here are not well suited to be included in a scientific publication as they use colours and can be quite complex (i.e. many panels). The average correlation is much easier to report as it is just a single value. We have seen above that this single value may not be biologically meaningful. We argue that a complex biological concept such as temporal asynchrony cannot be characterised by a single number that condenses the whole biological information. In our opinion, a meaningful graph is better suited to summarise temporal asynchrony. The next subsection illustrates an alternative

[19]The corresponding **.R** file does show that part of the code. This can easily be adapted to fit new situations. The original example can be found at page 240 of Sarkar's book.

solution that does not rely on complex graphs. This may be preferred when brevity is essential.

3.3 Partitioning time asynchrony into two components

Although a single graph may be a good option to display temporal asynchrony, there are cases where we may want to make a simple but robust claim on how much species react in the same way to annual fluctuations. One possible option is to formally estimate the variability accounted for "species-specific year deviations" and by "common year effects". The obvious way to do that is to use a Linear Mixed Effects Model where these two components are modelled as random effects. This enables us to estimate their variability within a well-accepted framework. We now fit this model to our data and look at the estimated variance components. To do this, we remove these two variables from the fixed effects part and add them as random effects. We then look at their estimated standard deviations.

```
mod.4 <- update(mod.2, . ~ . - date.fac - species:date.fac +
                  (1 | date.fac) + (1 | species:date.fac))
VarCorr(mod.4)
```

```
 Groups           Name        Std.Dev.
 tree             (Intercept) 0.1322
 species:date.fac (Intercept) 0.0748
 site             (Intercept) 0.0594
 date.fac         (Intercept) 0.0779
 Residual                     0.1459
```

Not surprisingly, the two estimated variances are almost the same (i.e. 0.075 for `species:date.fac` and 0.078 for `date.fac`). Indeed, this is what we observed in Figure 1 where we plotted the raw data against time while grouping for species. This figure showed that "common year" and "species-specific" fluctuations are of about the same magnitude. Therefore, we can claim that we are not in a situation where one of the two components is much stronger than the other. This means that this system does not show very strong asynchrony or very strong positive synchrony.

We can compare these results with the mean average correlation obtained when looking at the species-specific year deviations (i.e. `mean(lt.3)` $= -0.325$). At first sight, this average correlation value might seem to contradict the results obtained by partitioning the temporal fluctuations into two components. Indeed, the average correlation is very close to its lower boundary (i.e. $-0.\overline{3}$). In reality, these results do not contradict each other as they measure two different things. The ratio of variances computed with the estimates of `mod.4` quantifies how much of the temporal variability can be explained by a common response versus species-specific responses. Whereas the correlation matrix obtained with the species-specific deviations is useful to investigate the pairwise correlations. Note that the mean pairwise correlation of these deviations is meaningless[20].

[20]Indeed, there is no way that we can get positive mean correlations as this would mean that there is still some "common year effect" left. This is not possible as we have removed that by including date as a factor in our model.

The advantages of estimating these two variance components within the Mixed Effects Models framework is that we can make use of all the tools of this established methodology. For example, we can compute confidence intervals for the estimated variances and formally test if they are different from zero. Alternatively, we may test if species are perfectly synchronous by testing whether the `species:date.fac` variance component is different from zero. On top of that, we can make use of all graphical tools used to interpret variance components (see Subsection 4.3). Finally, it is possible to quantify the practical or biological relevance of these variance components by comparing them with the residual variance. In this case both variance components are about half of the residual variance, which means that they are not negligible. In other situations, where sample size is large, we may have a significant variance component that is several magnitudes smaller than the residual variance. In that case, we may conclude that these fluctuations are practically irrelevant.

One of the greatest advantages of estimating variance components, is that we can easily interpret them biologically. Here the standard deviation for year fluctuations is 0.078. This implies that the vast majority of the years (95%) lies within a $+/-$ 0.153 interval. As we are modelling the log-transformed ring width, this value has to be interpreted in a multiplicative way, which is not straightforward. However, we can make predictions in the back-transformed scale, so that these results can be interpreted biologically. For example, we can make a prediction for `Larix decidua` in the full-species composition when all the other fixed effect predictors are at their mean value.

```
## 1. create the data frame for predictions
d.pred.1 <- data.frame(species = "Larix decidua",
                       diversity.stand = 4,
                       density.stand = mean(d.20$density.stand),
                       age = mean(d.20$age),
                       basal.area = mean(d.20$basal.area))
##
## 2. make predictions (all random effects set to zero)
pred.1 <- predict(mod.4, newdata = d.pred.1, re.form = NA)
##
## 3. extract the sd of years
sd.years <- sqrt(unlist(VarCorr(mod.4)))["date.fac"]
##
## 4. back-transform interval
10^(pred.1 - sd.years * 1.96); 10^(pred.1 + sd.years * 1.96)
```

```
    1
0.784
   1
1.58
```

```
## these results are in the original "ring" scale (i.e. are in mm)
```

The results above imply that there can be quite some variability among years. Indeed, the interval obtained goes from 0.784 to 1.584 (these results are in the original untransformed scale, i.e. in mm)[21].

[21]Note that naively back-transforming (10^x) yields biased estimates (i.e. too small). Solutions exist (Tanadini and Mehrabi, in review). However, our aim here is to give a rough idea of the variability of years, not precise predictions.

Interestingly, we can compare the effect of a random effect with the one of a fixed effect. For example, we could look at the effect on ring width when the diversity of a site is increased from 1 to 4.

```
## 1. create the data frame for predictions
d.pred.2 <- data.frame(species = "Larix decidua",
                       diversity.stand = c(1, 4),  ## changing bit
                       density.stand = mean(d.20$density.stand),
                       age = mean(d.20$age),
                       basal.area = mean(d.20$basal.area))
##
## 2. make predictions (all random effects set to zero)
pred.2 <- predict(mod.4, newdata = d.pred.2, re.form = NA)
##
## 3. back-transform interval
( pred.2.back <- 10^(pred.2) )
```

```
    1     2
0.934 1.115
```

```
## these results are in the original "ring" scale (i.e. are in cm)
```

So the variability brought in by year differences is larger than the effect of increasing diversity by three units (from monoculture to full-species composition). This is quite an important and interesting point in terms of biology. This also highlights the fact that if we were to ignore the noise brought in by climatic fluctuations we may miss the effect of other variables of interest such as diversity.

Unfortunately, there are also some disadvantages to this approach. Making the choice that `date.fac` and `species:date.fac` are random effects has some consequences. Firstly, we cannot compute any post-hoc contrasts and thus say anything about years (e.g. "the 1997 drought was the worst year for all species"). When using the package *lme4* we also assumed that these effects were normally distributed which may not be the case[22]. In addition to normality, we assumed independence of the random effects. This is very unlikely to be fulfilled in reality as, for example, years close together are more alike than those far apart[23].

To be able to estimate a variance component, a sufficient number of levels is required[24]. Here, we have 20 years and 80 species-years combinations, which is largely enough. This may not always be the case, especially in grassland biodiversity studies were fewer temporal measurements are carried out. In these cases, the "fixed effects approach" presented in Subsection 3.1 is better suited. Note that this latter approach enables the user to test for asynchrony with as little as two time points. Classical metrics are based on correlations which need at the very least five or six observations to be estimated.

Finally, we want to stress the fact that this approach can be considered to sit half way between measuring temporal asynchrony and temporal stability (Appendix D expands

[22]There are solutions to this problem: The **R** packages *robustlmm* and *INLA* do not need to make these assumptions (Koller, 2016; Rue et al., 2009).

[23]Again, there are packages that can deal with that, e.g. *regress* and *INLA* (Clifford and contributions by HJ Auinger., 2014; Rue et al., 2009).

[24]The package used here (i.e. *lme4*) requires "at least six levels" (Bates, 2010).

on that topic). This highlights once again, how different aspects of diversity-function studies are interconnected.

3.4 Finding groups of species based on their responses to years

If the number of species is large, it may be interesting to focus on how species group together rather than to look at all pairwise correlations. To obtain these groups we can take advantage of clustering techniques. In this example, we are dealing with only 4 species. There is thus little interest in finding groups. However, the data set used in Appendix A (i.e. the Sabah Intensive Experiment) involves 16 species. Two examples of clustering are shown in Appendix D ("Multivariate extensions").

3.5 Relaxing the "discreteness requirements" of classical metrics (**)

One of the main "technical" requirements of classical asynchrony measures is time or space discreteness. To compute them we need to create vectors of biomasses measured at the same time point. In practice, however, carrying out all measurements at the same time points is not always feasible. As an example, the Sabah Biodiversity Experiment (Hector et al., 2011b) tracked the diameter of about 100,000 seedlings over time. In this case, it is clear that a full census of measurements can take a very long time (i.e. years). So for large studies, which are likely the most valuable, we are not allowed to use classical asynchrony metrics[25].

To illustrate this issue, we use the Czech Republic data set to simulate a situation in which measurements were not carried out at discrete time points[26]. To do that we use the `jitter()` function to add a small amount of noise to dates.

```
set.seed(2)
d.20$date.sim <- jitter(d.20$date, factor = 1.5)
##
xyplot(log10(ring) ~ date.sim | species, data = d.20,
       type = c("p", "smooth", "g"),
       span = 0.05,
       xlab = "Simulated date",
       col.symbol = "gray", cex = 0.1,
       lwd = 2)
```

[25] Obviously, it is always possible to discretise these measurements into groups (e.g. censuses). However, in so doing we are loosing important biological information and the obtained groups may be heterogeneous.

[26] Of course this makes little sense with dendrochronological data. It is just for illustrative purposes.

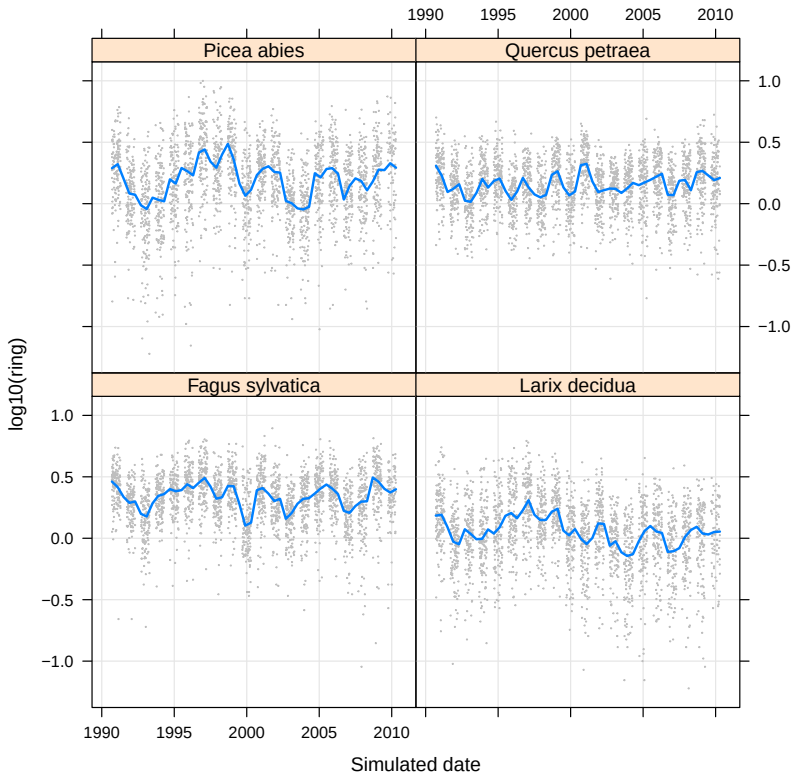


Figure 6: Response variable plotted against jittered dates.

This graph shows the response variable plotted against the simulated dates with a smoother on it. It was not possible to add a line connecting years as these are not discrete any more. Years are still recognisable on this graph, and it may be argued that we could discretise dates into years and use classical asynchrony metrics. However, is so doing we are loosing information and this is not always feasible (e.g. when dates of censuses overlap).

The obvious extension of Linear Models to be used here is smoothing (also known as Generalised Additive Models (GAMs); Wood (2006)). Essentially what we want to do is to test whether these smooth lines in the figure above can be considered to differ among species. This would tell us that there is temporal asynchrony (i.e. species react differently to time).

In **R** the most powerful package to fit these models is *mgcv* (Wood, 2016). This package enables the user to fit Generalised Additive Mixed Models (GAMMs) which are needed here (due to the presence of random effects). Line `(a)` of the call below states that we want a smooth term for simulated dates (i.e. `s()`), we allow this smoother to be computed independently for each species (i.e. `by = species`) and we specify how smooth this term must be (i.e. `k = 40`). Random effects are declared on line `(b)` [27].

```
require(mgcv)
gamm.0 <- gamm(log10(ring) ~ s(date.sim, by = species, k = 40) + ## (a)
```

[27]The syntax needed to declare random effects follows that used in the package *nlme* (Pinheiro et al., 2016).

```
                  species + diversity.stand + density.stand + age + basal.area +
                  species:age + species:basal.area,
                random = list(tree = ~ 1, site = ~ 1, species.site = ~ 1), ## (b)
                data = d.20)
```

To test whether there is evidence for time asynchrony, we fit the same model but drop the possibility to have different smoothers for jittered dates. Subsequently we perform statistical inference[28].

```
gamm.1 <- gamm(log10(ring) ~ s(date.sim, k = 40) +
                  species + diversity.stand + density.stand + age + basal.area +
                  species:age + species:basal.area,
                random = list(tree = ~ 1, site = ~ 1, species.site = ~ 1),
                data = d.20)
anova(gamm.0$lme, gamm.1$lme)
```

```
           Model df   AIC   BIC logLik   Test L.Ratio p-value
gamm.0$lme     1 26 -8720 -8530   4386
gamm.1$lme     2 20 -7351 -7205   3696 1 vs 2    1381  <.0001
```

Not surprisingly, there is very strong evidence that species fluctuate differently in time. We can visualise the estimated fluctuations by using the `plot()` method function for `gam` objects (i.e. `plot(gamm.0$gam, pages = 1)`). However, these effects are essentially the same as those shown on Figure 6.

So through a powerful extension of Linear Models we are able to cope with non-discrete measurements. Note that smoothing can be used for other terms of this model too. We may want to test whether the effect of species diversity is truly linear or not (i.e. we fit a model with `s(diversity)`). In this case, there was no evidence that the effect of diversity is not linear[29].

Generalised Additive Mixed Models are very powerful tools for ecologists (Wood, 2006). They can model non-linear relationships without making any assumptions on its form (unlike Non-Linear Models that make very strong assumptions on the model equation). This flexibility comes with a price, that is that we need to specify how smooth is the function to estimate. Above in our call we set the smoothness through the parameter `k` that controls the number of bases used for that smooth term.

3.6 Measuring asynchrony for grassland biodiversity studies

Our approach to model temporal asynchrony can be applied to grassland biodiversity studies as well. Nevertheless, these studies often have designs differing from the Czech Republic example presented here. It is thus important to interpret the results correctly. Long-term grassland biodiversity studies such as the Cedar Creek Experiment (Tilman et al., 2006) can be analysed without any further distinction.

[28] In *mgcv* GAMMs are fitted through the use of *nlme* package. For that reason the fitted objects have two slots, one for gam and one for the Mixed Model (i.e. $lme).

[29] The **R**-code to fit this model is omitted from this pdf document but present in the provided **R** file.

Shorter studies, however, may need additional attention. As an example, if we were to measure the height of grasses within one growing season, we would not be able to make strong claims about temporal asynchrony. Indeed, if there was a non additive effect between species and time, we may attribute that to two different, not mutually excluding, mechanisms. It may be that species have different growth patterns, but it may also be that species react differently to time fluctuations or it may be that both act simultaneously. If we wish to tell these two mechanisms apart we may run the same experiment over different years. This is what was done in our example where trees had different ages, such that it was possible to tell apart `date` and `age` effects.

Another fundamental difference between forest studies and grassland studies is that variables of interest are often measured at a different level. Tree rings are measured for each tree, while plot biomass may only be measured at "ecosystem level". Therefore, if sorting of species is not applied, we may not look at interactions between time and species. Nevertheless, we may want to look at temporal asynchrony for species composition. Again both approaches presented here can be used. In other words, we may model `time` and `time:composition` as fixed or as random effects. If few replicates of each composition are available, we may prefer to model temporal asynchrony using random effects.

3.7 Summary

We saw that within the framework of Linear Mixed Effects Models it is very easy to test temporal asynchrony. In essence, we just need to fit a model with an interaction term between the time variable and species (or any other grouping variable of interest, e.g. functional group). By testing these interaction terms we are actually testing whether species fluctuate in a synchronous way over time or not. To test these effects we propose to use established statistical methods which are based on very solid theory. In addition to looking at p-values and confidence intervals, we also suggest to produce meaningful graphs. This will help the user to understand whether a statistically significance result is also relevant from the biological point of view.

Currently used temporal asynchrony metrics try to summarise this complex biological pattern into a single value. We showed that meaningful graphs are better suited to understand temporal asynchrony. Very large correlation matrices can be meaningfully represented with a simple graph. One of the main issues of averaging correlations is that the obtained values carries very little biological information. Below we show two additional cases where the average pairwise correlation is exactly the same as that one observed for the "species-specific year deviations" (i.e. -0.326). Nevertheless these three situations are biologically very different. The first is the real case observed with species-specific deviations. In the second case all pairwise correlations are -0.326, while in the third case there are three values that fill the correlation matrix (-0.526, -0.326 and -0.126). These correlation matrices are printed below[30], and graphically displayed with level plots.

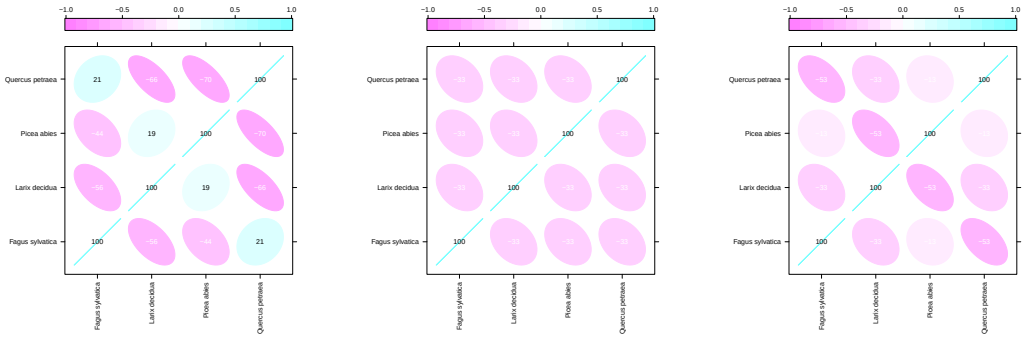
```
4 x 4 Matrix of class "dtrMatrix"
                Fagus sylvatica Larix decidua Picea abies Quercus petraea
```

[30]Note that for clarity only the lower triangle is printed.

```
Fagus sylvatica           1.000           .           .           .
Larix decidua            -0.556       1.000           .           .
Picea abies              -0.437       0.189       1.000           .
Quercus petraea           0.212      -0.662      -0.699       1.000
```

```
4 x 4 Matrix of class "dtrMatrix"
                Fagus sylvatica Larix decidua Picea abies Quercus petraea
Fagus sylvatica           1.000           .           .           .
Larix decidua            -0.326       1.000           .           .
Picea abies              -0.326      -0.326       1.000           .
Quercus petraea          -0.326      -0.326      -0.326       1.000
```

```
4 x 4 Matrix of class "dtrMatrix"
                Fagus sylvatica Larix decidua Picea abies Quercus petraea
Fagus sylvatica           1.000           .           .           .
Larix decidua            -0.326       1.000           .           .
Picea abies              -0.126      -0.526       1.000           .
Quercus petraea          -0.526      -0.326      -0.126       1.000
```



(a) Observed situation with the species-specific year deviations.

(b) Simulated situation where all pairwise correlations are 0.326.

(c) Another simulated situation.

Figure 7: Levelplots of the correlations matrices showing very different biological situations. Their average correlation is exactly the same (i.e. -0.326) though.

This figure shows with no doubt that the same average correlation may arise in very different situations. Indeed, these three situations are not simply a rearrangement of the pairwise correlation, but very different values populate these correlation matrices. Note that the metric proposed by Gross does not yield exactly the same value here[31], but suffers from another issue which is the fact that the same average correlation has a different meaning in studies with different number of species. As presented above when using the measure proposed by Gross and colleagues an average correlation of -1 with four species implies strong negative correlation, while the same value for sixteen species implies almost no correlation.

In this section we also showed that depending on the research questions and on the design of the study we may want to estimate asynchrony via fixed effects or via random effects. This latter is better suited when we aim to partition the temporal variability

[31] This is not shown here, but is present in the **R** code

into common and species-specific responses. Using fixed effects is more interesting when pairwise correlations are of interest.

Although this will only be shown in Appendix D ("Multivariate extensions"), we mentioned that we may take advantage of clustering techniques to find groups of species based on their response to time.

In Subsection 3.5 we used an extension to Linear Models to cope with non-standard situations where measurements were not carried out at discrete time points. This option is particularly valuable when dealing with spatial asynchrony as it will be shown in the next section.

Finally, although not explicitly shown, it is very important to note that the approach presented here can be applied to essentially any kind of data. Indeed, we may want to estimate temporal asynchrony for mortality rates (i.e. binomial data) or number of invading species (i.e. counts data) via Generalised Linear Models. None of the currently used metrics can be used in these situations. This is unfortunate as only a small subset of measured traits are actually analysed for temporal asynchrony. This may give us a biased picture of ecosystem functioning.

4 Spatial asynchrony

In this section we will show how to use our approach to test whether species react differently to space and how to quantify this biological interaction when present. The procedure used is very similar to the one used for temporal asynchrony. Both fit a model with an interaction between species and time or space respectively.

Again, we are going to use the Czech republic data set. In this context `site` is taken to be representative of space. Thus, we are looking whether sites can be particularly good or bad for species. There are 45 sites in this study and each of them contains several trees. Each site has a specific species composition.

In the previous section we showed that there are two alternative ways to fit a model that estimates temporal asynchrony. We may take both terms, here `site` and `species:site`, as random or as fixed effects. In this case, choosing sites to be random effects comes more naturally as we are not interested in their actual effects. In other words, we are not interested, for example, in the effect of site 24 or in comparing site 17 against site 4[32].

However, in other studies, in particular in multi-site studies like the Biodepth Experiment (Hector et al., 1999), we may prefer to consider site to be a fixed effect. Indeed, we may want to make claims about single places or groups of places (e.g. Southern Europe sites against Northern Europe sites). Often there may be several "space" indicators. In Biodepth, we had location, block and plot. It may be interesting to look at different spatial scales to see where spatial asynchrony is the strongest.

[32]On top of that, a model with site as fixed effect may not be identifiable as this latter can happen to be fully nested in species richness, or species compositions (Tanadini and Hector, in prep.).

4.1 Visualising spatial asynchrony

In analogy to what we have done in the previous section we may want to start by looking at the residuals of a model that does not contain the interaction term `species:sites`. If species were to react differently to space (i.e. sites), we would expect to see this on a residual plot. In particular, by plotting the residuals against a new predictor which is a combination of `species` and `sites`, we would see that residuals do not randomly scatter around zero, but group together. However, this procedure cannot be used here as our model contains the random effect `tree` which is fully nested in both `species` and `site`[33]. For that reason, if we were to use that graph mentioned above we would not see any pattern and we could mistakenly conclude that there is no spatial asynchrony. This graph is not shown here, but present in the **R** code.

We are using here the estimated variabilities to decide whether spatial asynchrony is present at site level or not. To do that we must add the interaction term to the model. After adding it, we compare the standard deviation of the random effects. For the sake of brevity and clarity we use the model that does not contain temporal asynchrony (i.e. `mod.2`).

```
mod.5 <- update(mod.2, . ~ . + (1 | species:site))
VarCorr(mod.5)
```

```
 Groups       Name        Std.Dev.
 tree         (Intercept) 0.1264
 species:site (Intercept) 0.0538
 site         (Intercept) 0.0476
 Residual                 0.1459
```

There is evidence that species may react differently to site conditions. Indeed, the standard deviation of `species:sites` is not dramatically smaller then the residual variance (i.e. 0.05 and 0.14 respectively) and similar to that one estimated for `site`. Residual variance can be taken as reference for the relevance of a random effect (Pinheiro and Bates, 2006).

Note that as we noticed for temporal asynchrony, the two variance components `site` and `species:site` are very similar (i.e. both about 0.05). This implies that the variability brought by the differences between sites is of similar magnitude as the variability explained by species-specific responses to sites. These results are thus in clear agreement.

4.2 Formal statistical inference for spatial asynchrony

P-values

To formally test whether species react differently to sites, which implies spatial asynchrony at site level, we may carry out statistical inference. In analogy to what we have done with fixed effects, we may want to compute a p-value for the additional random effect `species:site`.

[33] Obviously, any one tree can only belong to one species and come from one site.

```
anova(mod.2, mod.5)
```

```
Data: d.20
Models:
mod.2: log10(ring) ~ species + density.stand + diversity.stand + date.fac +
mod.2:     age + basal.area + (1 | tree) + (1 | site) + species:density.stand +
mod.2:     species:date.fac + species:age + species:basal.area
mod.5: log10(ring) ~ species + density.stand + diversity.stand + date.fac +
mod.5:     age + basal.area + (1 | tree) + (1 | site) + (1 | species:site) +
mod.5:     species:density.stand + species:date.fac + species:age +
mod.5:     species:basal.area
      Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
mod.2 96 -9306 -8605   4749    -9498
mod.5 97 -9312 -8604   4753    -9506  7.89      1      0.005 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This test tells us that there is evidence that species react differently to sites. However, statistical significance does not imply practical or biological significance (Cox, 1982). Nevertheless, we have already seen that this effect is not irrelevant as its standard deviation is about one third of the residual variance.

It is important to note that a non-significant result does imply that spatial asynchrony is absent or negligible (Altman and Bland, 1995). Indeed, we may be looking at the wrong spatial scale. For example, soil conditions may greatly vary within each site and species may react differently to it. However, when analysing spatial asynchrony at site level, we miss that pattern because we are inspecting that pattern at a spatial scale that is not fine enough.

In the previous section we analysed temporal asynchrony at year level, which makes biological sense. In the other hand, here we are using sites as indicator of space. Sites were defined based on purely practical aspects (i.e. they had to contain all interesting trees), but not on a biological basis.

Confidence intervals

We have already mentioned that using Likelihood Ratio Tests to test random effects is inappropriate as some of the fundamental assumptions of the test are not fulfilled, and these tests are conservative (Bates, 2010). An interesting and more powerful alternative is to compute confidence intervals for the variance components. When using the package *lme4* the recommended way to do that is by using profiling likelihood methods (Held and Sabanés Bové, 2014). This method is implemented in the `profile()` function. Note that this method is computationally expensive as it involves refitting a large number of models (see Bates (2010) for an explanation of how this method works).

We are now estimating confidence intervals for all random effects present in the model. This is specified via the argument `which`. Subsequently, we use the generic function `confint()` to get confidence intervals estimated by profiling likelihood.

```
prof.5 <- profile(mod.5, which = "theta_") ## theta_ is for random effects
VarCorr(mod.5)
```

```
 Groups       Name        Std.Dev.
 tree         (Intercept) 0.1264
 species:site (Intercept) 0.0538
 site         (Intercept) 0.0476
 Residual                 0.1459
```

```
confint(prof.5)
```

```
       2.5 % 97.5 %
.sig01 0.118  0.135
.sig02 0.023  0.068
.sig03 0.022  0.069
.sigma 0.143  0.147
```

Unfortunately, the output of the `confint()` function does not use the same labelling for random effects that we used when fitting the model. Instead, the variance components are named `.sig--`, where `--` is a number varying from `01` to the total number of random effects present in the model (i.e. three → `03` for `mod.5`). However, we can evince the order of the terms by looking at output of the `VarCorr()` function. So `.sig01` is actually the random effect `tree`, `.sig02` is `species:site`, `.sig03` is `site` and `.sigma` is the residual variance.

4.3 Graphically displaying confidence intervals (*)

We can visualise the estimated confidence intervals through the corresponding `xyplot()` method function. As you can see by looking at the next graph, the method functions used by *lme4* are actually using *lattice* package that was widely used in Appendix A.

We modified slightly the default call to better highlight some important patterns to be noted here. In particular, we stacked all panels on a single column to better compare across them `(a)`. For the same reason we forced all panels to have the same x-range `(b)`. We use 95% confidence intervals `(c)` and finally, we order panels with the ordering found in the output of the `VarCorr()` function (i.e. residual variance at the bottom) `(d)`.

```
xyplot(prof.5, layout = c(1,4), ## (a)
       xlim = c(0, 0.15), ## (b)
       conf = c(2.5, 97.5) / 100, ## (c)
       as.table = TRUE) ## (d)
```

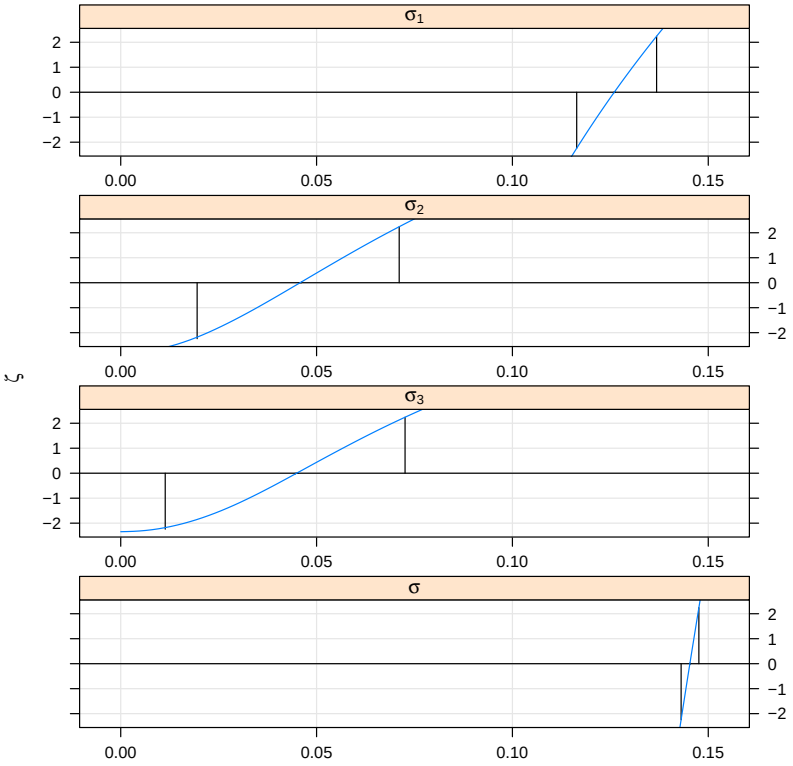


Figure 8: Profile-zeta plot for the random effects of mod.5. Confidence intervals at 95 percent level are shown with black vertical lines.

This graph, named "profile-zeta plot" (see Bates (2010) for more information), is well-suited to compare the variability explained by the different random effects. The two black vertical lines in each panel define the confidence intervals for the estimated standard deviation. For example, the 95% confidence interval for `tree` (top panel; σ_1) goes from about 0.12 to 0.14.

If a confidence interval was to contain or be very close to zero, then we would assume that this random effect is non-significant. Given that the residual standard deviation is also plotted along with its confidence interval, we can also judge the relevance of the other random effects. Indeed, the confidence interval for a given random effect may be clearly distinct from zero, but its actual value so much smaller than the residual variance, that it would be considered irrelevant.

In this case, we see that the random effect `tree` (top panel) explains a fair amount of the variation present in the data as its variance estimate is close to the residual one (bottom panel). The two spatial components (i.e. `site` and `species:site`; panels σ_3 and σ_2 respectively) are not to be considered very important as they are considerably smaller than the residual standard deviation and not too far from containing 0.

On this graph we can also see that the number of observations used to estimate the variability of a given random effect strongly influences the width of the obtained confidence interval. This makes perfect sense, as for example, we have more than 10,000 observations to estimate the residual variance and its confidence interval is very narrow (bottom panel). On the other hand, there are only 45 sites and thus the estimated

standard deviation for this random effect is much less precise (i.e. wider; see panel σ_3). We may want to reproduce the same graph for temporal asynchrony to see whether its effect is actually larger than spatial asynchrony[34].

The graph above very well represents the philosophy of the approach presented in this publication. We are making use of powerful existing tools developed by world-leading statisticians, tested by millions of users. We make best use of these techniques by tailoring them to our needs.

4.4 Quantifying pairwise correlations

In analogy to what we did for temporal asynchrony at page 13, we could use a scatterplot matrix to visualise pairwise correlations. In particular, we could take the conditional modes of the random effects `species:site` and correlate them. These represent the species-specific site preferences. Note that obviously not all sites have the four-species compositions, therefore, the number of observations is not 180 (i.e. 45 sites times four species), but 96.

When dealing with temporal asynchrony, we estimated the species-specific year deviations from the residuals of the model without interaction between `species` and `date.fac`. We claimed that we could get these values also from the model that contains the interaction. Essentially, we would achieve that by taking the predicted values for each `species-date.fac` combination.

Here we are dealing with random effects. So we can get the predicted values of the random effects (i.e. the so called conditional modes, Bates (2010)). These can be extracted with the `ranef()` function. In order to display them with the `splom()` function to inspect pairwise correlations, we must rearrange them. These steps are a bit involved and long.

```
## 1. extracting the conditional modes
re.5.sp.site <- ranef(mod.5)[["species:site"]]
##
## 2. adding two colums at this data frame (species and sites)
re.5.sp.site$species <- factor(sub("\\:.*", "", rownames(re.5.sp.site)))
re.5.sp.site$site <- factor(sub("^.*?\\:", "", rownames(re.5.sp.site)))
rownames(re.5.sp.site) <- NULL
##
## 3. renaming one column for plotting purposes
names(re.5.sp.site)[1] <- "cond.mod"
##
## 4. checking everything again
str(re.5.sp.site)
```

```
'data.frame': 96 obs. of  3 variables:
 $ cond.mod: num  -0.0279 0.0581 -0.017 -0.0179 -0.0125 ...
 $ species : Factor w/ 4 levels "Fagus sylvatica",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ site    : Factor w/ 45 levels "1","12","15",..: 1 12 2 3 4 8 16 17 18 19 ...
```

```
##
## 5. reshaping the data frame into wide format (a matrix)
re.5.sp.site.wide <- reshape(data = re.5.sp.site, v.names = "cond.mod",
```

[34]The **R**-code is hidden.

```
                              timevar = "species",
                              idvar = "site", direction = "wide")
  ## head(re.5.sp.site.wide)
##
## 6. for plotting reasons we use shorter names
names(re.5.sp.site.wide)[2:5] <- c("Fagus sylvatica", "Larix decidua",
                                   "Picea abies", "Quercus petraea")
```

```
splom(~ re.5.sp.site.wide[ , -1], pscales = 0,
      abline = list(h = 0, v = 0, col = "gray"))
```

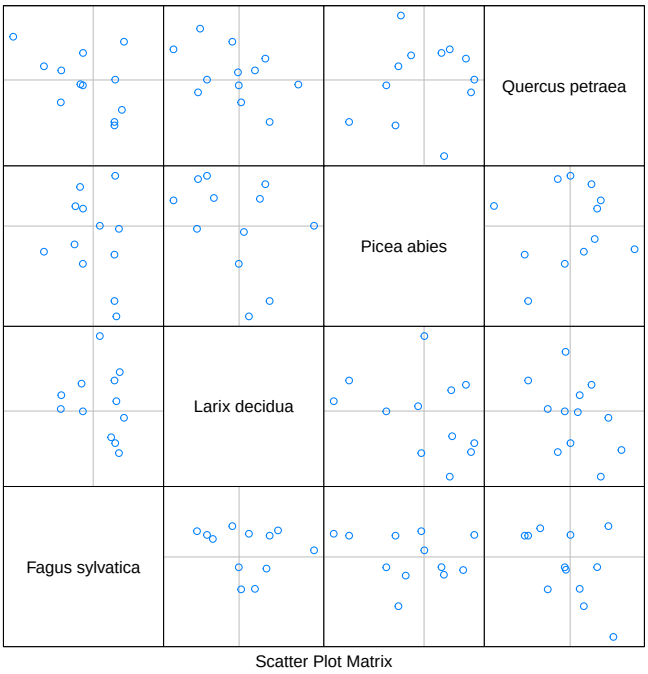


Figure 9: Scatterplot matrix of species-specific site preferences. Each dot represents a site.

This scatterplot does not show any striking pairwise correlation. We may interpret these results as "there is some evidence that species react differently to sites, as the variance component of `species:site` is not close to zero. This implies that spatial asynchrony is present at that spatial scale (i.e. site). However, there is no evidence of strong pairwise correlations between species."

This is a graphical way to analyse pairwise correlations. We may want to formally estimate Pearson correlations. We can do that with the `cor()` function. Note that we specify the `use` argument to prevent the function dropping all non-four-species sites. By default the function would just use the full-mixture sites, which are actually only three.

```
( cor.spat.asy <- cor(re.5.sp.site.wide[ , -1], use = "pairwise.complete.obs") )
```

```
                Fagus sylvatica Larix decidua Picea abies Quercus petraea
Fagus sylvatica            1.00         -0.19       -0.21           -0.49
Larix decidua             -0.19          1.00       -0.36           -0.42
Picea abies               -0.21         -0.36        1.00            0.15
```

```
Quercus petraea          -0.49          -0.42          0.15          1.00
```

By looking at these numbers is very difficult to judge their relevance. Some correlations are as high as -0.49, but if we look at the graph we clearly understand that they are not relevant. We may compute p-values for these correlations, but it is not entirely clear how we should correct for multiple testing and due to the low sample size a clear pattern may not be significant after correction. For example, the classical metric used to estimate asynchrony would estimate the mean correlation as being -0.25. However, it is not clear how we could test for this being different from zero, and this does not tell us anything about the single pairwise correlations.

In addition, when estimating Pearson correlations and testing them we assume normality of the observations and linearity of the pairwise relationship. In reality this may not be true. The graph below simulates this situation.

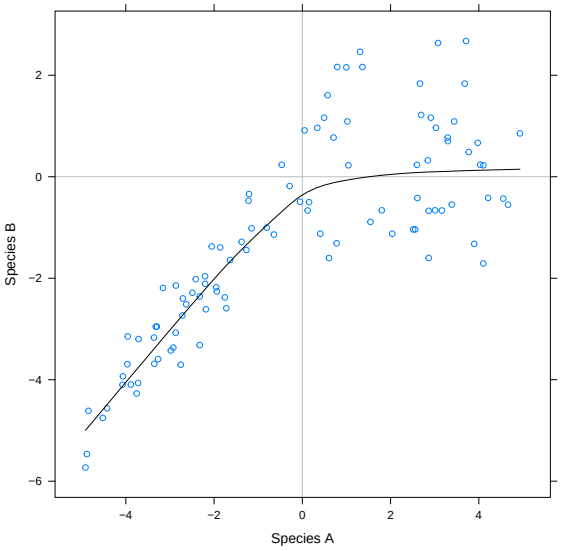


Figure 10: Simulated non-linear relationship between two species.

If the relationship between two species is not linear, we must be very careful when using Pearson correlation. As an example, dry and nutrient-poor sites may negatively affect species A and B. This is reflected in the linear correlation in the lower left quadrant of the graph above. But at the same time, good sites for species A are not necessarily good sites for species B. This is reflected by the absence of correlation in the upper right quadrant. We added a smoother to this graph to facilitate the reading of the correlation.

In this case, by using a classical Pearson correlation, we would miss this biologically interesting pattern. We may fix this problem, by using other correlation coefficients (e.g. Spearman). However, this would just create yet another metric that still suffers from other problems listed previously.

It is important to note that the metric proposed by Gross (Gross et al., 2014) cannot be used to estimate spatial asynchrony. Indeed, only sites that have the four species compositions can be used. All other sites are dropped as it is not possible to compute

the sum of all species as some are not present. In this particular case, we would drop 42 of the 45 sites. This is very inconvenient and obviously estimating a correlation based on three observations is nonsensical.

4.5 Generalised Additive Models to capture spatial asynchrony

Similarly to what we presented for temporal asynchrony, we can also use smoothing techniques to assess spatial asynchrony. The main difference is that we are now working in two dimensions (i.e. with latitude and longitude)[35]. We may model the effect of space with a smooth surface. Essentially, instead of giving a value to each site, that assume that the conditions are exactly the same at each point in one site, we allow for the effect of space to vary smoothly.

This technique may be particularly useful with design such as that one of the Sabah Biodiversity Experiment. In this large scale experiment seedlings were planted at very high density over the entire area. Plots to which species diversity treatments were applied are very large (i.e. 200m by 200m). It likely that space conditions do vary within plots. Therefore, using a smooth surface instead of assuming a constant plot effect may be very important to characterise spatial asynchrony.

Fitting a Generalised Additive Model to model space may not always be easy (Simon Wood personal communication). This is the price to pay for dropping the spatial discreteness assumption (i.e. conditions are not varying within plot). Interestingly, the use of a smooth surface indicates, in a objective way, the estimated scale at which spatial asynchrony acts[36]. We applied this technique to the Sabah Biodiversity Experiment data set previously analysed (Tuck et al., 2016) and found that spatial asynchrony was the strongest at a scale of about 20 metres (results not shown). In this particular experiment, plots had a size or 200 metres by 200 metres, so they are much larger than the actual scale at which the ecological process studies works. Using classical spatial asynchrony metrics we may thus miss or underestimate the strength of spatial asynchrony.

4.6 Summary

This section showed how to estimate spatial asynchrony by using the Mixed Effects Models framework. The differences with the approach shown for temporal asynchrony are minimal. In Subsection 4.2 we discussed how to perform formal inference when estimating spatial asynchrony.

Subsequently, we showed how to graphically display the confidence intervals obtained for the random effects. Nevertheless, the plots used, known as "profile-zeta plots" (Bates, 2010), are tailored to the needs of the analysis of biodiversity function experiments.

[35]This can be modelled using `s(latitude, longitude, by = species)`.

[36]The Estimated Degrees of Freedom (edf) of the smooth surface carry this information. Roughly speaking, we divide the size of the study area modelled with the smooth surface by the corresponding edf.

5 Extending the Gaussian Linear Model and model interpretation

5.1 Modelling non-normal data

Up to this point we assumed that the data modelled followed a Gaussian distribution (i.e. was normally distributed). In reality, this is not always the case. We may want to model count data or other types of non-normal data. The approach presented here enables the user to do that very easily. Generalised Linear Models, GLMs for short, are used to extend the Linear Model in these situations.

If we were to use the *lme4* package, then we should fit a model with the `glmer()` function and specify the `family` argument (e.g. `glmer(..., family = poisson)`). A real example will be presented in Appendix D ("Multivariate extensions").

Using Generalised Linear Models allows us to model a large variety of data types by using the distributions present in the exponential family (e.g. poisson, binomial, ...). However, in some situations we do not want to make any assumption about the distribution of our data. Nevertheless, we still want to be able to make inference and make strong conclusions on the obtained results.

This situation may arise when the data do not naturally fit into any of the several families available. Bootstrapping techniques are very useful here, as they enable the user to drop most distributional assumptions made when using Generalised Linear Models.

As an example, the tree ring data modelled in this document does not perfectly follow a normal distribution (i.e. residuals are slightly asymmetric and long-tailed). Therefore, it may of interest to use bootstrap to perform statistical inference. Below we use bootstrap to estimate confidence intervals for the effect of diversity which we saw was close to significance level when assuming normality.

Several types of bootstrap exist. Below we use semi-parametric bootstrap where both residuals and random effects are resampled (type `?bootMer` for more information). Subsequently, the `boot.ci()` function of *boot* package is used to compute 95% confidence intervals (Canty and Ripley, 2016).

```
boot.5.div <- bootMer(mod.5, FUN = function(x) fixef(x)["diversity.stand"],
                      nsim = 10^3,  ## 1.000 simulations
                      seed = 2,  ## for reproducibility
                      use.u = TRUE,  ## resamples random effects
                      type = "semiparametric")  ## resamples residuals
require(boot)
boot.ci(boot.5.div, type = "basic")

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot.5.div, type = "basic")
```

```
Intervals :
Level      Basic
95%   ( 0.015,  0.025 )
Calculations and Intervals on Original Scale
```

The output above shows that the confidence intervals for the effect of `diversity` does not contain zero (i.e. 0.015-0.025). This implies this predictor has a statistically significant effect at the 5% level. Note that this way to estimate confidence intervals can be applied to any parameter of the model. As an example, to get confidence intervals for the variance components we just need to adapt our code[37].

5.2 Partial residual plots and conditional effects

Going back to the effect of `diversity`, the interval obtained is not very far from zero. As we already noted, statistical significance is just a tiny part of the whole picture when we interpret variable effects (Wasserstein and Lazar, 2016). Ideally, we would like to be able to make statements about the biological or practical relevance of the predictor. A meaningful graph may be helpful for this task. Appendix A showed that by plotting the raw data against each predictor we see their marginal relationship. We may want to visualise the conditional effect of `diversity` instead. In applied statistics the most common way to do this, is to use the so called "partial residual plots" (Faraway, 2014). These are very useful as they illustrate the conditional effect of a predictor. In other words, it displays the effect of e.g. `diversity` when we control for all other predictors in the model. To be able to judge the relevance of an effect, residuals are added to these graphs.

```
## 1. creating the data frame where everything but diversity is kept constant
d.div <- data.frame(diversity.stand = d.20$diversity.stand,
                    ## fixed predictors:
                    species = "Larix decidua",
                    age = mean(d.20$age),
                    date.fac = "2000",
                    density.stand = mean(d.20$density.stand),
                    basal.area = mean(d.20$basal.area))
##
## 2. making predictions for this data set
t.div <- predict(mod.5, newdata = d.div, re.form = ~ 0)
##
## 3. adding residuals to predictions
d.div$fit.div <- t.div + residuals(mod.5)
##
## 4. producing the partial residual plot
xyplot(fit.div ~ diversity.stand, data = d.div,
       ylab = "Predicted values + residuals (log-scale)",
       xlab = "Diversity [exp(Shannon Index)]",
       type = c("p", "g", "r"),
       cex = 0.5,
       col.line = "black")
```

[37] This is not shown here, but present in the **R**-code.

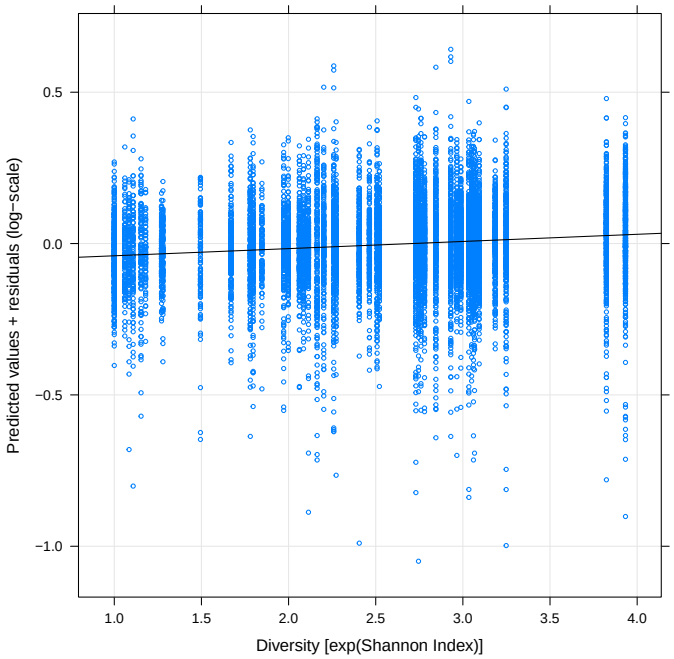
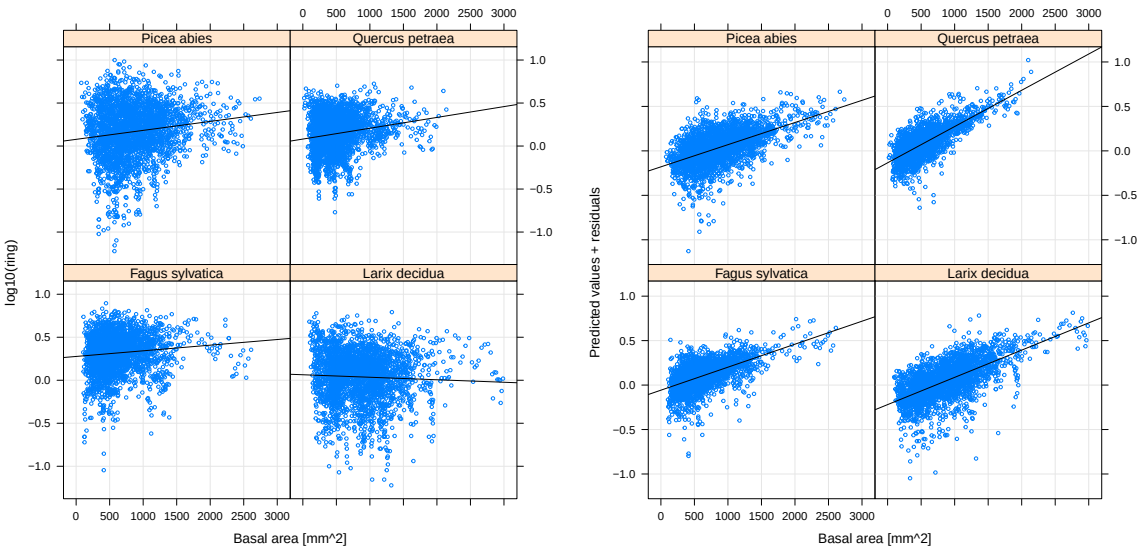


Figure 11: Partial residual plot for diversity.

This graph confirms what we have seen by looking at the p-value and confidence interval for `diversity`. This predictor has a weak positive effect on the response variable. Note that we are able to make claims about the relevance of an effect because the residuals are plotted along the fitted values. Whether this effect is biologically relevant or not depends on the context. If showing an effect of biodiversity, even if small, is important than this effect may be considered relevant. On the other hand, the effect shown here may be practically irrelevant for foresters. Note that if we were interested in a quantitative interpretation of the `diversity` effect in the original scale, we may back-transform the obtained values.

We now turn our attention to another predictor. When were looking at the effect of `basal.area` in Appendix A we did not see a strong and clear relationship. However, it is well known that larger trees grow faster (Chamagne et al., 2016). The figure below shows the marginal (left) and conditional relationship (right) between the predictor `basal.area` and the response variable.



(a) Marginal relationship (raw data) (b) Conditional relationship (partial residual plot)

Figure 12: Marginal and conditional effects of basal area.

By comparing these two graphs, we understand the power of partial residual plots. They are able to remove the effect of the others predictors in the model and display the real effect of a predictor. When looking at raw data (graph (a)), we do not see a clear relationship between `basal.area` and the response variable. This is likely due to the confounding effect of the other predictors. In particular, `age` may be obfuscating the relationship. As the right graph (b) correctly shows, larger trees grow faster. However, older trees, that are obviously larger, have lower growth rates. Thus, what we see on the left graph is a sum of these effects. Whereas what we see on the right panel is just the effect of `basal.area`.

Note that it is tempting to add confidence intervals for the regression lines plotted on these partial residual plots (e.g. using the facilities of *ggplot2* package). However, computing these confidence intervals is harder that it may first appear. To do so, it is advised to use bootstrapping techniques to get the confidence intervals for the predicted values (type `?predict.merMod` for more information). This procedure can be computationally expensive and for that reason was omitted here. Is worth mentioning that usually on partial residual plots a smoother is also presented along the regression line so that model deficiencies (non-linearities) can be spotted.

Partial residual plots are well suited to report the effect of the model predictors in scientific publications. Here, the model coefficients and the partial residual plot indicate undoubtedly that, for all species, larger trees grow faster. However, if we were to use the left graph (a) for a publication, it may seem that for European larch (right lower panel) the relationship is actually negative. This would appear in contradiction with our modelling results. Indeed, the `basal.area` regression coefficient for European larch is positive.

5.3 Summary

We showed that using the approach presented in this publication, we were able to make statistical inference in essentially any situation. We may want to model non-normal data for which a reference distribution exists (e.g. count data with a Poisson distribution). Or we may even want to drop most distributional assumptions and use bootstrapping techniques. This is a real advantage over the majority of the classical asynchrony metrics that do not have a clear inferential path and solely work with continuous data. We also saw that partial residual plots are a very powerful tool to display the conditional relationship between predictors and the response variable.

6 Final remarks

In this appendix we showed how to use our approach to mode temporal and spatial asynchrony. We also discussed the many advantages and few disadvantages of this new approach.

It is important to note that to model asynchrony we propose to simply extend the models usually fitted to data to quantify the effect of biodiversity by including an interaction between species and time and space respectively. Here, we used an example where we focus on species asynchrony, nevertheless, this approach is applicable to any categorical variable of interest (e.g. functional groups). We also showed that our approach can be used to quantify asynchrony for essentially any kind of data. This implies that we can easily drop the stringent requirements made by current asynchrony metrics in terms of ecosystem functions that can be analysed. In addition, we also showed how other requirements, such as measurement discreteness can easily be dropped.

Another important aspect of the proposed approach is that it estimates temporal and spatial asynchrony in a conditional way. This means that the effect of biasing or confounding factors is controlled for. This property is of fundamental importance as ignoring these effects can lead to wrong and misleading conclusions.

Finally, it is worth mentioning that this approach greatly improves the biological information that can be extracted from diversity-function studies as it allows users to compare the effect of very different elements that influence the modelled ecosystem function. Here we compared the effect of temporal asynchrony modelled as year fluctuations with the effect of biodiversity.

In summary, the method proposed is extremely flexible as it can adapt to a huge variety of situations, and takes into account important external factors that are omnipresent in modern diversity-function studies. The strong graphical focus and the ease of biological interpretability of the results makes it a very powerful tool to better understand ecosystem functioning. All this is achieved by using, in a tailored way, modern mainstream statistical techniques.

Reproducibility

Checkpoint date was set to 2016-05-01.

```
sessionInfo()

R version 3.3.2 (2016-10-31)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Debian GNU/Linux 8 (jessie)

locale:
 [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
 [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
 [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] mvtnorm_1.0-5   ellipse_0.3-8   gdata_2.17.0    lattice_0.20-34
[5] lme4_1.1-12     Matrix_1.2-7.1  knitr_1.14

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.7   gtools_3.5.0  digest_0.6.10 MASS_7.3-45   grid_3.3.2
 [6] nlme_3.1-128  formatR_1.4   magrittr_1.5  evaluate_0.10 highr_0.6
[11] stringi_1.1.2 minqa_1.2.4   nloptr_1.0.4  splines_3.3.2 tools_3.3.2
[16] stringr_1.1.0
```

References

D. G. Altman and J. M. Bland. Statistics notes: Absence of evidence is not evidence of absence. *British Medical Journal*, 311(7003):485, 1995.

D. Bates, M. Maechler, B. Bolker, and S. Walker. *lme4: Linear Mixed-Effects Models using 'Eigen' and S4*, 2016. URL `https://CRAN.R-project.org/package=lme4`. R package version 1.1-12.

D. M. Bates. lme4: Mixed-Effects Modeling with R. *URL http://lme4. R-forge. R-project. org/book*, 2010.

A. Canty and B. Ripley. *boot: Bootstrap Functions (Originally by Angelo Canty for S)*, 2016. URL `https://CRAN.R-project.org/package=boot`. R package version 1.3-18.

J. Chamagne, M. Tanadini, D. Frank, R. Matula, C. Paine, C. D. Philipson, M. Svátek, L. A. Turnbull, D. Volařík, and A. Hector. Forest diversity promotes individual tree growth in central European forest stands. *Journal of Applied Ecology*, 2016.

D. Clifford and P. M. A. contributions by HJ Auinger. *regress: Gaussian linear models with linear covariance structure*, 2014. URL `https://CRAN.R-project.org/package=regress`. R package version 1.3-14.

D. Cox. Statistical significance tests. *British Journal of Clinical Pharmacology*, 14(3):325–331, 1982.

J. J. Faraway. *Linear Models with R.* CRC Press, 2014.

K. Gross, B. J. Cardinale, J. W. Fox, A. Gonzalez, M. Loreau, H. W. Polley, P. B. Reich, and J. Van Ruijven. Species richness and the temporal stability of biomass production: A new analysis of recent biodiversity experiments. *The American Naturalist*, 183(1):1–12, 2014.

F. Harrell. *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis.* Springer, 2015.

A. Hector, B. Schmid, C. Beierkuhnlein, M. Caldeira, M. Diemer, P. Dimitrakopoulos, J. Finn, H. Freitas, P. Giller, J. Good, et al. Plant diversity and productivity experiments in European grasslands. *Science*, 286(5442):1123–1127, 1999.

A. Hector, T. Bell, Y. Hautier, F. Isbell, M. Kery, P. B. Reich, J. van Ruijven, and B. Schmid. BUGS in the analysis of biodiversity experiments: Species richness and composition are of similar importance for grassland productivity. *PLoS One*, 6(3):e17434, 2011a.

A. Hector, C. Philipson, P. Saner, J. Chamagne, D. Dzulkifli, M. O'Brien, J. L. Snaddon, P. Ulok, M. Weilenmann, G. Reynolds, et al. The Sabah Biodiversity Experiment: A long-term test of the role of tree diversity in restoring tropical forest structure and functioning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1582):3303–3315, 2011b.

L. Held and D. Sabanés Bové. *Applied Statistical Inference.* Springer, Berlin Heidelberg, 2014.

M. Koller. *robustlmm: Robust Linear Mixed Effects Models*, 2016. URL `https://CRAN.R-project.org/package=robustlmm`. R package version 1.8.

J. Pinheiro and D. Bates. *Mixed-Effects Models in S and S-PLUS.* Springer Science & Business Media, 2006.

J. Pinheiro, D. Bates, and R-core. *nlme: Linear and Nonlinear Mixed Effects Models*, 2016. URL `https://CRAN.R-project.org/package=nlme`. R package version 3.1-128.

H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

D. Sarkar. *Lattice: Multivariate data visualization with R.* Springer Science & Business Media, 2008.

M. Tanadini and A. Hector. A modern model-based framework to compare ecosystem functioning. in prep.

M. Tanadini and Z. Mehrabi. Does no-till agriculture limit crop yields? *Nature*, in review.

D. Tilman, P. B. Reich, and J. M. Knops. Biodiversity and ecosystem stability in a decade-long grassland experiment. *Nature*, 441(7093):629–632, 2006.

S. L. Tuck, M. J. O'Brien, C. D. Philipson, P. Saner, M. Tanadini, D. Dzulkifli, H. C. J. Godfray, E. Godoong, R. Nilus, R. C. Ong, B. Schmid, W. Sinun, J. L. Snaddon, M. Snoep, H. Tangki, J. Tay, P. Ulok, Y. S. Wai, M. Weilenmann, R. Glen, and A. Hector. The value of biodiversity for the functioning of tropical forests: Insurance effects during the first decade of the Sabah Biodiversity Experiment. *Proceedings of the Royal Society B*, 283 (1844):20161451, 2016.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-PLUS.* Springer Science & Business Media, 2013.

G. R. Warnes, B. Bolker, G. Gorjanc, G. Grothendieck, A. Korosec, T. Lumley, D. MacQueen, A. Magnusson, J. Rogers, and others. *gdata: Various R Programming Tools for Data Manipulation*, 2015. URL `https://CRAN.R-project.org/package=gdata`. R package version 2.17.0.

R. L. Wasserstein and N. A. Lazar. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 2016.

R. H. Willis. Lower bound formulas for the mean intercorrelation coefficient. *Journal of the American Statistical Association*, 54(285):275–280, 1959.

S. Wood. *Generalized Additive Models: An introduction with R*. CRC press, 2006.

S. Wood. *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*, 2016. URL `https://CRAN.R-project.org/package=mgcv`. R package version 1.8-15.

Appendix C: Stability

Chapter 3

Contents

1 Introduction

The aim of this appendix is to show how stability of diversity-function experiments can be analysed using our approach. This latter is compared to other metrics used to quantify stability.

After this brief introduction, Section 2 introduces a commonly used metric of stability, the coefficient of variation (CV). Subsequently, Section 3 highlights some issues related to the use of marginal stability metrics such as the coefficient of variation. In Section 4 we discuss one of the most important assumptions made by the coefficient of variation, a specific mean-variance relationship. In Section 5 we introduce the model-based approach that we propose. After that, in Section 6 we briefly explain how to interpret biologically the results obtained with our approach and we compare that with the interpretation of CV. Section 7 briefly introduces the methods used to perform statistical inference with our approach and when using CV or related metrics. Section 8 presents an important advantage of our model-based approach, its flexibility in terms of the type of data that can be analysed. Finally, Section 9 summarises the appendix and makes some general biological claims.

2 Quantifying ecosystem stability

Stability is an important aspect quantified in the analyses of diversity-function experiments. Often, it is compared among species or among species richness levels to understand how species identity or species richness can influence this ecosystem property. Several metrics have been proposed to estimate ecosystem stability. The coefficient of variation is very likely the most popular choice although other metrics are used too (Carnus et al., 2015; Tilman, 1999; Lehman and Tilman, 2000; Loreau and de Mazancourt, 2008). CV is computed within each group of interest as the ratio of the standard deviation and the mean response multiplied by 100 ($CV = \sigma/\mu * 100$). This measures the variability of a given response variable while taking into account its mean value. The rationale for this is that groups with higher mean values are naturally expected to show higher variability.

Example 1
To illustrate how CV works we illustrate an example with simulated data. Assume that we want to compare the stability of two species, A and B, when looking at the dry biomass produced in each plot.

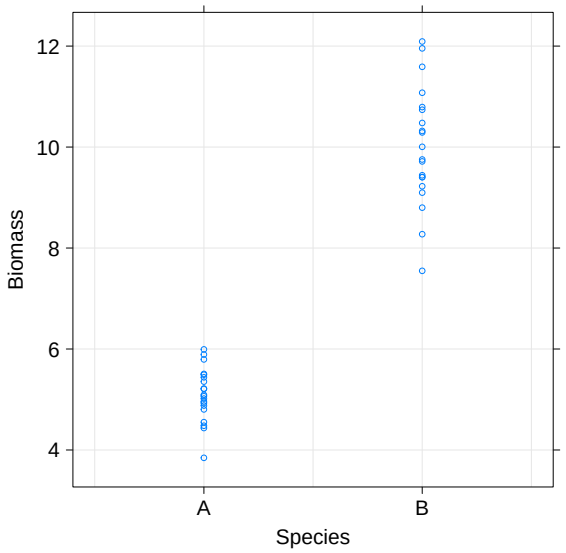


Figure 1: Simulated dry biomass of two species (A and B).

Species B, that has greater mean biomass, also shows greater variability. CV accounts for the fact that the variability increases with the mean and yields a value that enables the user to compare stability of two ecosystems differing in their mean values. In particular, CV assumes that the standard deviation increases linearly with the mean. For example, here we simulated data with the following parameters: The mean value of A is $\mu_A = 5$ and its standard deviation $\sigma_A = 0.5$, thus CV_A is $0.5/5 * 100 = 10$. Species B was simulated to have the same CV coefficient (i.e. $CV_B = 1/10 * 100 = 10$). In Section 4 we are going to discuss theoretical and practical aspects underlying the mean-variance assumption made when using CV and other similar metrics.

These metrics to quantify ecosystem stability have helped scientists to understand the effects of biodiversity on functional stability. The coefficient of variation is also used heavily in other fields (e.g. agricultural studies; Piepho (1998)). The logarithm of CV has also been indicated as a good metric to be used in meta-analyses to compare stability across studies (Nakagawa et al., 2015). Unfortunately, there are situations where CV can lead to misleading results or may miss important patterns due to its limited power. In the next section we are going to give a few examples where CV and other similar metrics work in a suboptimal way.

3 Marginality of CV and similar metrics

As we discussed in Appendix B, possibly the main disadvantage of the majority of currently-used metrics in biodiversity-ecosystem functioning studies is that they are marginal metrics. Put simply, they estimate an ecological process, here stability, implicitly assuming that no other unwanted factor can confound or bias the obtained values.

For the sake of brevity, in this section we focus on CV. However, any other marginal metric to quantify stability (e.g. 1/CV; Lehman and Tilman (2000)) is affected in the

same way by the highlighted issues. Conditional metrics, by which we mean those using a model-based approach (e.g. Carnus et al. (2015)), are discussed later in this document.

Example 2: Biasing effects (growth)

Figure 2 shows a situation where the response variable, say biomass, is increasing over time. For the sake of simplicity, we look at two species with no replication. Real experiments are obviously more complex, nevertheless, this does not alter the conclusions drawn here.

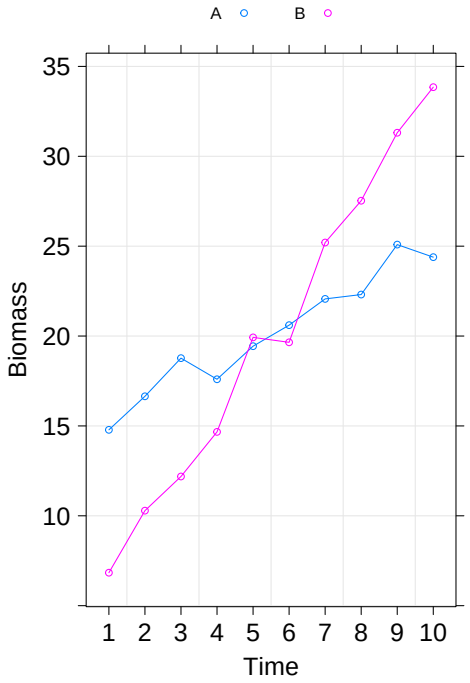
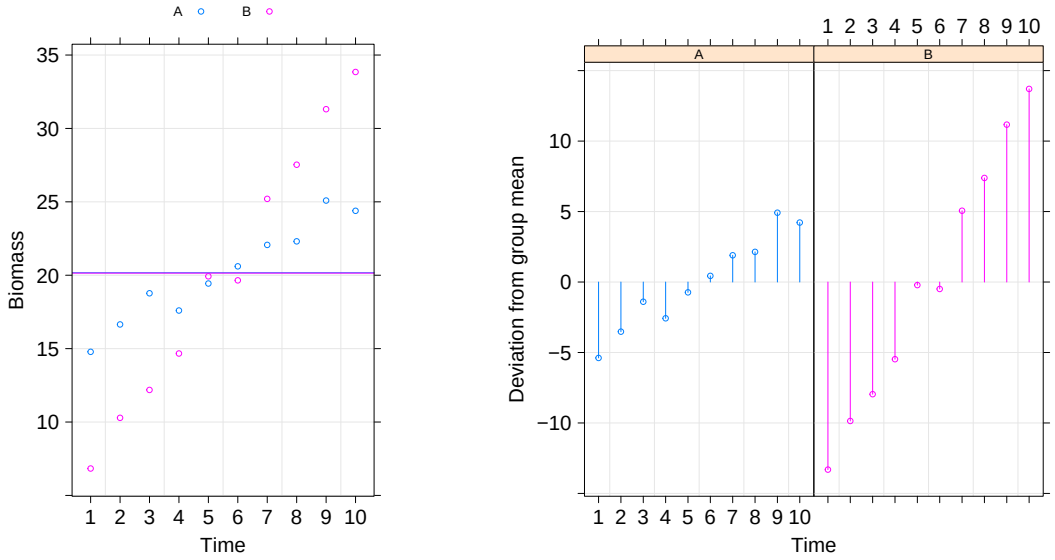


Figure 2: Biomass of two species (A and B) plotted over time.

Figure 2 shows that species B is growing faster. Data were simulated such that the average biomass of the two species was the same. The reason for that is that, for now, we want to remove the effect that different mean values have on variability. In other words, we simulate data that does not differ in mean values so that it is easier to visually compare their variability. Note that in this example we are looking at the variability of biomass over time, however, the reasoning applies to any other ecosystem function or type of stability (e.g. spatial stability).

If we were to apply blindly CV to this data we would conclude that fast growing species B is far less stable ($CV_A = 16.6$ while $CV_B = 45.63$). Indeed, CV_B is almost three times larger than CV_A. As the two species have the same mean value, CV boils down to simply measuring the standard deviation within each group. We can visualise that by highlighting the distance between each observation and the corresponding group mean.

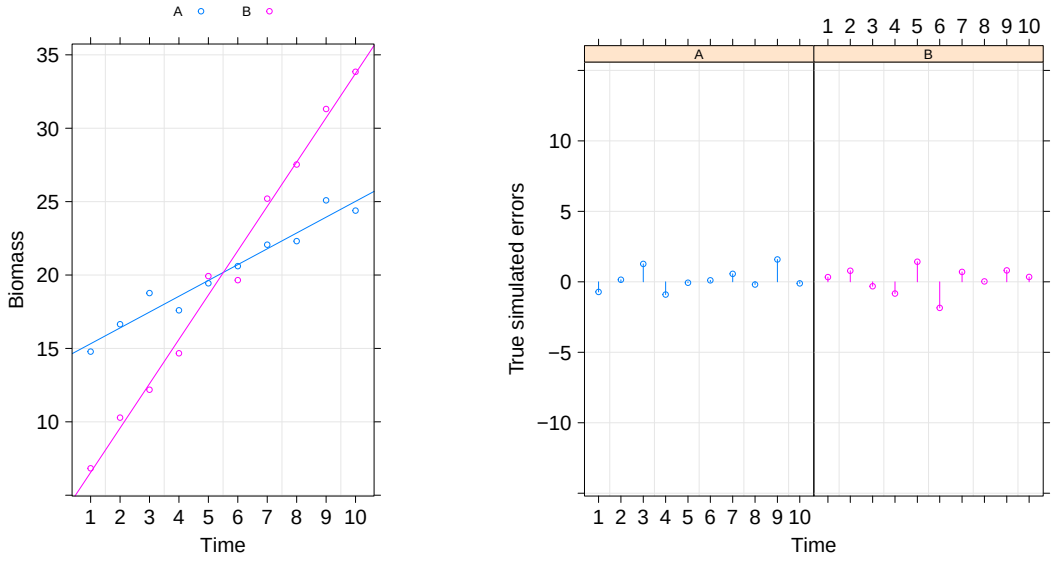


(a) Raw data and group means

(b) Deviations from average group biomass

Figure 3: Measuring CV (ignoring growth).

This figure makes it clear that CV cannot discriminate between growth and actual stability. As a matter of fact, the data was simulated such that both species have exactly the same variability, they just differ in growth. We can thus look at the actual simulated errors and see that there is no difference between species, in terms of variability.



(a) Raw data with trend modelled (regression)

(b) True simulated errors

Figure 4: Measuring CV while considering trend (growth).

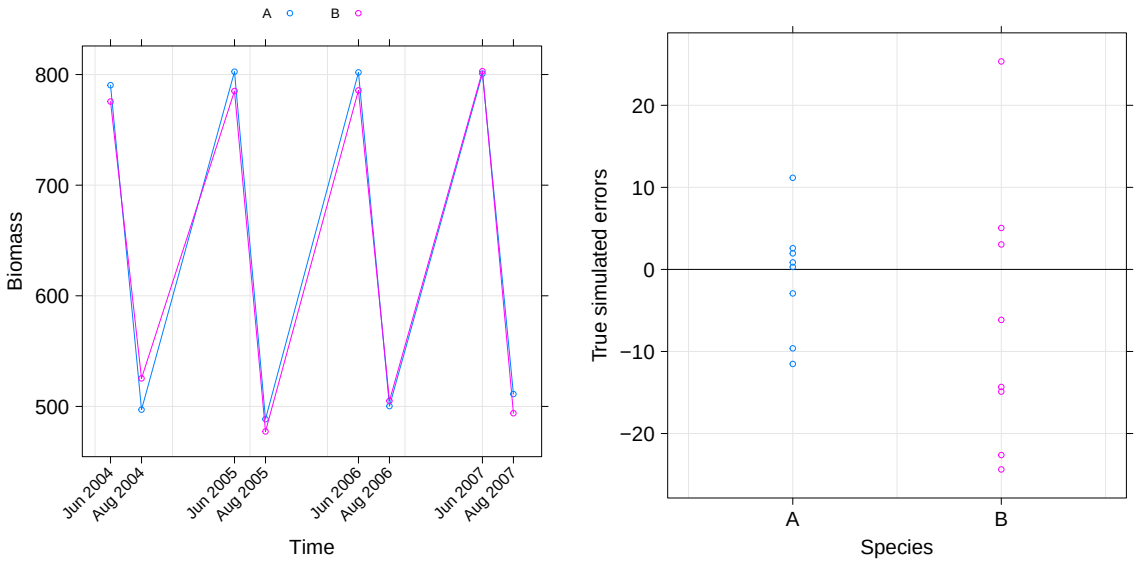
Ideally, we would apply CV after the biasing effect of growth is removed. If we were to measure CV on the actual errors we would conclude correctly that the two species barely differ in terms of stability ($CV_A = 466.48$ while $CV_B = 656.97$). Note that we

set the y-range of graph (b) to be equal in both Figures 3 and 4. This to highlight that estimating the variability after the trend has been removed is more powerful. Indeed, in this latter case, we are focusing on the real process of interest, stability. To better explain this concept, we introduce another example where where an external factor, annual fluctuations, add noise to the comparison of species stability.

Finally, it is worth noting that here we assumed equidistant time measurements. However, if time points happened to be clustered, the bias introduced by growth would be exacerbated. Fast growing species would appear to be even less stable than they really are.

Example 3: Confounding effects (time fluctuations)
Above we illustrated how using CV, that implicitly assumes a stable mean, can lead to misleading results. In that case, because of the biasing effect of growth, we greatly underestimated the stability of the fast growing species B. In other cases, assuming a stable mean where this is not the case, can confound the obtained results (without introducing bias).

For example, the data used in a related publication shows a strong human-induced variability (Tanadini and Hector, in prep.; Vojtech et al., 2008). This data shows strong temporal fluctuations that are due to the design of the experiment. In this study herbaceous mixtures were left to grow over time for four years. In each year two harvests were conducted, one in June and another in August. Figure 5 illustrates this situation. For the sake of simplicity, data were simulated, however, the parameters used to simulate data were estimated from the real data set.



(a) Biomass of two species plotted over time. No trend, but strong induced fluctuations are present

(b) Actual errors used to simulate data

Figure 5: Effect of experimenter-induced fluctuations on CV.

The left graph (a) in Figure 5 clearly shows the recurring pattern due to the harvesting scheme. Yields in August are lower than those in June simply because they take place

two months after the other (destructive) harvest. If we were to naively apply CV to this data, we would conclude that species A and B have similar stabilities. However, data were simulated such that the variability in one group is twice the other, see graph (b) on the right hand side.

To capture the difference in variability we should remove the confounding effect of the harvesting scheme. This is feasible by fitting a Linear Model, where harvest time is modelled as a categorical variable[1], and then by looking at the residuals.

There are several situations in which interventions carried out by the experimenters can induce fluctuations that add noise and even bias the real ecosystem variability. In grassland biodiversity studies these fluctuations can be introduced by, for example, weeding. Indeed, after weeding space becomes available and we may observe a boost in biomass of the target species. Analogously, in tree studies thinning may be applied (e.g. Satakunta Forest Diversity Experiment; Julia Koricheva personal communication). Thinning is likely to have a short-term positive effect on tree growth. This would artificially introduce fluctuations and thus lower ecosystem stability. Differences among species are also underestimated in that case.

Noteworthy, the issue presented here is affecting the estimation of the closely related concept of temporal asynchrony as well. Indeed, these experimenter-induced fluctuations may be of greater magnitude than the actual natural ecosystem fluctuations. This would imply that temporal asynchrony is underestimated (species appear to be respond more similarly to temporal fluctuations than they actually do).

Example 4: Decomposing stability into its constitutive components (I)
Up to this point we have been looking at stability over time. However, we may want to quantify spatial stability as well. This quantity represents the variation that, for example, species show over space. We may think of a generalist species that adapts easily to different local conditions and thus has low spatial variability (i.e. high spatial stability). On the other hand, a species with a narrower niche may respond strongly to spatial heterogeneity.

Figure 6 shows a hypothetical situation where two species react in a different manner to space. For each species, ten replicates (plots) are measured over time. Plots are highlighted by connected observations and the use of colours.

[1] Obviously, estimating and the removing the harvest effects would not be feasible here as we there is no replication and only two species. This is possible with the original data set.

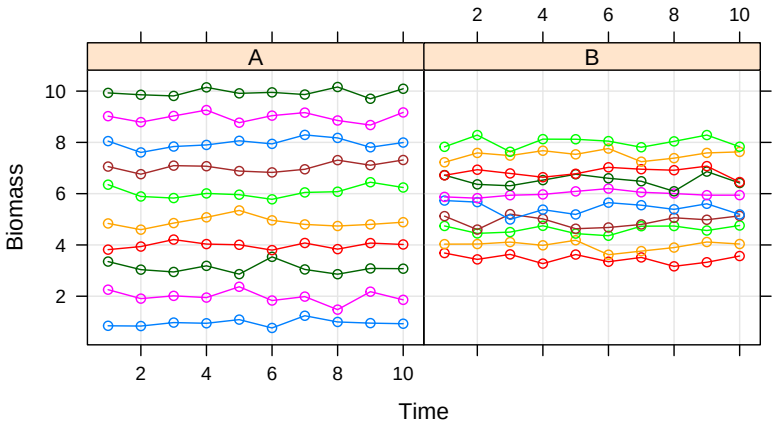


Figure 6: Two species showing different spatial stability. Plots are highlighted by colour.

Figure 6 indicates that the variability among plots is much larger in species A than in species B. This implies that species A is more strongly affected by spatial heterogeneity than species B. On the other hand, there is no indication that temporal stability differs among species. Indeed, the variability within plots appears to be similar in both species. This is indeed how data were simulated.

If we were to apply CV to plots (i.e. temporal CV), we would conclude correctly that these two species have the same temporal stability. When looking at spatial stability, we would expect that there is strong evidence for differences between these two species. Graphically, spatial CV is the variability observed when looking at each time point in the two groups separately (i.e. the variability along the y-axis in Figure 6). In other words, the vertical variability present on the graph. To quantify spatial stability we could measure CV for each species at each time point separately and then compare the ten values obtained for the two species[2].

A fundamental remark here is that by using this way to compute spatial stability, we are not telling apart the variability that comes from within and among plots. In other words, the measured spatial CV is the sum of temporal stability and spatial stability. To make this clearer we introduce another example similar to that one just discussed above (Figure 6).

Example 5: Decomposing stability into its constitutive components (II)
As we mentioned above, to measure spatial CV, we would simply estimate the vertical variability at each time point for each species (remember the effect of differing means on the variability is ignored for the moment, but this has no influence on the reasoning made here). We can now display the newly simulated data. Note that we are not highlighting plots (each species was replicated ten times), as this is not taken into account when measuring spatial CV.

[2]Alternatively, we could take plot averages and then compare their variability in the two groups. Conceptually, this leads to the same results.

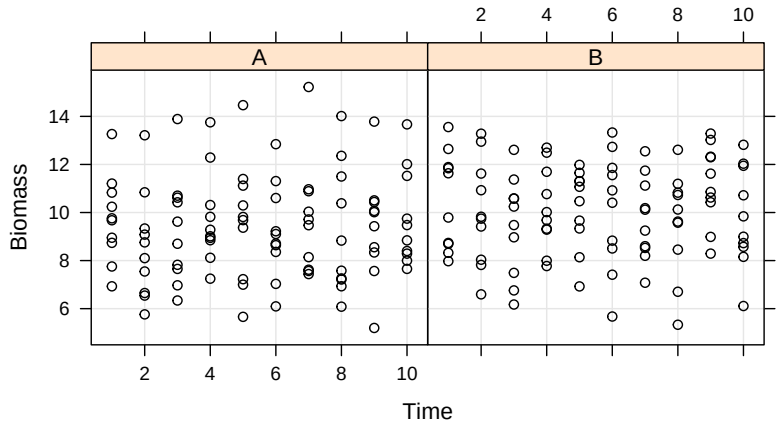


Figure 7: Two species that, apparently, show the same spatial stability.

By looking at this graph, we would conclude that these two species have the same spatial stability. Indeed, the (vertical) variability perceived does not appear to differ between the two panels. However, this is not the reality as these two species were simulated to differ in terms of spatial stability. To better understand that, we reproduce the same figure above, but with plots highlighted.

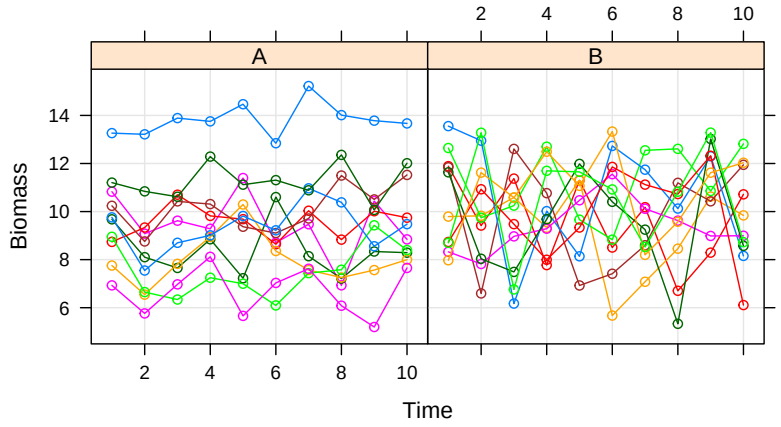


Figure 8: Two species showing different spatial and temporal stability.

Figure 8 makes it evident that species A and B differ in terms of spatial stability. Species A has low spatial stability (large differences among plots), but high temporal stability (little variation within plots), while species B shows the opposite pattern. In this example data was simulated so that the total variability measured as spatial CV is the same in the two groups.

In real settings, it is unlikely to encounter two species showing opposite behaviours in two closely related ecological processes as here. This example serves to highlight that spatial stability measured with marginal metrics such as CV is actually the sum of spatial and temporal stability.

It is important to note that this problem is exacerbated when measuring spatial stability

at higher spatial scales. Multi-site experiments such as Biodepth (Hector et al., 1999), are the most valuable to test hypotheses regarding spatial stability. Indeed, these experiments have several nested levels at which we may want to test for spatial stability. For example, there may be no evidence for differences in spatial stability in one site simply because the spatial conditions are too homogeneous. So the generalist and the more specialised species have both a fairly consistent productivity over space. However, when zooming out and comparing species productivity among sites, we may find differences in stability. Unfortunately, when comparing spatial stability at site level using CV, we are actually looking at the sum of temporal stability, spatial stability at block level and spatial stability at site level. Telling these effects apart and being able to quantify each component, is fundamental to interpreting these results accurately, biologically.

In large scale biodiversity experiments such as the Satakunta Forest Experiment (Koricheva, 2002), the Jena experiment (Weigelt et al., 2010) or the Sabah Biodiversity Experiment (Hector et al., 1999) we may want to investigate at which spatial scale the difference among species, in terms of stability, is the greatest. For this type of question it is fundamental to tell apart the variabilities measured at each spatial level. This is the fundamental idea behind the Analysis Of Variance (ANOVA) and its modern extensions such a Linear Mixed Effects Models with categorical variables. Our approach is based on these techniques and will be introduced in Section 5.

Summary
In this section we highlighted the disadvantages of using marginal metrics to measure ecosystem stability. The use of these marginal metrics is expected to underestimate the differences among groups, as well as to underestimate stability in general. In addition, spatial stability measured with marginal metrics has rather to be interpreted as the sum of spatial and temporal stability. The examples shown here made it clear that, biologically, this can have serious consequences.

4 Modelling the mean-variance relationship

4.1 Mean-variance relationship assumed by CV

In the previous section we put aside that variability is expected to grow with the mean value, by holding the mean constant in our simulations. Observations in groups with higher mean values have also larger variability and we explore that now.

Example 6
Figure 9 shows two groups with different variabilities and different mean values. CV takes into account the fact that when looking at real measurements, the variance of the observations is expected to increase with mean. Indeed, its formula divides the standard deviation by the mean value ($CV = \sigma/\mu * 100$). CV assumes thus that the variability measured as the standard deviation increases linearly with the their mean value (this assumptions is referred as the "mean-variance assumption"). If group A had standard deviation 40 and mean 10, while group B had standard deviation 30 and

mean 7.5, both groups would appear to have the same stability sensu CV (i.e. 400). The two groups shown in Figure 9 below are simulated to have exactly the same CV value.

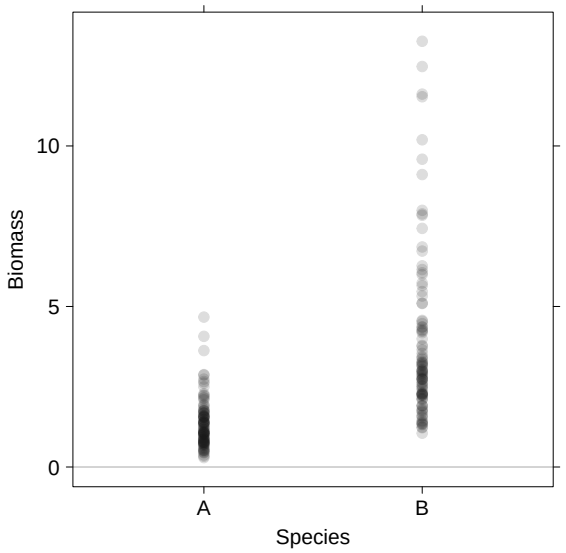


Figure 9: Simulated biomass of two species (A and B). Both mean and variability differ between groups. However, stability (measured with CV) is the same. Note that transparency is used to reduce overplotting.

The question that arises naturally when looking at the mean-variance assumption made by CV, is whether this is a sensible choice in reality. As a matter of fact, another similar metric was proposed, ($CV2 = \sigma^2/\mu * 100$; Tilman et al. (1998)). This latter assumes that the variance, and not the standard deviation, increases linearly with the mean.

Before answering this question, we look at the data simulated in the two groups above. These data have a property that is very often encountered when analysing real data sets in ecology: They are skewed. The distribution of both groups is asymmetric with few observations with large values and many observations with low values grouping together. This is typical behaviour of continuous response variables that can take only positive values such as weight or height. The usual way to cope with skewness of the data is to log-transform the measured values. The use of the log-transformation was introduced by possibly the well-known statistician, John Tukey (Mosteller and Tukey, 1977). This transformation is strongly encouraged for both the analysis and the display of "physical measurements" (Venables and Ripley, 2013; Limpert and Stahel, 2011; Sarkar, 2008).

When fitting a Linear Model to log-transformed data, we assume normality of the errors. This means that we implicitly assume that the original untransformed response variable follow a log-normal distribution. Interestingly, the coefficient of variation is a concept that was originally borrowed from the field of statistics where it is used to quantify the relative variability of log-normal random variables[3].

[3]CV was likely borrowed from agricultural studies first and then subsequently from biodiversity-ecosystem functioning studies (Smith et al., 2005).

To summarise this subsection, CV is an appropriate measure of relative variability when the data follow a log-normal distribution. Many response variables measured in diversity-function studies can be expected to follow such a distribution. It is worth mentioning that using CV on the original scale or simply measuring the standard deviation in the log-transformed scale is fully equivalent. Graphically, it is much easier to compare groups in the log-scale as we can directly interpret the observed variabilities. We can thus reproduce Figure 9 in the log-scale to see that the two groups have indeed the same variability.

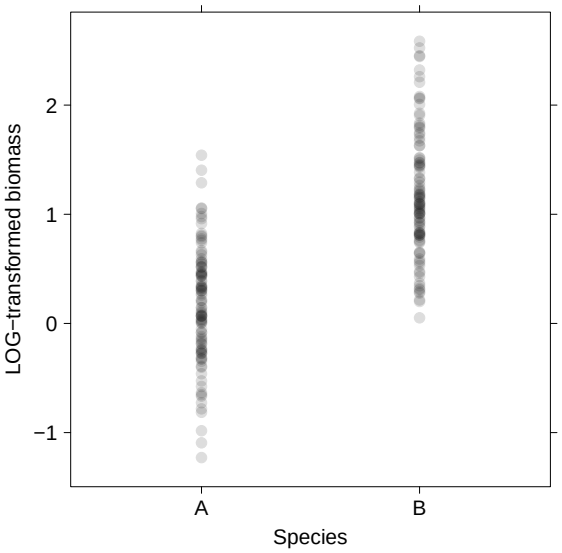


Figure 10: Simulated biomass of two species (A and B). The response variable has been log-transformed.

This graph confirms that these two groups have the same (relative) variability. Note that the equivalence between CV and the variability of log-transformed values does not limit itself to the natural logarithm, but extends to any logarithm.

4.2 Mean-variance relationship for other types of data

Although many response variables analysed in diversity-function studies are continuous variables that can only take positive values, there are other types of response variables. For example, we may be interested in the protective effects of biodiversity against pathogens. We may want to measure the number of pathogen galls on leaves, hence count data (Muiruri and Koricheva, 2016). Alternatively, we may want to quantify the pathogen load as seen in the leaf damage, hence a percentage (Vehviläinen and Koricheva, 2006; Muiruri et al., 2015).

Count data

Count data are often modelled with the Poisson distribution (Faraway, 2005; Venables and Ripley, 2013). This distribution is fully characterised by a single parameter (λ) which describes both the mean and the variance (the variance is assumed to equal the

mean). In other words, if the average of a group of observations is 20, then its variance is expected to be 20 too. This obviously means that the variance increases linearly with the mean[4]. Therefore, using CV on Poisson-distributed data leads to misleading results. In particular, the stability of highly productive ecosystems is overestimated.

On the other hand, CV2 does assume this particular mean-variance relationship and is therefore appropriate for Poisson-distributed data. In analogy to what we have done with data that follows a log-normal distribution, we may want to apply an appropriate transformation to count data so that we can directly compare the variabilities with a graphic. For log-normal data we used the log-transformation, this latter is indeed the so called "variance stabilising transformation" for log-normal data (Held and Sabanés Bové, 2014). In other words, this is the transformation that stabilises the variance so that it does not depend on its mean any more. These transformations are based on solid theory and can be derived for any type of data (i.e. for any distribution; Held and Sabanés Bové (2014)). For Poisson distributed data the appropriate transformation is the square root (Held and Sabanés Bové, 2014). Thus, an appropriate way to measure stability with count data is to use CV2 or to square-root-transform raw data and then estimate standard deviations.

Binomial data or proportions

We may be interested in quantifying the stability of response variables that are assumed to follow a binomial distribution or represent a proportion. The mean-variance relationship of both these two types of data is assumed to have a bow-shaped form. Indeed, the variance is small for small values and large values (around 0% and 100%), but large at values around 50%[5]. This is also what we often see in practice with little variation in groups close to 0% and 100%, and greater variability for groups close to 50%.

Both CV and CV2 would give misleading results for this type of data. Indeed, both would greatly overestimate the stability of groups that have proportions close to 100%. We are not aware any stability metric for this kind of data, however, this could easily be developed. This is anyway very inefficient as it would lead to a large number of data-specific metrics, which values would not be comparable among them.

Alternately, we can also apply the corresponding variance stabilising transformation to these data and estimate standard deviations. For binomial data (and proportions) this latter is the arcsine-square-root transformation ($arc - sinus(\sqrt{\%})$; Held and Sabanés Bové (2014)).

Normal data

We have seen that many currently measured ecosystem functions can be assumed to follow a log-normal distribution, however, if data is averaged, data would then likely follow a normal distribution. For example, the researchers of the Jena Experiment measured grass height ten times on a transect within each plot (Weigelt et al., 2010). These data were then averaged. If the number of measurements used to compute a

[4]Count data are very often considered to be overdispersed. This fact does not alter the reasoning made here.

[5]In particular, the variance of a binomial random variable has a $p * (1 - p)$ form, where p is the probability.

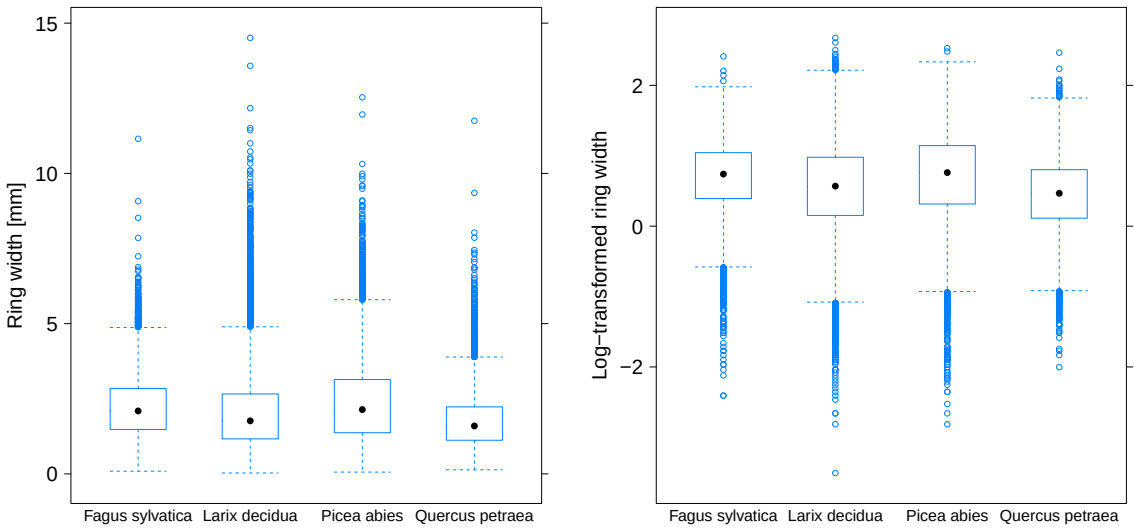
mean value is large enough, we can expected this latter to converge towards a normal distribution.

In this situation CV would give misleading results and overestimate the stability of the ecosystems with high productivity (tall grasses). In this case, we should simply compare the standard deviations without the need to divide them by the mean value to obtain an appropriate measure of stability.

4.3 Mean-variance relationship in practice: The Czech Republic data

In the previous subsection, we discussed how to appropriately quantify stability for those that can be considered the most common types of data (Faraway, 2005). If data were to follow one of these distributions, we could easily quantify stability and draw solid conclusions by using the appropriate variance-stabilising transformation. However, in reality data does not always fit into one of these discrete classes of distributions and may thus have a different mean-variance relationship. In that case assuming the wrong mean-variance relationship can lead to misleading results.

In Appendix B ("Asynchrony") we used extensively the Czech Republic data set to quantify asynchrony. The response variable modelled was `ring`. This latter represents the annual growth measured as tree ring widths. This variable can be considered to be a "physical measurement" and is therefore expected to follow a log-normal distribution. As a consequence, CV would be an appropriate measure of stability. We display the raw values along with the log-transformed data with boxplots. In this study we are interested in comparing the stability of the four tree species.



(a) Raw data

(b) Log-transformed data

Figure 11: Annual ring growth plotted against species identity.

Graph (a) on the left of Figure 11 indicates that `ring` could indeed follow a log-normal distribution (right skewness). If this was correct, then graph (b) on the right hand side

of this figure, which is in the log-transformed scale, is supposed to correctly illustrate the relative stability of each group. However, any conclusion on stability is based on the assumption that the transformation truly stabilised the variance (i.e. the effect of the mean on the variance is removed). If this was the case, the graph on the right would display normally distributed data[6]. By looking carefully at graph (b) we note that all four groups show a slight (right) skewness. Indeed, the bottom tails of these distributions appear to be slightly longer. As the normal distribution is symmetric, this pattern is unexpected.

At this stage is not entirely clear whether the log-transformation stabilised the variance. The unexpected skewness of the transformed response variable indicates that this was not the case. But how to check this in a more formal way? The answer is with "residual diagnostics plots". This may sound unexpected, however, it is not. Indeed, when fitting a Linear Model or any of its extensions we make assumptions about the mean-variance relationship. For example, when fitting a Linear Model, we assume that the variance does not depend on the mean value as it is stable (homoscedasticity). To test this assumption diagnostics plots are produced with the residuals of the fitted model (Faraway, 2014, 2005).

Inspecting the mean-variance relationship

To inspect the mean-variance relationship assumption, two graphs are used, the "Tukey-Anscombe plot" and the "scale location plot"[7].

We produce here the two plots used to inspect the mean-variance relationship with the model fitted to the Czech Republic data set in Appendix B. To do that, we refit the corresponding model.

```
mod.cz <- lmer(log(ring) ~ species * (density.stand + diversity.stand +
                                      date.fac + age + basal.area) +
                 (1 | tree) + (1 | site) + (1 | species:site),
               data = d.20)
```

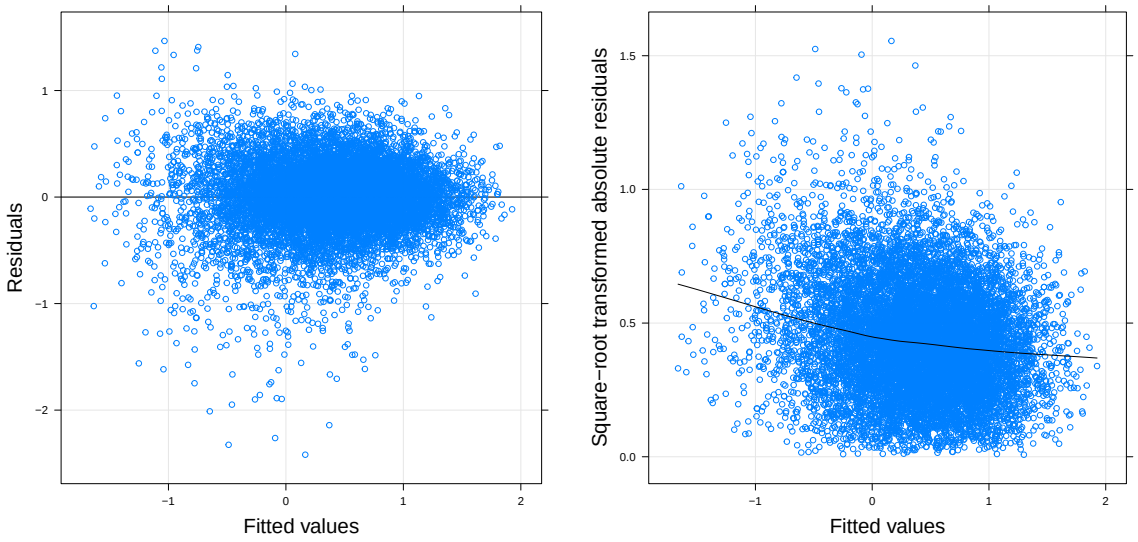
After fitting the model, we produce the two diagnostic plots. The Tukey Anscombe plot is a scatter plot of the residuals plotted against the fitted values ($\hat{\varepsilon} \sim \hat{y}$). On this plot a line at zero is added to spot asymmetries of the data. The scale-location plot displays the square-root-transformed absolute values of the residuals against the fitted values ($\sqrt{|\hat{\varepsilon}|} \sim \hat{y}$). On this plot a smoother is added to spot a non-constant variance. If the variance of the residuals is stable (i.e. does not depend on the mean; homoscedasticity), the smoother on the sale-location plot will be a flat horizontal line.

```
## 1. Tukey-Anscombe plot (a)
xyplot(resid(mod.cz) ~ fitted(mod.cz),
       ylab = list(label = "Residuals", cex = 1.5),
       xlab = list(label = "Fitted values", cex = 1.5),
       type = c("p" ,"g"),
       abline = 0)
```

[6] CV implicitly assumes that data is log-normal. By definition by taking the logarithm of this kind of data, we get normally distributed data.

[7] These two plots, along with the "leverage plot" (to spot influential observations) and the "quantile-quantile plot" (to check normality) are fundamental to check the assumptions of a Linear Model (Faraway, 2014).

```
##
## 2. Scale-location plot (b)
xyplot(sqrt(abs(resid(mod.cz))) ~ fitted(mod.cz),
       ylab = list(label = "Square-root transformed absolute residuals", cex = 1.5),
       xlab = list(label = "Fitted values", cex = 1.5),
       type = c("p", "g", "smooth"),
       col.line = "black")
```



(a) Tukey-Anscombe plot

(b) Scale location plot

Figure 12: Residual diagnostics for the Czech Republic model.

Both graphs (a) and (b) in Figure 12 indicate that the the variance is not stable. In particular, the smoother on plot (b) shows that the variability of the residuals is slightly decreasing with the fitted values. Thus, the data does not seem to follow a normal distribution after log-transformation (i.e. the raw untransformed data is not log-normally distributed). Therefore, with these data, comparing species stability with CV would not be appropriate. Analogously, we cannot simply look at the residual variability in each species group and compare them. Indeed, if a species was to be associated to higher fitted values (right hand side of plot (b)), we would expect to see smaller variability in residuals (and vice-versa).

Fortunately, we have an estimate of the mean-variance relationship. Indeed, the smoother on the scale-location plot (b) is modelling this relationship. Therefore we can use this information to compare species. The easiest way to do this graphically is to highlight species on the scale location plot. Because we are interested mostly in the mean-variance relationship, we highlight smoothers.

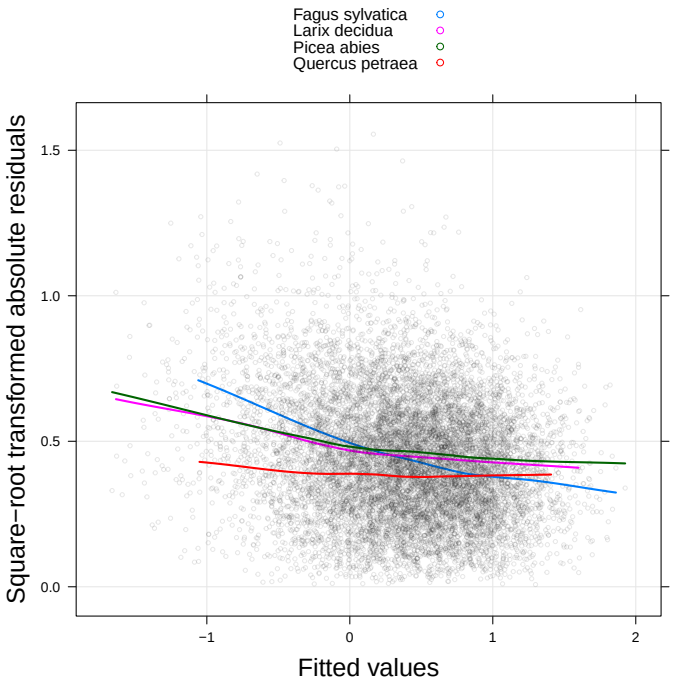


Figure 13: Scale-location plot for the Czech Republic model (mod.cz). Species-specific mean-variance relationships are highlighted with smoothers.

Figure 13 shows that all four species have a similar mean-variance relationship. In particular, for all groups the variance decreases with the fitted values. On top of that there are no huge differences among species. Indeed, if a species happened to be far less stable, its residuals would be much larger, and its smoother would run well above the others. The reason that stability measured with residuals does not show any striking pattern is very likely that the model used contains terms that model spatial and temporal stability already. Indeed, our model contains an interaction term `species:date.fac`, as well as an interaction term `species:sites`. Therefore, what we are looking at here, is the remaining stability that is not accounted for by space or time(i.e. "intrinsic" species stability).

Nevertheless, there are some interesting patterns that we may want to inspect further. For example, the sessile oak appears to have, on average, smaller residuals (i.e. its smoother (red) is always below the others). This implies that this species is more stable. Interestingly, the two conifers have almost the same mean-variance relationship. Finally, European beech appears to be the most unstable species for low predicted ring values, but the most stable for large fitted values.

However, given the large variation present in these residuals, we can safely assume that there are no great differences in stability among these four species. In Section 7 we will show how to model formally the mean-variance relationship so that p-values can be computed. This is useful, for example, if we want to test formally that the sessile oak is more stable than the other species. In Section 6 we will discuss further how the results of this graphical approach can be interpreted and reported. This graph can be produced using other groups. For example, we may want to look at the effect of species diversity on stability.

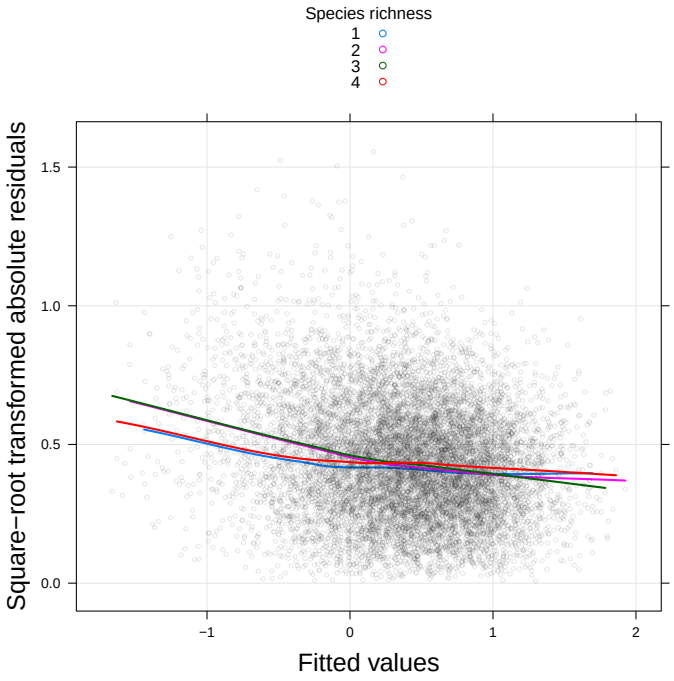


Figure 14: Scale-location plot for the Czech Republic model (mod.cz). Mean-variance relationships are highlighted for different groups of species richness with smoothers.

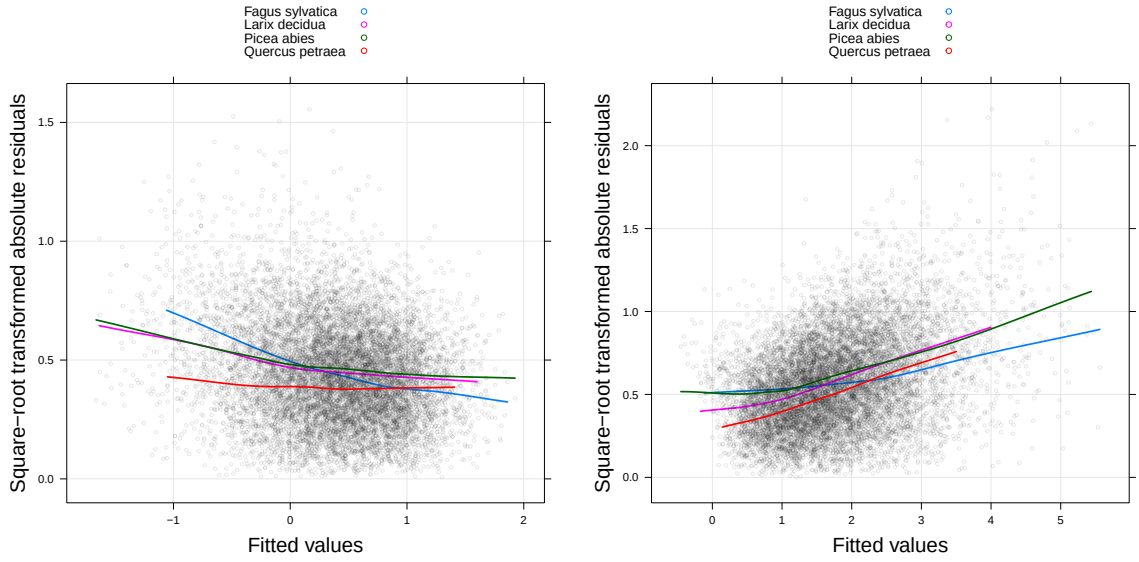
When looking at the differences in stability among species richness levels, we don't see any practically relevant difference.

CV and other related metrics make strong assumptions on the mean-variance relationship. In addition, the relationship is assumed to be identical for all groups inspected (here species). To make a graphical analogy, CV assumes that the smoothers on Figure 14 are straight lines that run parallel to the x-axis.

Our graphical approach drops all these assumptions and estimates the mean-variance relationship separately for each group in an objective way. Obviously, this may complicate the interpretation of the results as it may not be possible to state which species is the most stable as this, as shown by European beech, can vary. However, this should not seen as an hindering factor, as it represent what the data actually tell us.

Invariance to transformations of the graphical approach

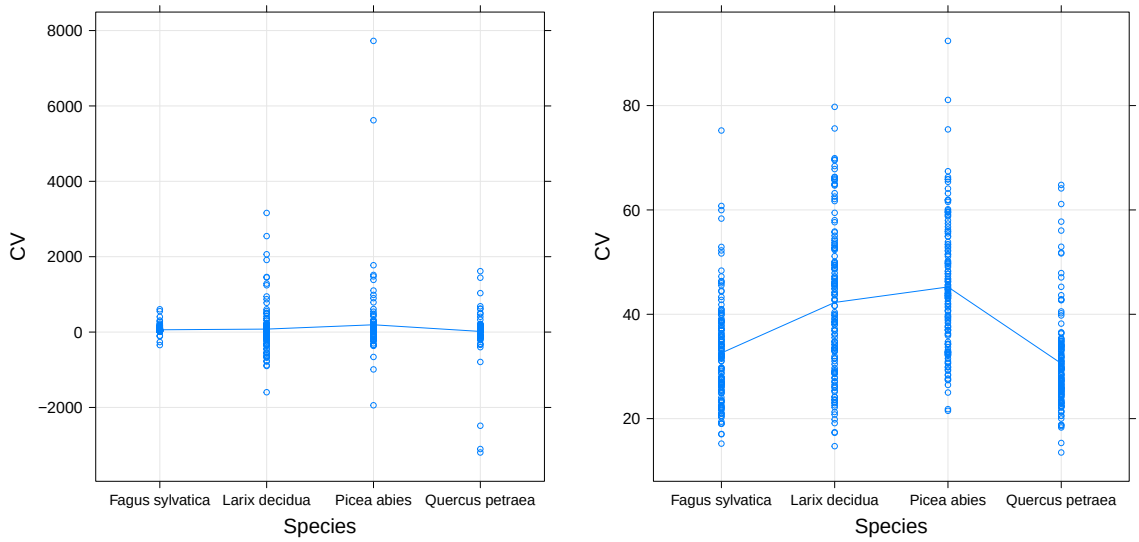
An advantage of the graphical approach is that it is invariant to transformation applied to the modelled response variable. For example, we saw that taking the logarithm of the `ring` does not perfectly stabilise the variance. Therefore, we may prefer to fit a model to the untransformed response variable so that the results (e.g. regression coefficients) are easier to interpret. If we were to do that and apply our graphical approach to estimate stability to the untransformed data set, we would expect to reach the same conclusions. Figure 15 illustrates this situation.



(a) Log-transformed response variable (b) Untransformed response variable

Figure 15: Scale-location plots for the tranformed response variable and the orginal reponse variable (ring).

Figure 15 shows that the graphical approach leads indeed to the same conclusions. Obviously, the variance is decreasing in one case and increasing in the other. However, the differences among groups remain virtually the same. On the other hand, if we were to apply CV to both response variables we would not obtain agreeing results. The reason for that is that when we transform data, we are modifying the mean-variance relationship. Below we show the estimated CV values obtained in the two cases.



(a) Log-transformed response variable (b) Untransformed response variable

Figure 16: CV for the transformed response variable and the orginal reponse variable (ring).

Figure 16 shows undoubtedly that CV is not invariant to transformations that affect the mean-variance relationship. Changing the scale of the response variable (e.g. going from grams to kilograms) has no influence on CV. However, taking the square of a response variable (e.g. modelling diameter or basal area) does modify the mean-variance relationship and thus gives different results.

Taking the precision of the observations into account
When modelling the response variable `ring` we concluded that the raw data is not normally distributed (as the variance increased with the mean), but at the same time the data did not appeared to be log-normally distributed (as the variance of the residuals decreased with the mean). These data can thus be considered to lie somewhere between these two distributions. In reality, we note that this does not come unexpected as the response variable modelled here (`ring`) is not a direct physical measurement, but is actually the average of two cores performed on each tree. This well explains the hybrid behaviour observed when looking at the mean-variance relationship.

In this study, it was not possible to use both cores of all trees (see Chamagne et al. (2016) for details). Indeed, for about 6% of the trees only one core was used. This could imply that these measurements are less precise and reliable than those based on two cores. As a consequence, they are expected to be associated with larger residuals. We can formally inspect this hypothesis by estimating the variance of the residuals in the two groups. Not unexpectedly, the observations based on one core have a variance which is double the one of the observations based on two cores ($\hat{\sigma_\varepsilon}^2 = 0.038$ and $\hat{\sigma_\varepsilon}^2 = 0.019$ respectively).

When measuring stability this type of information is fundamental. CV and similar marginal metrics cannot include this information when estimating stability. There are situations where this can have dramatic consequences.

For example, Vojtech and colleagues ran a biodiversity experiment with five perennial herbaceous species where the size of the harvested area varied over time (Vojtech et al., 2008). Due to logistical constrains the harvested area was reduced from 0.25 m^2 to 0.9 m^2 in the last surveys. Tanadini and colleagues showed that by taking this information into account the results concerning transgressive overyielding changed significantly (Tanadini and Hector, in prep.). If we were to apply CV to that data set, we would wrongly conclude that in the last phase of this experiment, where sampled surfaces were smaller, the stability of all ecosystems decreased significantly.

5 Model-based framework

Possibly the most important aspect of the comprehensive approach to analyse diversity-function experiments presented in this publication is the use of a model-based approach. Instead of running several disconnected analysis we propose to fit a single meaningful model where all biological aspects of these analyses are considered. When looking at stability this approach has the great advantage that we can remove the effect of biasing factors (e.g. growth), but also that we can decompose stability into its constitutive elements and estimate them simultaneously.

In Example 5, presented at page 9, we highlighted how CV, being a marginal metric, cannot tell the effects of spatial and temporal stability apart. Below we show the data again.

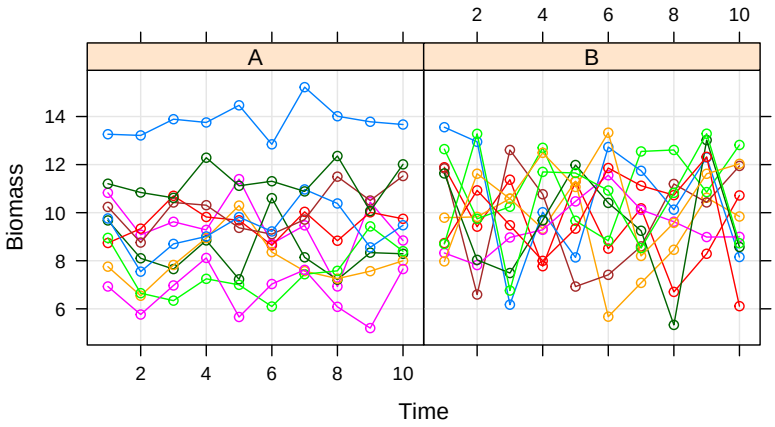


Figure 17: Two species showing different spatial and temporal stability.

Here we are going to fit a model that tells spatial and temporal variability apart. Distinguishably, this approach can also be used to tell other elements of stability apart. For example, in grassland biodiversity studies, we may want to estimate stability at population level (i.e. species level) and at community level (i.e. plot level). We fit a Linear Mixed Effects Model to the data introduced in Example 5.

```
mod.ex.5 <- lmer(y ~ time + species + (1 | plot), data = d.example.5)
```

The model was fitted with `time` and `species` as fixed effects and `plot` as a random effect.

We start by visualising the variability within plots. To do this, we use the residuals of the fitted model.

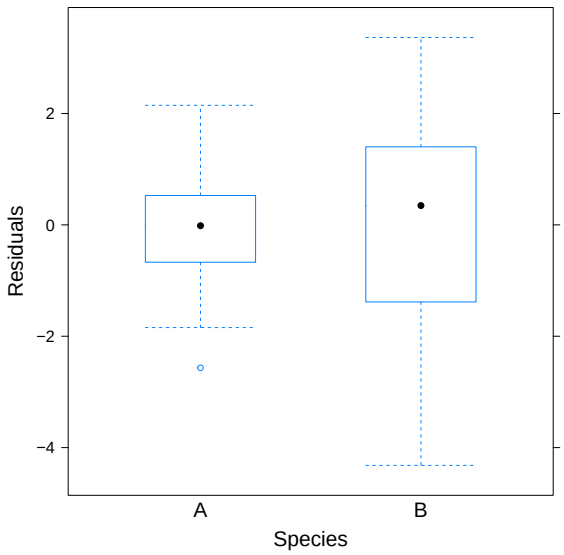


Figure 18: Variability of the residuals of species A and B (mod.ex.5).

As expected the variability of the residuals in species B is much larger than in species A. Note that here we visualised variability by simply plotting the residual of the observations of the two species.

We now look at the variability among plots, which in this example can be considered to represent spatial stability. To estimate this stability for both species we look at the predicted values of the random effect plot (i.e. the so-called conditional modes; Bates (2010)).

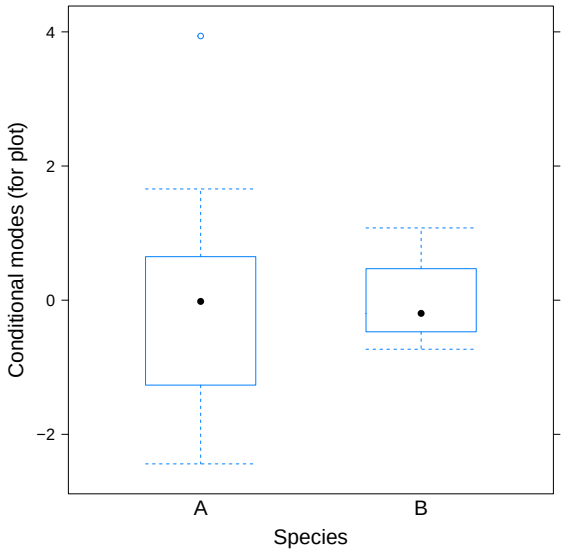


Figure 19: Variability of the conditional modes of species A and B (mod.ex.5).

Figure 19 indicates that the variability among plots in species A is larger than in species B. This confirms what we saw when visualising the raw data (Fig. 17).

The data used in Example 5 was simulated as follows: Both groups (A and B) had a mean biomass of 10. Among-plots variation for A $\sigma_{plot,A} = 2$ and $\sigma_{plot,B} = 1$. Within-plots variation for A $\sigma_{\varepsilon,A} = 1$ and $\sigma_{\varepsilon,B} = 2$. This is exactly what we observe when visualising the conditional modes and residuals of the two species.

6 Interpretation of the obtained values

6.1 Biological interpretation of the estimated variances of random effects

We concluded the previous section by listing the estimated standard deviations for the residuals of each species. This is indeed, what our graphical method visualises. A natural question that arises here is whether we can interpret these numbers biologically. The answer is "Yes!".

As the among-plots variability for species A is $\sigma_{plot,A} = 2$. Thus, we can claim that 95% of the plots of species A are expected to be contained within an interval of $-1.96 * 2$

and $+1.96 * 2$ around the species average. Roughly speaking, the vast majority of the plot averages lie in the interval 10 ± 4. We add this information to Figure 17 (only the panel for species A is shown).

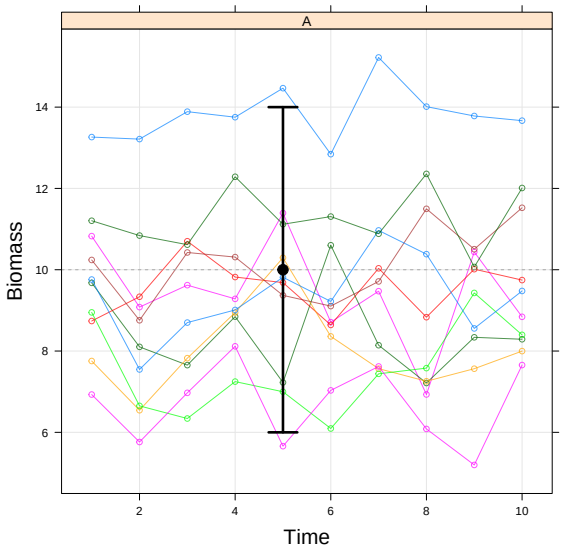


Figure 20: Raw data for species A (example 5). The dashed horizontal line is the species average, while the black vertical line is the 95 percent confidence interval for plot averages.

The confidence interval plotted on Figure 20 highlights the one-to-one relationship between what we see on graphs and stability estimated through standard deviations (Sarkar, 2008).

We can repeat the same procedure with the residual standard deviation. For species A this was 1, therefore, we can claim that the vast majority of the observations within each plot are to be found in an interval ± 2 around the plot mean. In other words, we do not expect to find many observations falling outside the ± 2 interval around the plot average.

It is worth mentioning that standard deviations can also be compared with estimated fixed effects. For simplicity we look at the estimates of model `mod.ex.5` where both, residual standard deviation and plot standard deviation are assumed to be the same in both species[8].

```
fixef(mod.ex.5)

(Intercept)        time    speciesB
      9.327       0.017       0.643

##
VarCorr(mod.ex.5)

 Groups   Name        Std.Dev.
 plot     (Intercept) 1.44
```

[8]It is possible to fit more complex models, where these parameters are estimated separately for each species Bates (2010).

```
Residual             1.54
```

The difference in averages between the two species is 0.643 (**R** uses treatment contrasts by default), while the (pooled) standard deviation for plots is 1.44. This means that, we expect plots averages to be within a $\pm 1.96 * 1.44 \approx \pm 3$ around the species mean. The difference between the two species is small compared to that interval, we can thus conclude that the variability among plots is of greater importance than the difference in averages. This qualitative reasoning are helpful to comment biologically the results of such models. Interestingly, overyielding can be compared with spatial asynchrony.

We conclude this section mentioning that here we used a simple (simulated) design, with only two levels (within plot and among plots). Nevertheless, present-day experiments can be much more complicated with several nesting levels. We may want to estimate stability at each level simultaneously without confounding the stability of each level with the ones below. This technique allows us to estimate stability at different level of organisation (organism, population, community).

6.2 Biological interpretation of CV values

On the other hand, CV cannot be directly compared with fixed effects estimates such as the difference among two species or overyielding of the mixture over the monocultures. In addition, the link between CV values and what is seen on graphics is not so straightforward. Obviously, CV values are only meaningful when the data shows a linear relationship between the mean and the standard deviation.

7 Formal comparison of variances

7.1 Formal inference for estimated variances

The concept of comparing ecosystem stability through the comparison of residual variances is not new. Indeed, this approach has been used for decades in agriculture (Smith et al., 2005), and was more recently introduced in the context of diversity-function experiments (Carnus et al., 2015). Here we simply extended this idea to random effects (and to fixed effects), so that stability can be modelled at various levels.

In practice, Carnus and colleagues (2015) proposed to use modern Linear Mixed Effects software such as the add-on package *nlme* (Pinheiro et al., 2016) to formally estimate different residual variance for the groups of interest (e.g. species). If the variance is stable (as seen in the scale-location plot), this approach is very powerful as it enables us to compare the stability of different groups and even perform statistical inference (i.e. estimate p-values). However, if the variance is not stable, estimating a separate residual variance for each species can be misleading. If data were log-normal, then the species associated with higher mean values (larger productivity) would display larger variability. However, this does not imply that these species have a lower relative stability.

Using packages such as *nlme* it is possible to include a non-stable residual variance in the model and to model that separately for each species. However, this way of proceeding can quickly lead to complicated models that are hard to fit and to interpret. This is indeed, one of the main reasons that the authors of the successor `lme4` package dropped some of these functionalities (Bates et al. (2014), Martin Maechler personal communication).

7.2 Formal inference with CV values

To conclude this section we make a few important remarks about making statistical inference with estimated CV values. At page 19 we displayed the CV measured at tree level for the four species of the Czech Republic study (Fig. 16). To compare whether species really differ in terms of stability we may want to fit an Linear Model where the estimated CV value is the response variable and species the predictor. There are a few important issues to consider when doing that.

To start, it has been shown theoretically that CV does not follow a normal distribution (neither does log(CV)) (Hendricks and Robey, 1936). In addition, we may need to adapt our model based on actual design elements of the study. In this example, trees used to estimate CV values cannot considered independent as they are grouped in sites. So we may want to extend our model by incorporating a random effect `site`. On top of that, trees differ greatly in terms of age and site density, and these variables may affect stability. Trees may also differ in number of observations. So we would end up fitting a complex model when, actually, all this information has already been modelled when looking at productivity or asynchrony. In addition, all variables that influence stability and vary over time (e.g. annual fluctuations) cannot be used in this latter model. Indeed, in this case CV aggregates the data over time. Thus, all predictors that vary over time cannot be considered.

Finally, it worth mentioning that CV is based on the estimation a standard deviation. In some circumstances the sample size of groups may not be enough to estimate a standard deviation. Consider a situation where most trees have two or three observations, but unfortunately some have just one single observation. When using CV we must drop all trees observed only once (as it is not possible to compute sd based on one observation). On the other hand, when using a model-based framework, each observation is associated with a residual, and can therefore be used to estimate residual variability.

8 Extending the Linear Mixed Effects Model

8.1 Non-Gaussian data: Generalised Linear Mixed Models

When dealing with non-normal data for which a distribution is implemented in the software used, we do not necessarily need to use the variance stabilising transformation. Indeed, we can simply apply the appropriate model (e.g. a binomial model) and then use the standardised residuals and perform exactly the same graphical procedure

described above. There are several types of standardised residuals and definitions may slightly differ among software (Faraway, 2005, 2014). However, most software returns, by default, standardised residuals when Generalised Linear Mixed Models are fitted (e.g. Bates (2010)).

8.2 Generalised Additive Models

In analogy with what we have seen in Appendix B ("Asynchrony") we may want to use Generalised Additive Models to drop discreteness assumptions. In that appendix, we proposed to use a smooth surface to model spatial heterogeneity found in the large-sized plots of the Sabah Biodiversity Experiment. Indeed, these plots have a size of 200 metres by 200 metres. We mentioned that trees seem to react to spatial heterogeneity at a much lower scale (i.e. 20 metres). Therefore, it would be misleading to consider the conditions within these large plots to be homogeneous. In other words, fitting a Linear Mixed Effect Model where space is modelled with the random effect `plot` would not be sufficient. From the biological point of view if we were to ignore this information, we could miss that spatial asynchrony is actually present as we are looking at the wrong scale (i.e. we should look at a finer scale).

The results of Generalised Additive Models can be used to quantify stability as well. Indeed, if for the Sabah example we were to fit a smooth surface to each plot, we could use their estimated degrees of freedom (edf) to understand how much variation is present within plots. For example, we could think of species A that is a generalist, and copes well with a large variety of local conditions. Thus, within each plot of that species we would not see strong variation. For this species the smooth surfaces estimated in each plot would be very smooth and thus their estimated degrees of freedom low. On the other hand, species B may react more strongly to heterogeneous space and thus have a smooth surface that is more wiggly and need a larger number of estimated degrees of freedom.

9 Summary and final remarks

All marginal stability metrics discussed here estimate variability and correct these estimates for the effects of the mean. The assumptions made on the mean-variance relationship by these metrics determine the results obtained.

The coefficient of variation, which is probably the most widely used stability metric in diversity-function studies, assumes a linear relationship between the mean and the standard deviation. We highlighted that this assumption is, in theory, very likely to be fulfilled with most response variables analysed in this field. Indeed, all physical measurements such as weight or height are expected to have this mean-variance relationship.

However, it is important to note that data aggregation (e.g. averaging several measurements) or data transformation (e.g. measuring diameter but modelling tree basal area) modify the original mean-variance relationship. In addition, it can happen that real data does not follow the expected distribution (Sandrock et al., 2014).

The coefficient of variation is not an appropriate measure of stability for data such as counts, or proportions. The consequences of assuming a wrong mean-variance relationship can very serious. In particular, if data follow a normal distribution, or represents counts or proportions, then the stability of ecosystems associated with high mean value is overestimated. For proportions the overestimation of the stability of ecosystems associated with high mean values is dramatic.

We presented a graphical approach to inspect, quantify and interpret biologically stability. For physical measurements, which are expected to follow a normal distribution after log-transformation, this approach is fully equivalent to the use of the coefficient of variation.

The graphical approach presented here is inspired by the approach used in agriculture to measure the stability of crop productivity. This latter approach was recently introduced in the field of diversity-function experiments. The main difference with our approach is the focus on formal modelling of variances (our focuses rather on the use of graphs). We extended the approach used in agriculture by looking at the estimated variances of different nesting levels and not just at the lowest level (i.e. residuals). In addition, we also showed how to drop the assumptions of data normality and stable variance.

The advantages of our model-based approach over marginal metrics such as the coefficient of variation are manifold. To start with, the effect of biasing or confounding effects (e.g. growth) can be easily removed. Stability can be decomposed in its constitutive elements in a manner analogous to an ANOVA procedure, where all components are estimated simultaneously. The biological relevance of the obtained results can be easily judged with graphs or simple calculations (e.g. using confidence intervals). Our approach is usable with essentially any kind of data, and does not rely on any mean-variance relationship (this latter is objectively estimated from the data).

It worth noting that this approach is more powerful than other non model-based approaches. Differences in stabilities are likely to be currently underestimated. This is especially true for higher nesting levels of these experiments (e.g. differences in stability among plots of a grassland study). To conclude, we want to stress the fact that quantifying and comparing ecosystem stability may be harder than currently believed. Our approach is thus expected to better spot difference in stability among ecosystems and yield better estimates of variability.

Reproducibility

Checkpoint date was set to 2016-05-01.

```
R version 3.3.2 (2016-10-31)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Debian GNU/Linux 8 (jessie)

locale:
 [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
 [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
 [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] lme4_1.1-12     Matrix_1.2-7.1  lattice_0.20-34 knitr_1.14

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.7   digest_0.6.10 MASS_7.3-45   grid_3.3.2    nlme_3.1-128
 [6] formatR_1.4   magrittr_1.5  evaluate_0.10 highr_0.6     stringi_1.1.2
[11] minqa_1.2.4   nloptr_1.0.4  splines_3.3.2 tools_3.3.2   stringr_1.1.0
```

References

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting Linear Mixed-Effects Models using lme4. *arXiv preprint arXiv:1406.5823*, 2014.

D. M. Bates. lme4: Mixed-Effects Modeling with R. *URL http://lme4. R-forge. R-project. org/book*, 2010.

T. Carnus, J. A. Finn, L. Kirwan, and J. Connolly. Assessing the relationship between biodiversity and stability of ecosystem function-is the coefficient of variation always the best metric? *Ideas in Ecology and Evolution*, 7(1), 2015.

J. Chamagne, M. Tanadini, D. Frank, R. Matula, C. Paine, C. D. Philipson, M. Svátek, L. A. Turnbull, D. Volařík, and A. Hector. Forest diversity promotes individual tree growth in central European forest stands. *Journal of Applied Ecology*, 2016.

J. J. Faraway. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC press, 2005.

J. J. Faraway. *Linear Models with R*. CRC Press, 2014.

A. Hector, B. Schmid, C. Beierkuhnlein, M. Caldeira, M. Diemer, P. Dimitrakopoulos, J. Finn, H. Freitas, P. Giller, J. Good, et al. Plant diversity and productivity experiments in European grasslands. *Science*, 286(5442):1123–1127, 1999.

L. Held and D. Sabanés Bové. *Applied Statistical Inference*. Springer, Berlin Heidelberg, 2014.

W. A. Hendricks and K. W. Robey. The sampling distribution of the coefficient of variation. *The Annals of Mathematical Statistics*, 7(3):129–132, 1936.

J. Koricheva. Sustainable forestry in temperate regions. In L. Bjork, editor, *Proceedings of the SUFOR International Workshop (Lund, Sweden. April 7-9)*, pages 152–153, 2002.

C. L. Lehman and D. Tilman. Biodiversity, stability, and productivity in competitive communities. *The American Naturalist*, 156(5):534–552, 2000.

E. Limpert and W. A. Stahel. Problems with using the normal distribution-and ways to improve quality and efficiency of data analysis. *PLoS One*, 6(7):e21403, 2011.

M. Loreau and C. de Mazancourt. Species synchrony and its drivers: Neutral and nonneutral community dynamics in fluctuating environments. *The American Naturalist*, 172(2):48–66, 2008.

F. Mosteller and J. W. Tukey. *Data analysis and regression: A second course in statistics*. Addison-Wesley Series in Behavioral Science: Quantitative Methods, 1977.

E. W. Muiruri and J. Koricheva. Going undercover: Increasing canopy cover around a host tree drives associational resistance to an insect pest. *Oikos*, 2016.

E. W. Muiruri, H. T. Milligan, S. Morath, and J. Koricheva. Moose browsing alters tree diversity effects on birch growth and insect herbivory. *Functional Ecology*, 29(5):724–735, 2015.

S. Nakagawa, R. Poulin, K. Mengersen, K. Reinhold, L. Engqvist, M. Lagisz, and A. M. Senior. Meta-analysis of variation: Ecological and evolutionary applications and beyond. *Methods in Ecology and Evolution*, 6(2):143–152, 2015.

H.-P. Piepho. Methods for comparing the yield stability of cropping systems. *Journal of Agronomy and Crop Science*, 180(4):193–213, 1998.

J. Pinheiro, D. Bates, and R-core. *nlme: Linear and Nonlinear Mixed Effects Models*, 2016. URL `https://CRAN.R-project.org/package=nlme`. R package version 3.1-128.

C. Sandrock, M. Tanadini, L. G. Tanadini, A. Fauser-Misslin, S. G. Potts, and P. Neumann. Impact of chronic neonicotinoid exposure on honeybee colony performance and queen supersedure. *PLoS One*, 9(8):e103592, 2014.

D. Sarkar. *Lattice: Multivariate data visualization with R*. Springer Science & Business Media, 2008.

A. Smith, B. R. Cullis, and R. Thompson. The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *The Journal of Agricultural Science*, 143(06):449–462, 2005.

M. Tanadini and A. Hector. A modern model-based framework to compare ecosystem functioning. in prep.

D. Tilman. The ecological consequences of changes in biodiversity: A search for general principles. *Ecology*, 80(5):1455–1474, 1999.

D. Tilman, C. L. Lehman, and C. E. Bristow. Diversity-stability relationships: Statistical inevitability or ecological consequence? *The American Naturalist*, 151(3):277–282, 1998.

H. Vehviläinen and J. Koricheva. Moose and vole browsing patterns in experimentally assembled pure and mixed forest stands. *Ecography*, 29(4):497–506, 2006.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-PLUS*. Springer Science & Business Media, 2013.

E. Vojtech, M. Loreau, S. Yachi, E. M. Spehn, and A. Hector. Light partitioning in experimental grass communities. *Oikos*, 117(9):1351–1361, 2008.

A. Weigelt, E. Marquard, V. M. Temperton, C. Roscher, C. Scherber, P. N. Mwangi, S. Felten, N. Buchmann, B. Schmid, E.-D. Schulze, et al. The Jena Experiment: Six years of data from a grassland biodiversity experiment. *Ecology*, 91(3):930–931, 2010.

Appendix D: Multivariate analyses and extensions

Chapter 3

Contents

1 Introduction

The aim of this appendix is to illustrate how analyses run on different response variables, but measured in the same experiment, can be interlinked. the rationale is that we can glean new information by allowing the results of different analyses to communicate. In Section 3 we will analyse both response variables of the Sabah Intensive Experiment (tree diameter and tree survival), which data was visualised in Appendix A. This experiment was run on six plots. This information is included in both models with a random effect `plot`. For both models we can get the predicted random effects, or conditional modes, for each plot. We can correlate these values to test whether a good plot to grow is also a good plot to survive.

After this brief introduction, Section 2 uses the the data of the well-known Biodepth experiment to illustrate how conditional modes of different models can be correlated to glean new information. Subsequently, in Section 3 we use the data of the Sabah Intensive Experiment to further develop this technique. Finally, Section 4 illustrates how clustering techniques can be used to extend the asynchrony analyses shown in Appendix B.

2 Example I: Biodepth Experiment

2.1 Modelling both response variables

In this first section we are going to use the data of the Biodepth experiment (Hector et al., 2011), a multi-site experiment that aimed to inspect the effect of species richness on productivity and on stability. The experiment was run at eight European sites. In all, but one site (i.e. Portugal) the experiment was replicated with two blocks. Each has site, but Portugal, has thus two replicates of any particular mixture (species composition). Not all mixtures are present in all sites. In particular, 200 different mixtures were used in the whole experiment. However, in each site only about 30 different mixtures are to be found. Some mixtures were used in different sites. Each plot in this experiment was measured for several response variables. The Biodepth data set is freely available on-line[1].

Here we use the response variables root biomass and shoot biomass. We start by visualising and modelling both response variables. Subsequently, we correlate the results of the two models. We start by looking at a scatterplot of the two response variables. As both response variables (`root` and `shoot`) are "physical measurements" we log-transform them (Venables and Ripley, 2013).

[1] `http://www.esapubs.org/archive/mono/M075/001/suppl-1.htm`

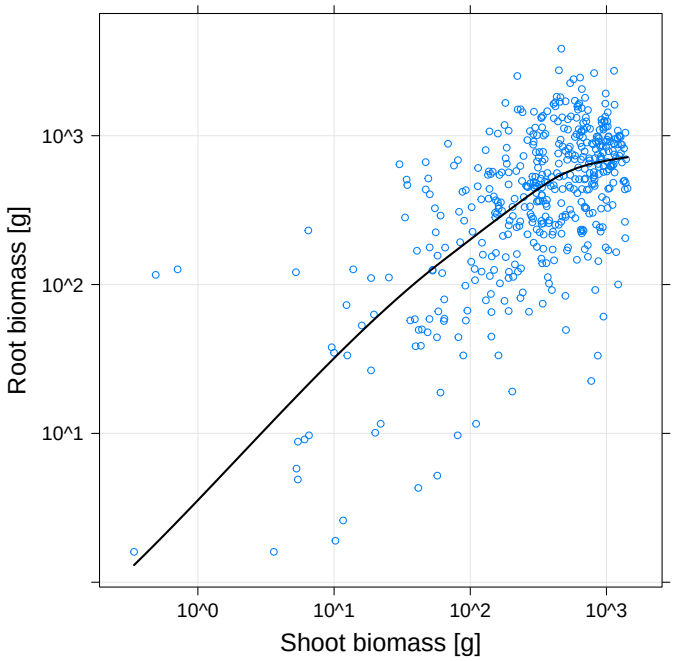


Figure 1: Scatterplot of the two response variables (shoot and root biomass). Scales are logarithmic.

Figure 1 shows that there is a positive relationship between shoot and root biomass. This is implies that, not unexpectedly, if the root biomass of a given plot is higher than the average, than this is true also for the shoot biomass measured in that same plot.

Modelling root biomass

Before fitting a model, we display the response variable against the predictor species richness. As mentioned above, we log-transformed the response variable. For ease of the interpretation, we use the logarithm base ten. Consistent with the original analysis, we also transform the species richness predictor with the logarithm base two (Hector et al., 2011).

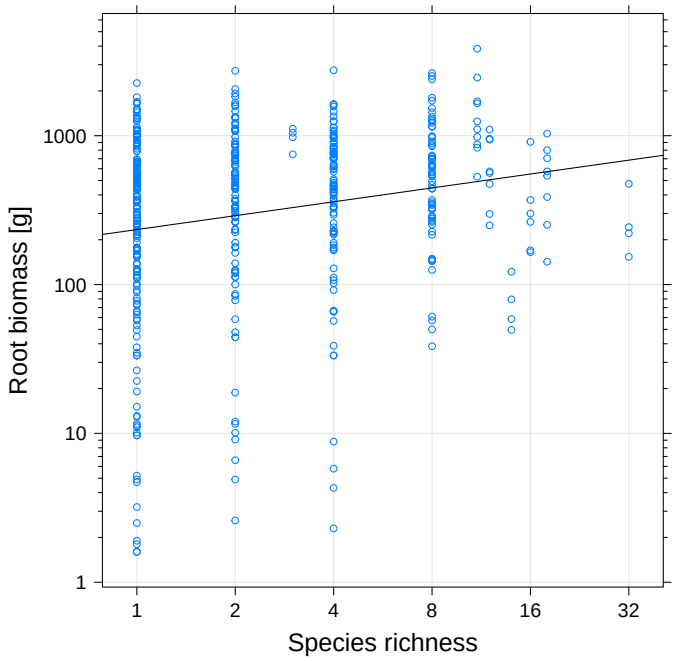


Figure 2: Root biomass plotted against species richness. Scales are logarithmic.

Root biomass seems to be positively influenced by species richness. Nevertheless, here we assumed linearity of the relationship (we added a regression line) and that the effect of species richness is the same in all sites. The next graph drops these assumptions and shows this relationship separately for each site with a smoother.

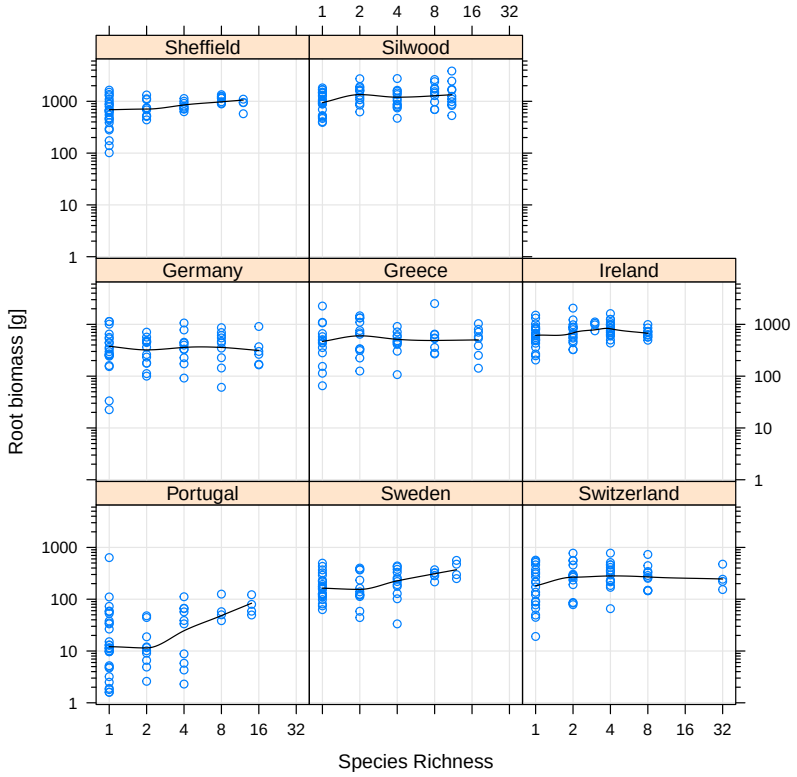


Figure 3: Root biomass plotted against species richness for each site. Scales are logarithmic, sites are ordered by their mean value (from left to right, bottom up).

Figure 3 shows that yields in Portugal were much lower than those in other sites. Indeed, in this site the mean values (see smoother) vary within 10 and 100 g. All the other sites vary, roughly speaking, between and 100 and 1.000 g. All sites, but Portugal, indicate that the effect of species richness can be considered to be linear.

We now fit a Linear Mixed Model to the log-transformed root biomass. The only fixed effects predictor is the log-transformed species richness. The random effects are: Mixture (or species composition), site and block. Mixture is taken as random as we are not interested in the particular effect of any specific species composition. We are rather interested in the effect of species richness, however, mixture is included into the model as it is part of the design.

```
require(lme4)
mod.root <- lmer(log10(root) ~ log2(SR) + (1 | Mix) + (1 | Block) + (1 | Site),
                 data = d.biodepth)
summary(mod.root)
```

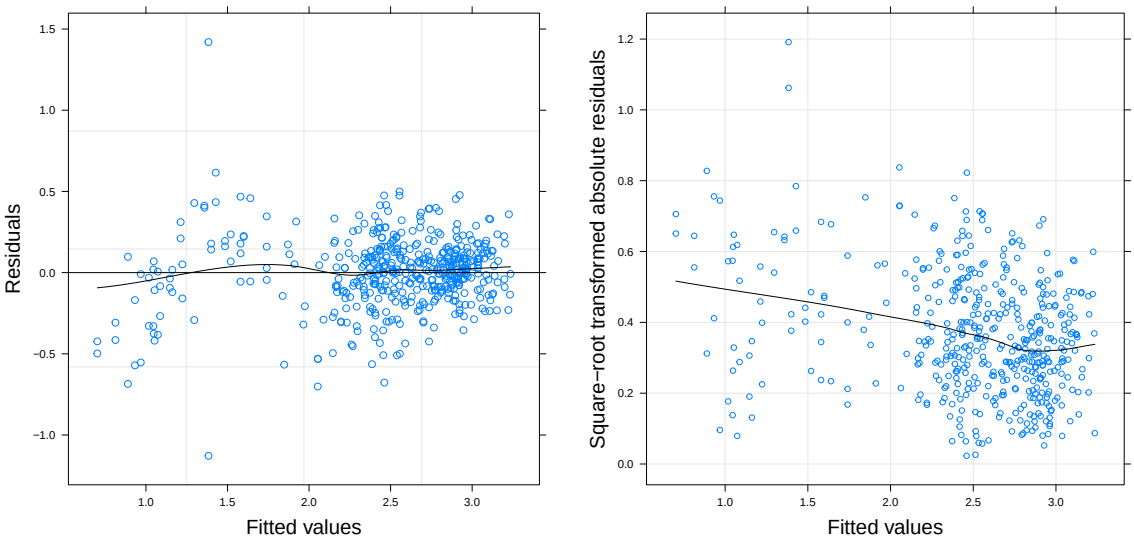
```
Linear mixed model fit by REML ['lmerMod']
Formula: log10(root) ~ log2(SR) + (1 | Mix) + (1 | Block) + (1 | Site)
   Data: d.biodepth

REML criterion at convergence: 277

Scaled residuals:
   Min     1Q Median     3Q    Max
```

```
-4.350 -0.430  0.057  0.487  5.473

Random effects:
 Groups   Name        Variance Std.Dev.
 Mix      (Intercept) 0.037807 0.194
 Block    (Intercept) 0.000255 0.016
 Site     (Intercept) 0.291551 0.540
 Residual             0.067276 0.259
Number of obs: 480, groups:  Mix, 200; Block, 15; Site, 8

Fixed effects:
            Estimate Std. Error t value
(Intercept)   2.3838     0.1931   12.34
log2(SR)      0.0645     0.0147    4.38

Correlation of Fixed Effects:
         (Intr)
log2(SR) -0.115
```

The summary output indicates that species richness has a positive effect likely to be statistically significant (t-value = 4.38). In addition, the estimated standard deviations of the random effects indicate that there is a some non-negligible variability among mixtures ($\hat{\sigma}_{Mix} = 0.19$), very little variation among blocks ($\hat{\sigma}_{Block} = 0.02$) and large variability among sites ($\hat{\sigma}_{Site} = 0.54$). The residual variability is estimated to be $\hat{\sigma}_{\varepsilon} = 0.26$. In practice, we can interpret these results as follow: There is large variation among sites, there is some variation within sites, but this is not explained by blocks and, finally, there is some variation explained by the Mixtures. Note that species richness is in the model. Therefore, the non-zero variability of mixtures, indicates that within a given level of species richness mixtures differ. In other words, species richness is not enough to fully explain the differences observed among mixtures.

Checking the model assumptions for `mod.root`

Before fitting a model to the other response variable, we check whether the model assumptions can be considered fulfilled. For the sake of brevity only some relevant steps of this procedure are shown. We fist look at two diagnostics plots introduced in Appendix C, the Tukey-Anscombe plot and the scale-location plot.



(a) Tukey-Anscombe plot

(b) Scale-location plot

Figure 4: Residuals diagnostics for mod.root.

Both graphs in Figure 4 indicate that the variance of the residuals is decreasing with the fitted values. Linking back to Appendix C ("Stability"), we can say that, with this data, the coefficient of variation is not fully appropriate to measure stability. Indeed, this data is also aggregated and does not represent the raw measurements (in some cases sub-samples have been combined; Andrew Hector, personal communication).

The Tukey-Anscombe plot (a) also shows that there are two observations with very large residuals. Both belong to the site Portugal. This does not come unexpected as this site appeared to behave quite differently from all the others. We may want to inspect this further by looking at the conditional modes of the random effects for sites (i.e. the predicted random effects). To do that, we use a so-called caterpillar plot (Bates, 2010). This plot is used to display the conditional modes of any random effect of interest. Conditional modes are shown along with a 95% credible interval[2]. On caterpillar plots, the conditional modes are ordered by their value. As seen above, Portugal had lowest yields and is thus expected to have the lowest conditional mode. The caterpillar plot here is used to identify outlying sites.

[2]Credible intervals are the analogue of confidence intervals in the Bayesian framework. This framework is used to make predictions on the estimated random effects (Bates, 2010).

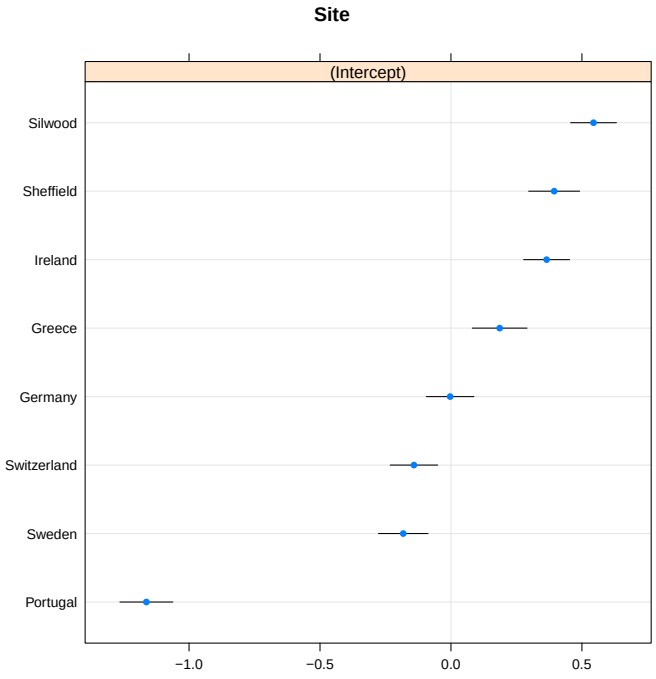


Figure 5: Caterpillar plot for the random effect sites.

Figure 5 confirms that Portugal is an outlying site. If this was of major concern, we could use Robust Mixed Effect Models (e.g. the add-on package *robustlmm*; Koller (2016)). For the analysis presented here, `mod.root` is our final model.

Modelling shoot biomass

We are now turning our attention to the other response variable measured: Shoot biomass. We start by plotting the data.

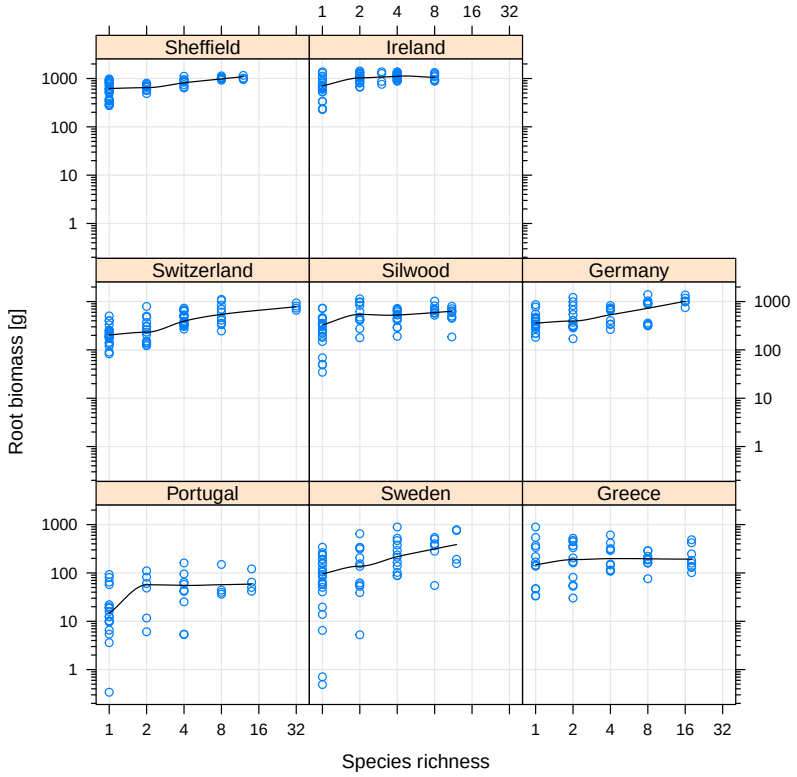


Figure 6: Shoot biomass plotted against species richness for each site. Scales are logarithmic, sites are ordered by their mean value.

Figure 6 shows a similar situation as observed for the other response variable (`root`). The effect of species richness appears to be positive. The later may vary among sites. Portugal shows again very low yields.

We now fit the same model as before, and compare the estimated coefficients of the two models.

```
mod.shoot <- lmer(log10(shoot) ~ log2(SR) + (1 | Mix) + (1 | Block) + (1 | Site),
                  data = d.biodepth)
summary(mod.shoot)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: log10(shoot) ~ log2(SR) + (1 | Mix) + (1 | Block) + (1 | Site)
   Data: d.biodepth

REML criterion at convergence: 292

Scaled residuals:
   Min     1Q Median     3Q    Max
-6.431 -0.344  0.027  0.369  2.607

Random effects:
 Groups   Name        Variance Std.Dev.
 Mix      (Intercept) 0.03395  0.1843
 Block    (Intercept) 0.00165  0.0406
```

```
 Site     (Intercept) 0.25253  0.5025
 Residual             0.07342  0.2710
Number of obs: 465, groups:  Mix, 193; Block, 15; Site, 8

Fixed effects:
            Estimate Std. Error t value
(Intercept)   2.2592     0.1804   12.52
log2(SR)      0.1156     0.0148    7.83

Correlation of Fixed Effects:
         (Intr)
log2(SR) -0.126
```

The results of the two models are very similar. Species richness has a positive effect, and there is evidence that this is statistically significant. Interestingly, also the estimated variances of the random effects are very similar (see below).

```
VarCorr(mod.root)
```

```
 Groups   Name        Std.Dev.
 Mix      (Intercept) 0.194
 Block    (Intercept) 0.016
 Site     (Intercept) 0.540
 Residual             0.259
```

```
VarCorr(mod.shoot)
```

```
 Groups   Name        Std.Dev.
 Mix      (Intercept) 0.1843
 Block    (Intercept) 0.0406
 Site     (Intercept) 0.5025
 Residual             0.2710
```

The residual diagnostics for `mod.shoot` (not shown) are not ideal. They are slightly worse than those obtained for `mod.root`. If the main purpose of this analysis was to estimate the effect of species richness, robust methods could be used. However, the aim of our analysis is to correlate estimated random effects, we can thus proceed (in this case, the analysis with robust methods yields same results).

2.2 Correlating estimated effects of the two analyses

The two analyses yielded similar results in terms of estimated for the fixed, but also in terms of variability of the random effects. Now we are interested in formally comparing these results to gather new biological information. We start by looking at the correlation of the conditional modes for sites. Put simply, we are now testing whether a good site for root biomass is also a good site for shoot biomass. We have already seen that one site (Portugal) seems to be associated with very low yields for both response variables. To inspect the relationship among site effects, we first extract the estimated conditional modes for both response variables.

```
( d.condMod.Site.root <- ranef(mod.root)[["Site"]] )
```

```
            (Intercept)
Germany         -0.0031
Greece           0.1860
Ireland          0.3652
Portugal        -1.1630
Sheffield        0.3940
Silwood          0.5443
Sweden          -0.1821
Switzerland     -0.1413
```

```
( d.condMod.Site.shoot <- ranef(mod.shoot)[["Site"]] )
```

```
            (Intercept)
Germany           0.245
Greece           -0.206
Ireland           0.533
Portugal         -0.977
Sheffield         0.487
Silwood           0.197
Sweden           -0.352
Switzerland       0.073
```

At first sight, the values of the conditional modes are not very informative. The reason is that the response variable is log-transformed. However, we can interpret the conditional modes for sites as being multiplicative on the untransformed scale of the response variable. To make an example, the site `Portugal` has a value close to -1 for both models, since we are using the logarithm base ten, we can conclude that the mean biomass in this site is about ten time smaller that the average of the other sites. This is exactly was we saw when plotting the data (Fig. 3 and Fig. 6). To make an example of this kind of computation, we can look at the estimated value for Portugal in the shoot model (i.e. -0.977). Its multiplicative effect on the mean is $10^{-0.977} = 0.11$. So, on average, the shoot biomass in Portugal is (about) ten times smaller than in all other sites.

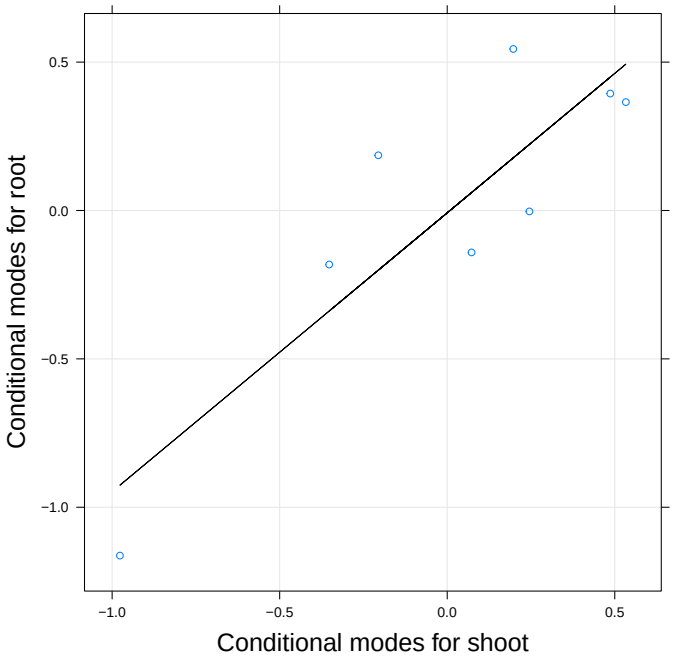


Figure 7: Scatterplot of the conditional modes for sites. A regression line estimated in a robust way is added to the graph.

Figure 7 shows a positive relationship between the `site` conditional modes of the two models. This implies that a good site for shoot biomass is also a good site for root biomass. This is not surprising, indeed, the aim of this first example is indeed to introduce the method. Note that the regression line added to this graph is based on robust methods (Todorov et al., 2016). This was done to deal with the outlying site Portugal.

We now turn our attention to another random effect shared by these two models, the mixture (`Mix`). If for sites a positive correlation was expected, this is not necessarily the case here. Indeed, there may be trade-off between what plants store in their roots and what contributes to shoot biomass. It may be that plants of some species tend to store have higher root biomass than shoot biomass, while other do the opposite. The data of this experiment is not available at population level (i.e. species within plot), but at community level (i.e. plot). Nevertheless, the effect of different storing strategies can be seen when looking at mixtures, as each mixture is by definition a sum of species.

If a trade-off between roots and shoots was present, we would expected to see a negative relationship when correlating the conditional modes for the random effect mixture. On the other hand, a particular species composition may boost productivity of all species in a given plot and thus have a positive effect on both biomasses[3]. Alternatively, some species may have a greater biomass for both shoot and root and thus lead to a positive correlation of these conditional modes[4]. Let's inspect this graphically.

[3]This can be seen as in analogy to the "niche partitioning" hypothesis in the context of overyielding (Loreau and Hector, 2001).

[4]Again, this can be seen as in analogy to the "sampling effect" in the context of overyielding (Loreau and Hector, 2001).

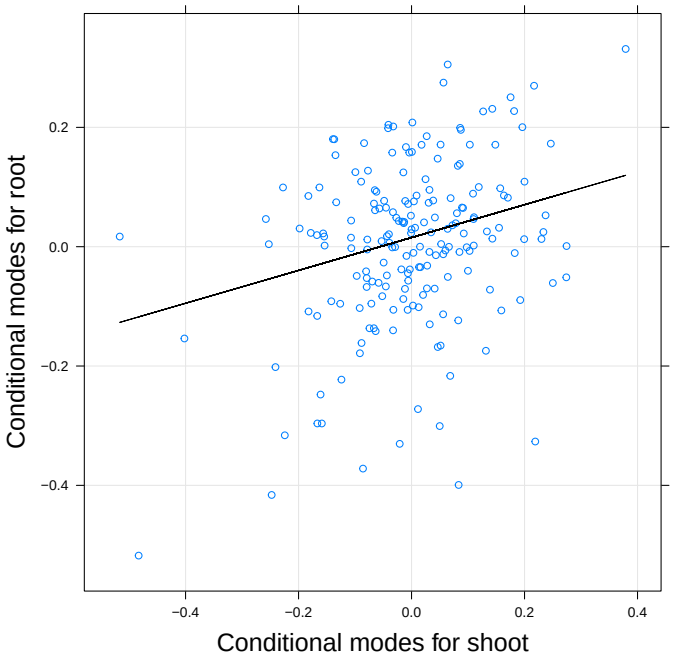


Figure 8: Scatterplot of the conditional modes for mixtures. A robust regression line is added to the graph. Each dot on this graph represents a particular mixture.

Interestingly, we see a weak positive relationship between the conditional modes of shoot and root. We can therefore conclude that the trade-off between root-shoot allocation has a smaller effect than the species identity effects and species interaction effects.

Testing more complex hypotheses

We may want check whether this correlation is modulated by functional richness. If the species interaction hypothesis had a stronger effect, then we would observe an increasing correlation of the conditional modes for higher values of functional richness. In other words, more functional groups would enable stronger species interactions and thus both root and shoot biomass were to be boosted. We inspect this hypothesis graphically.

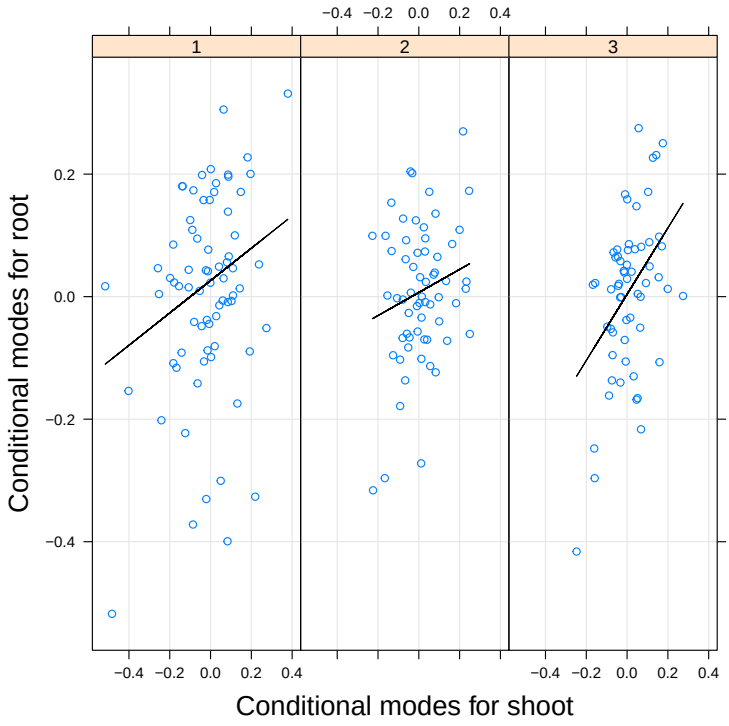


Figure 9: Scatterplot of the conditional modes for mixtures. Functional richness effect is highlighted with panels. Robust regression lines are added to the graphs.

Figure 9 shows that the correlation between root and shoot conditional modes is the strongest at the highest level of functional richness.

Testing the effect of sites on mixture conditional modes (**)
Sites appeared to the greatest variability (they had the greatest variance component). In addition, a site appeared to behave quite differently from the other. Therefore, we may want to inspect whether the relationship between the mixture conditional modes for root and shoot is the same in all sites. Or whether, different sites show different relationships. It could happen that the root-shoot trade-off is present in low-productivity sites only. If this was the case, we would expect the Portugal to show a strong negative correlation.

Remember that not all mixtures are present in all sites (there are 200 mixtures in total and each site has about 30 of them). Here we want to test whether the relationship of the conditional modes for mixture differ among sites. To do that, we must allow mixtures to behave differently in different sites. Put simply, we allow sites and mixtures to interact. To do that, we remove the random effect mixture and add a random effect mixture by site combination (this latter is named `MixNested`).

```
mod.root.2 <- update(mod.root, . ~ . - (1 | Mix) + (1 | MixNested))
mod.shoot.2 <- update(mod.shoot, . ~ . - (1 | Mix) + (1 | MixNested))
```

We then repeat the same procedure used above (i.e. we extract conditional modes and we correlate them). We look at the relationship of interest separately for each site.

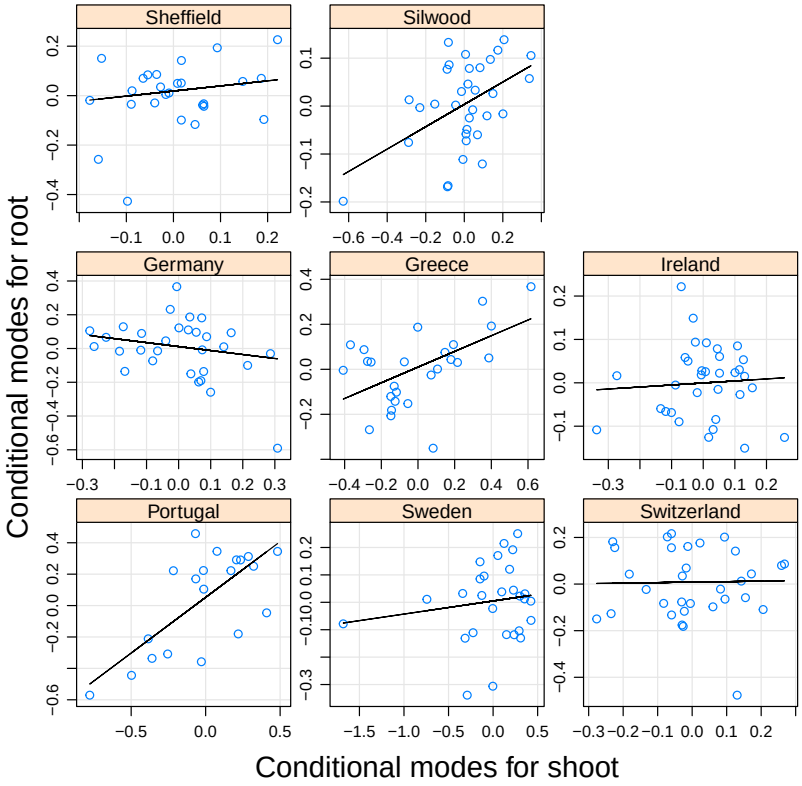


Figure 10: Scatterplot of the conditional modes for mixtures. Sites are highlighted with panels. Robust regression lines are added.

Interestingly, the relationship is changing among sites. Although most sites show a positive relationship, Germany shows a negative correlation. This may imply that the trade-off between the biomass put into roots and into shoots is stronger there.

3 Example II: Sabah Intensive Experiment (**)

3.1 Modelling both response variables

We now turn our attention to another data set, the Sabah Intensive Experiment. we already encountered it in Appendix A ("Graphs") where we looked at the spatial structure of this experiment. The are two response variables that were measured, tree survival and tree diameter. Both these variables were recorded at tree level, however, data was then aggregated at population level (i.e. species level in each plot).

Visualising survival
To best remember the structure of this data set we start by visualising the response variable survival.

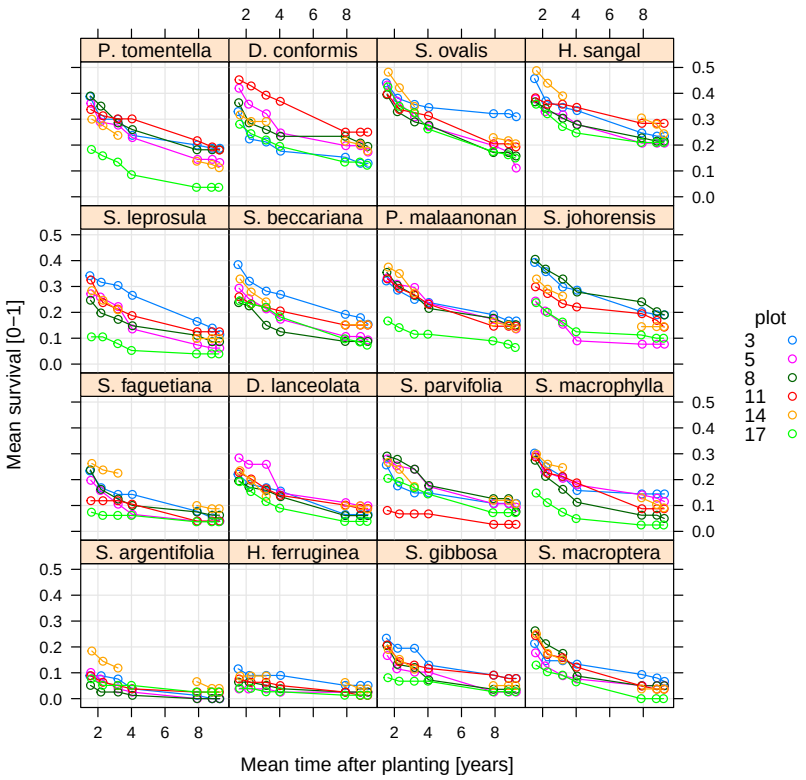


Figure 11: Survival of the 16 species in the Sabah Intensive Experiment. Each panel shows a species, each coloured line refers to a particular plot.

Figure 11 shows the survival of the 16 species used in the Sabah Intensive Experiment. Each observation represent the survival of a particular species, in a particular plot at given time point. Plots are rendered with colours in each panel. All species show a decline in survival, nevertheless, there are important differences among species but also within species (i.e. among plots of the same species). Species (i.e. panels) are ordered by mean survival.

Visualising diameter

We now inspect graphically the other response variable measured in this experiment, tree diameter. We produce the same graph as for survival (Fig. 11), but for the panel ordering (here based on growth) and the aspect ratio (here isometric). Growth is simply the slope observed in each panel. Diameter was log-transformed (base ten) prior to aggregate data at population level.

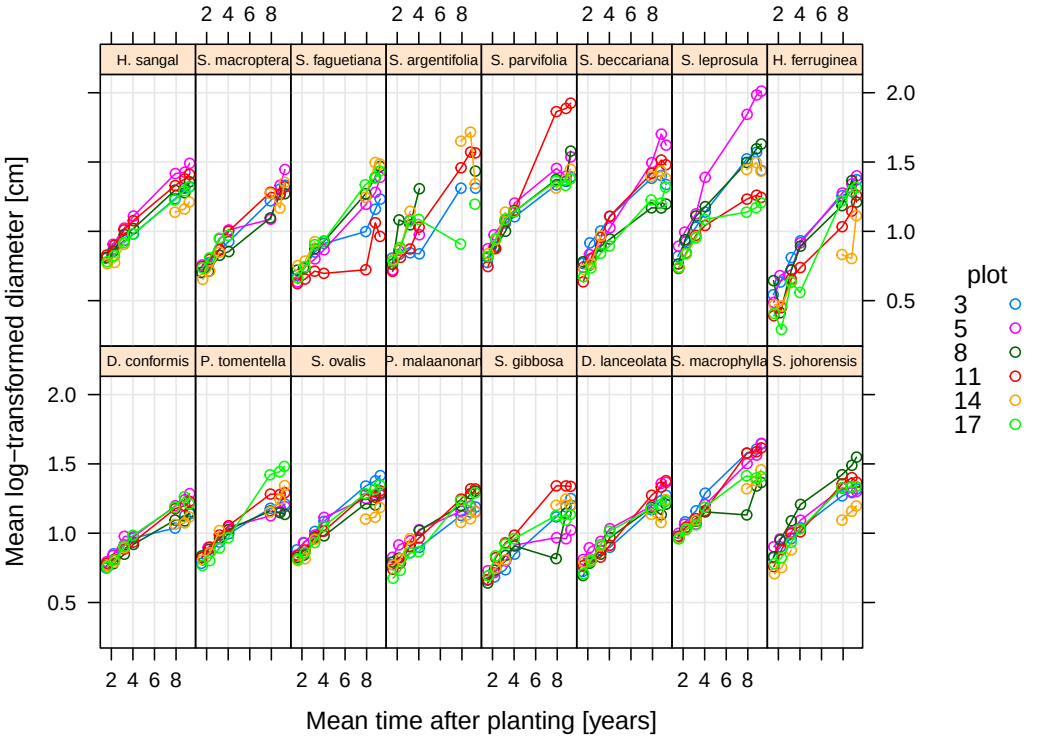


Figure 12: Diameter plotted over time for the 16 species of the Sabah Intensive Experiment. Each panel shows a species, each coloured line refers to a particular plot.

Figure 12 shows that there are important differences among species, but little variation within species. Plots do not seem to have a very strong influence on diameter. Interestingly, the variability within each panel seems to slightly increase with time. This would imply that the plot effect of diameter is small, but that there is some plot effect on growth. Figure 12 was produced to highlight differences in growth (i.e. slopes) among species. To test whether there is variability in growth within species (i.e. among plots), we can modify Figure 12 so that each panel has different y-ranges.

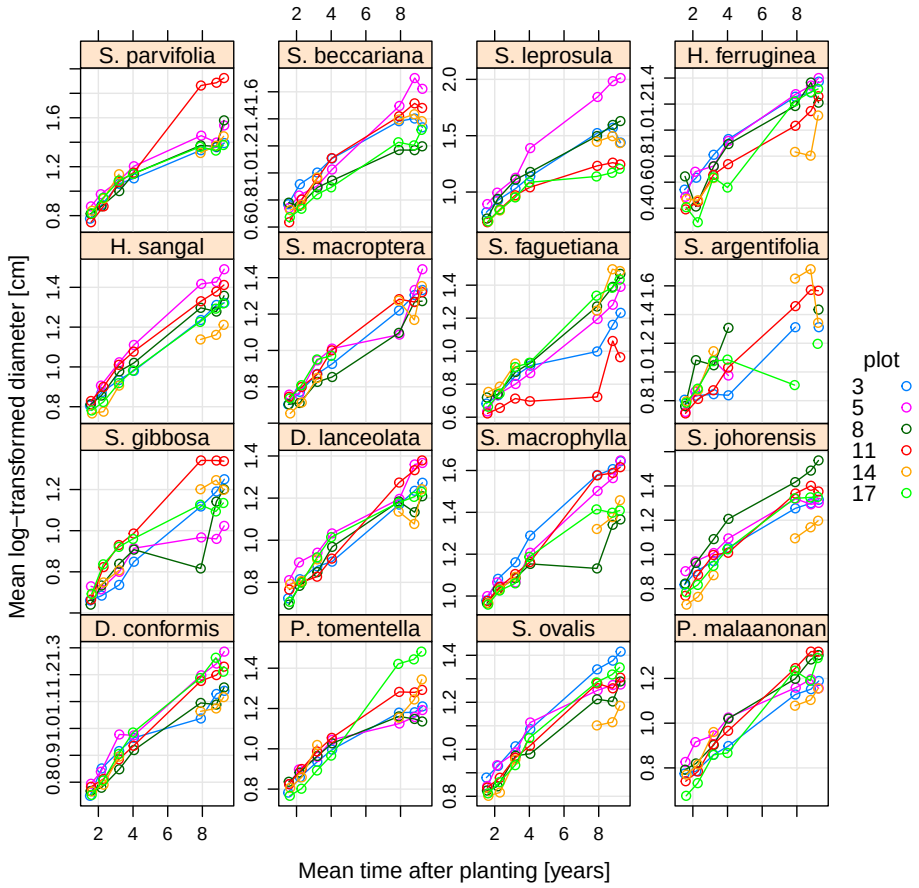


Figure 13: Diameter plotted over time for the 16 species of the Sabah Intensive Experiment. Each panel shows a species, each coloured line refers to a particular plot.

This graph indicates that plots may have an influence on growth. Indeed, lines within each panel are not parallel but diverge over time. We are going to formally test this hypothesis in the modelling phase.

Modelling survival

We now model survival visualised in Figure 11 at page 16 with a Linear Mixed Effects Model. To start with, we allow each species to have a different mean survival and also a different survival decrease over time. In other words, we fit a model with an interaction between species and time. We also add a random effect for plots.

```
mod.surv.0 <- lmer(m.surv ~ species.short * m.time.years + (1 | plot),
                   data = d.sbe.int)
## summary(mod.surv.0)  ## (not printed because long)
```

The model fitted above assumes that each species has a different mean a different trend over time. Figure 11 clearly indicated that species differ in their mean and trend values for survival. We may want to look at these two parameters for one species. We take the reference species of the `species.short` factor, D. conformis. We can thus look at its estimated coefficients[5].

[5]Remember that **R** uses treatment contrasts by default.

```
fixef(mod.surv.0)[c("(Intercept)", "m.time.years")]
```

```
 (Intercept) m.time.years
       0.360       -0.021
```

The second coefficient (`m.time.years`) indicates that for this species survival is decreasing over time at a rate of roughly -2% per year. The other coefficient (`Intercept`) indicates that at time zero (i.e. at planting) survival is modelled to be 36%. This obviously does not make sense as at planting all seedlings were very likely alive. The main problem when interpreting the species intercepts is that we are assuming that our model well represents reality outside the fitting region. Indeed, the first measurements were taken at 1.52 years. We cannot thus extrapolate our model to the planting date.

Nevertheless, we may want to compare the mean survival among species in a meaningful way. To do that, we can simply re-parametrise our model so that the intercepts do not indicate survival at time zero, but at a convenient time point that falls within the range of the observed data (i.e. 1.5, 9.3). A meaningful choice is often the mean value of the predictor (here 5.26). On page 21, we will show how to implement this re-parametrisation.

Going to back to the estimated survival at time zero (i.e. 36%), we notice that this greatly underestimates that real expected survival (i.e. 100%). However, time zero is just about out of the range of the data. The reason for such a large underestimation is that we assume a linear trend, however, Figure 11 indicated that the decrease in survival is not linear.

If assuming a linear trend is not correct, then the residual diagnostics will highlight this problem and help us improving our model. We can plot the residuals of the fitted model (i.e. `mod.surv.0`) against `m.time.years` to see whether any pattern is present.

```
xyplot(resid(mod.surv.0) ~ m.time.years, data = d.sbe.int,
       abline = 0,
       type = c("p", "smooth"),
       xlab = "Mean time after planting [years]",
       ylab = "Residuals",
       col.line = "black")
```

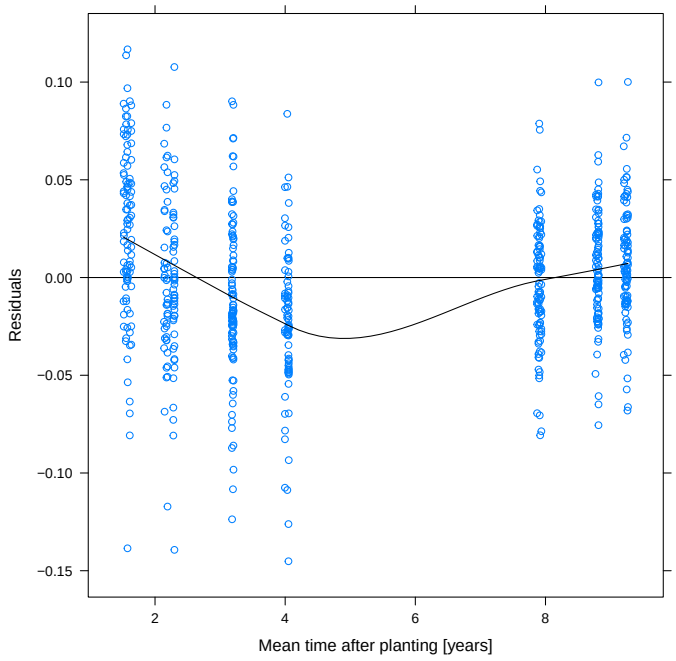


Figure 14: Residuals of mod.surv.0 plotted against time.

Figure 14 clearly shows that fitting a model that assumes a linear trend for survival is not enough to model this data. Indeed, there is a clear (non-linear) pattern present in the residuals (i.e. the black smoother is not flat on zero). The pattern observed indicates that by adding a quadratic term, we may better fit the data.

The next question to ask before fitting a model with the quadratic term is whether all species need the same quadratic term, or whether species show different patterns. To inspect that we reproduce Figure 14) with panelling for the 16 species.

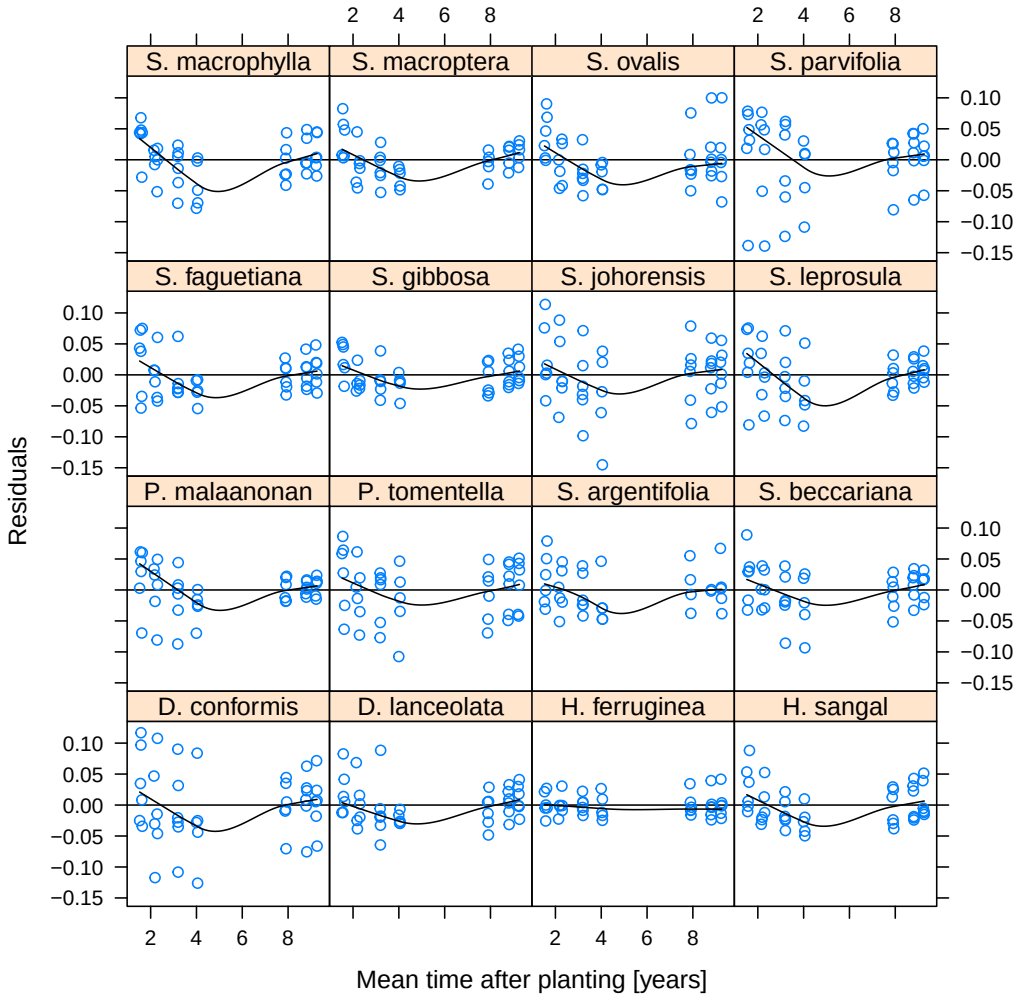


Figure 15: Residuals of mod.surv.0 plotted against time. Each species is plotted in a separate panel.

All species, but `H. ferruginea` show a non-linear relationship. However, the latter seems to be vary among panels (the curvature of the smoother varies across panels). We therefore need a model where species are allowed to have a different quadratic term in time.

Remember that for the ease of the interpretation of the regression coefficient of this model we are now modifying the time variable. In particular, we are centring time such that intercept of each species represents the "mean survival five years after planting". To model the non-linear effect of time, we are also adding a quadratic term to our model.

```
## 1. creating the centered "time" variable
d.sbe.int$m.time.5 <- d.sbe.int$m.time.years - 5
##
## 2. creating the quadratic term
d.sbe.int$m.time.5.squared <- d.sbe.int$m.time.5^2
##
## 3. fitting an improved model
```

```
mod.surv.1 <- lmer(m.surv ~ species.short * (m.time.5 + m.time.5.squared) +
                     (1 | plot),
                   data = d.sbe.int)
## summary(mod.surv.1)  ##  (not printed because long)
```

After fitting this improved model, we may want to look at the estimated coefficients to interpret them biologically. We are looking again at the reference species D. conformis. Note that in addition to an intercept and a slope, there will be a coefficient for the quadratic term here.

```
fixef(mod.surv.1)[c("(Intercept)", "m.time.5", "m.time.5.squared")]
```

```
     (Intercept)         m.time.5 m.time.5.squared
          0.2281          -0.0239           0.0032
```

Our model predicts the mean survival for species D. conformis five years after planting to be about 22%. This makes (biological) sense. This is indeed, what we see in Figure 11. Interestingly, we can also interpret the coefficient of `m.time.5`: Five years after planting survival is decreasing at a speed of about 2.4% per year. Note that as we fitted a model with a quadratic term in time, this value changes over time. Indeed, the speed at which survival decreases is now changing over time. Obviously at the beginning of the experiment the speed is at its highest, while it deceases towards the end. The coefficient for the quadratic term is, indeed, positive that indicates a curve open upwards (i.e. a convex curve).

We may want to check whether there is any structure left in the data. To do that we reproduce the same graph as above (residuals plotted against time). Here we use grouping to highlight plots.

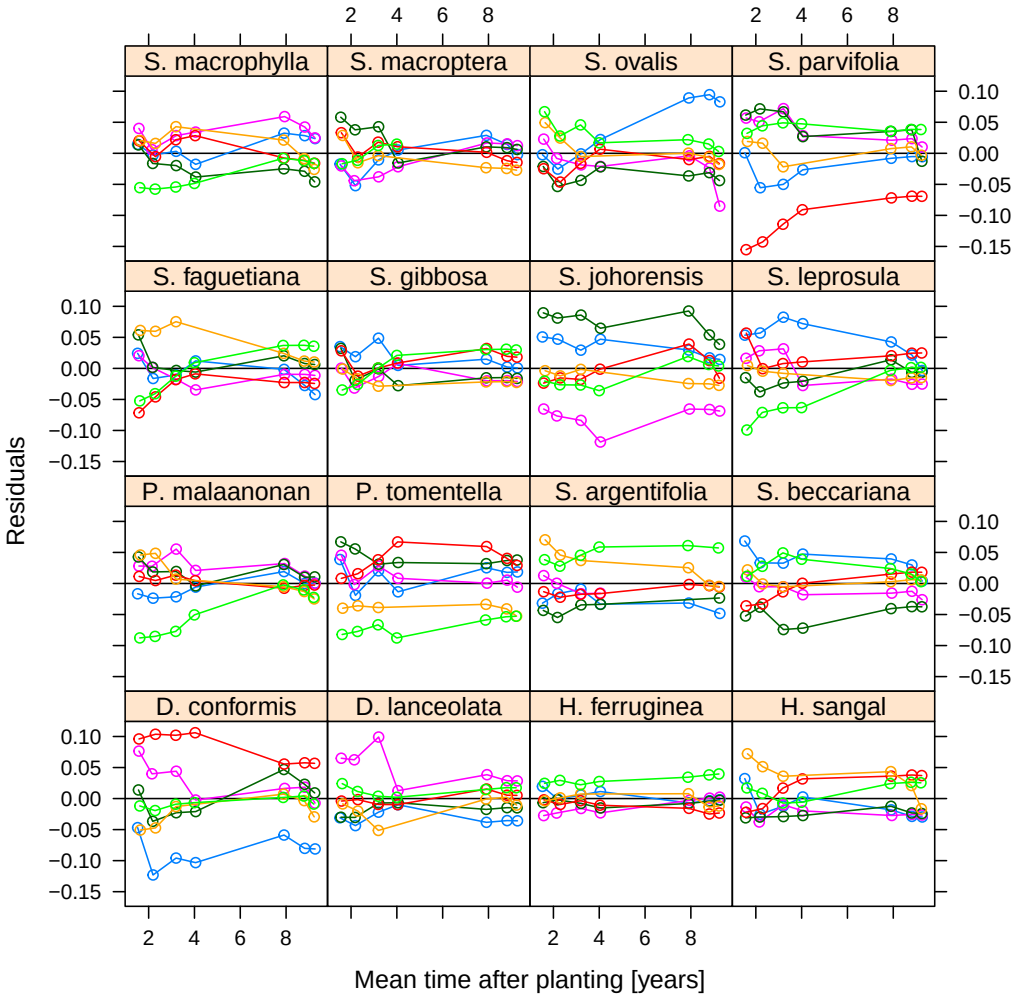


Figure 16: Residuals of mod.surv.1 plotted against time. Each species is plotted in a separate panel.

Figure 16 shows that there is still some structure left in the residuals. Indeed, in each panel, plot effects are very clear (i.e. plots shows clear horizontal shifts within panels). These plot effects differ among panels. For example, the best plot for `D. conformis` and `D. lanceolata` (bottom left panels) is not the same. There is thus a clear indication of a species by plot interaction. The effect of plots differ among species (i.e. species by environment interaction). There is thus evidence for spatial asynchrony among species[6]. We can improve our current model by adding the interaction among species and plots as a random effect.

```
mod.surv.2 <- update(mod.surv.1, . ~ . + (1 | plot:species.short))
```

After this interaction term is introduced in our model, there is no evidence that the model can be made more complex to better fit the data (not shown). Further residual and random effects diagnostics do not show any major violations of the model assumptions either (not shown). `mod.surv.2` is thus the final model fitted to the response

[6]Obviously, this claim must be supported by other types of evidence (e.g. statistical inference). Omitted for brevity.

variable survival that will be used to correlate random and fixed effects.

Final note

If stability was the main aim of this analysis, we may want to make sure that variance of the residuals is perfectly stable (to facilitate comparisons among groups). In this case the response variable is a proportion, therefore, the corresponding variance-stabilising transformation is the arc-sinus square-root transformation (Held and Sabanés Bové, 2014). However, if we were to transform the response variable the interpretation of the model coefficients would become more difficult. It is also worth mentioning that the sample diameter used to compute survival rates differ among species and also over time (obviously, species with lower survival have fewer trees left towards the end of the experiment). We could include this information in our model with weights. Again, for the sake of simplicity this was not done here (this does not alter the results of the correlation analysis shown below).

Modelling the log-transformed diameter

We now fit a model to the averaged log-transformed diameter. For the sake of brevity, we fit a starting model that already contains a quadratic time effect for each species. In this model the random effect plot is not yet allowed to interact with species identity. In other words, the effect of a given plot is considered equal for all species.

```
mod.diam.1 <- lmer(m.diam ~ species.short * (m.time.5 + m.time.5.squared) +
                     (1 | plot),
                   data = d.sbe.int)
## summary(mod.diam.1)  ##  (not printed because long)
```

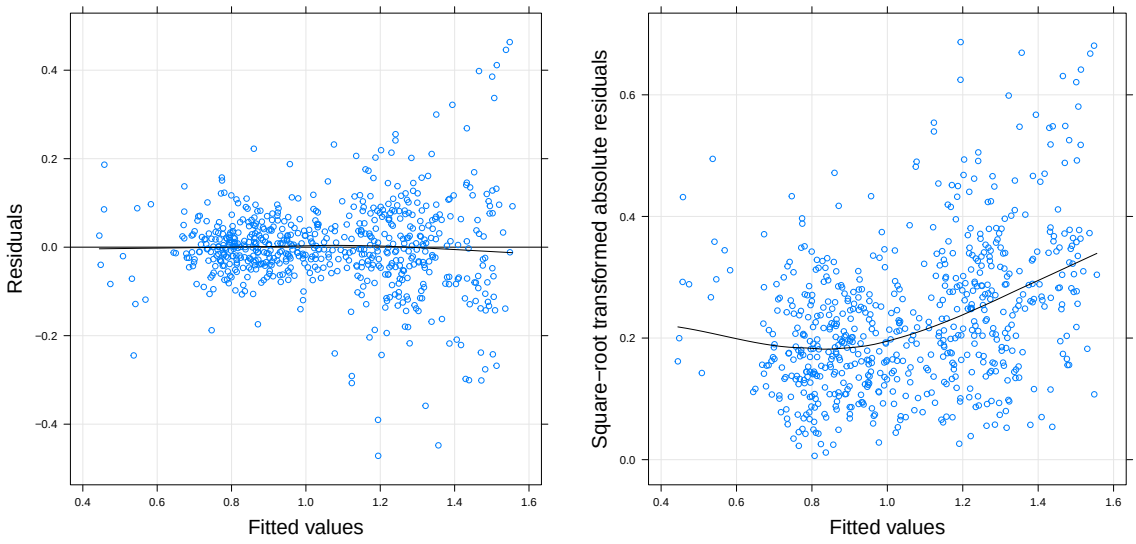
We now look at the estimated coefficients and interpret them. Again we look at the reference species D. conformis.

```
fixef(mod.diam.1)[c("(Intercept)", "m.time.5", "m.time.5.squared")]
```

```
     (Intercept)         m.time.5 m.time.5.squared
          1.0059           0.0565          -0.0037
```

All these coefficients have to be interpreted while keeping in mind that the response variable was log-transformed (base ten). To make an example, the intercept for this species is about 1. Therefore, the diameter of D. conformis after 5 years since planting is $10^1 = 10$ cm. The other two coefficients are harder to interpret in the original back-transformed scale. Nevertheless, we safely claim that after trees of this species are growing (i.e. slope is positive), but that growth is decelerating (quadratic term is negative).

At this stage we are checking the main model assumptions with a Tukey-Anscombe plot and a scale-location plot.



(a) Tukey-Anscombe plot

(b) Scale-location plot

Figure 17: Residuals diagnostics for mod.diam.1.

Figure 17 shows that the variance is increasing with the fitted values. This pattern may be due the real mean-variance relationship of this data set, or to a misspecified model. Remember that when we visualised this data on Figure 13 we noticed that plots appears to diverge over time. This may explain the increasing variance of the residuals. Indeed, pattern may be due to interaction between plots and species. To check that, we visualise the residuals of `mod.diam.2` against time for each species separately.

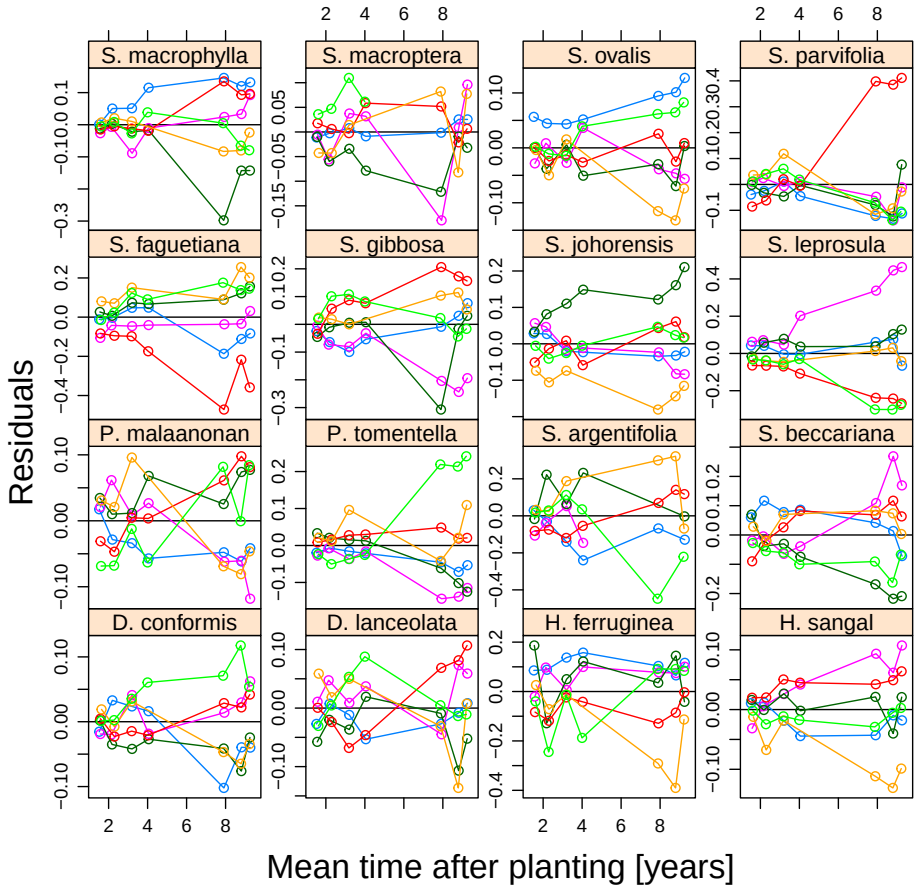
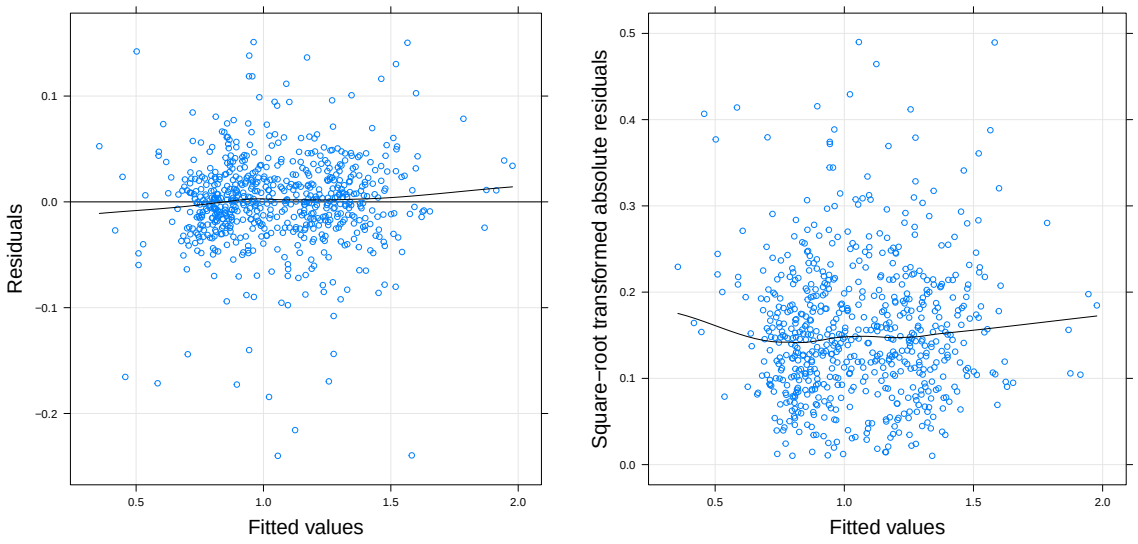


Figure 18: Residuals of mod.diam.1 plotted against time.

Figure 18 clearly indicates that plots have different slopes for each species. In other words, growth of a given species differs among plots. We can model that with an interaction among species, plot and time. To do that, we use the so-called random slopes (Bates, 2010).

```
mod.diam.3 <- update(mod.diam.1, . ~. + (m.time.5 | plot:species.short),
                     data = d.sbe.int)
```

After adding random slopes to our model the variance is stabilised (Fig. 19 below). This is an important point to take into account when using the approach presented in Appendix C to model stability. Indeed, a non-stable variance may just indicate that the model equation is incorrect.



(a) Tukey-Anscombe plot

(b) Scale-location plot

Figure 19: Residuals diagnostics for mod.diam.3.

Figure 19 shows that the mean-variance relationship is stabilised after the model equation is improved by adding random slopes for growth. `mod.diam.3` is the final model that will be used in the next subsection where estimated effects are correlated.

3.2 Correlating estimated effects for the two modelled response variables

Correlating the estimated effect of plot on diameter and survival
In analogy to what we did we the Biodepth data, we are going to correlate the plot random effects (there we used sites). In this experiment,the 16 species were grown in all six plots. We may thus want to inspect whether a good plot for survival is also a good plot for diameter. In other words, we are testing the hypothesis that some plots are of good quality in general and this affects both, mean survival and mean diameter. Alternatively, we may find that good conditions for survival are not the same as for tree diameter (diameter). To do inspect this hypothesis, we are again using the predicted values of the random effects `plot`, the so-called conditional modes.

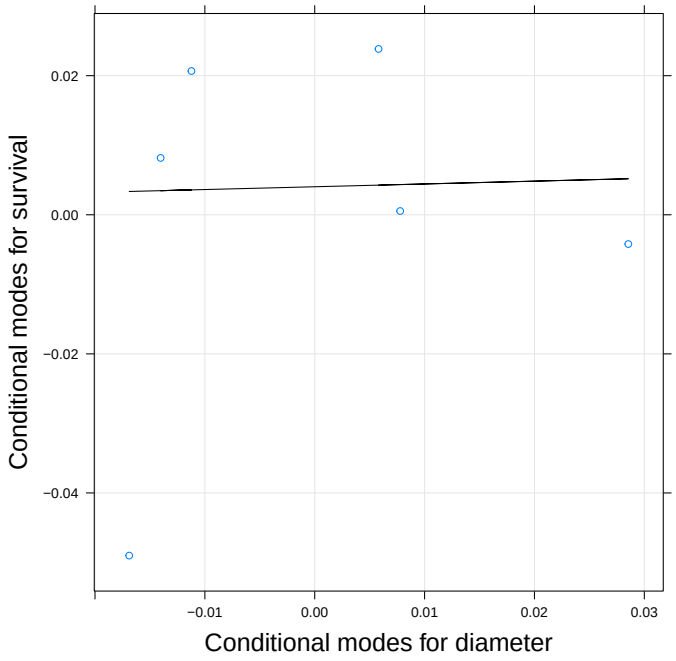


Figure 20: Scatterplot of the conditional modes for plot. Each dot on this figure represents a plot. A robust regression line is added to the graph.

Figure 20 shows that the is little correlation between the conditional modes of plots for survival and diameter. This implies that there is little support for the hypothesis that the ideal plot conditions for survival are the same for diameter.

Correlating survival and diameter

We may want to continue or correlative analysis and test whether species that have large diameter mean values have also high survival. To test this hypothesis, we are going to use the estimated fixed effects of our model. To be more precise, we are now correlating the survival with the mean diameter five years after planting (for each species).

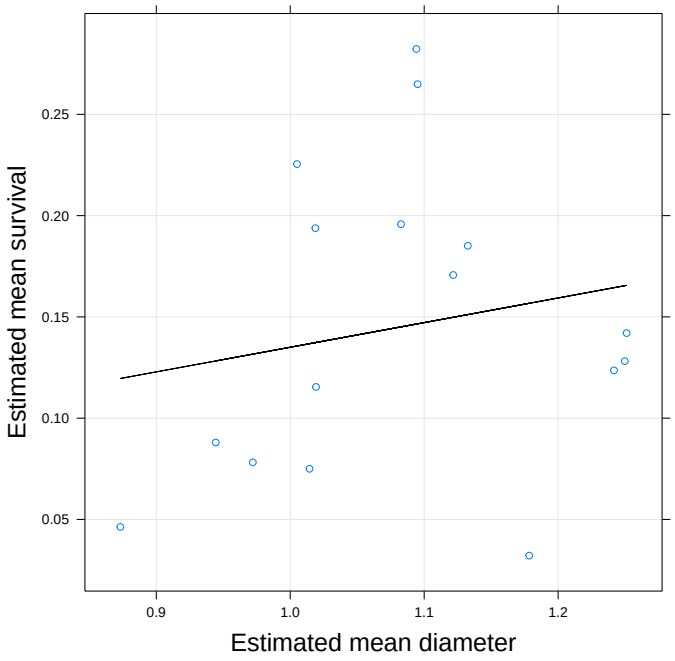


Figure 21: Scatterplot of the estimated mean survival rates and mean diameter five years after planting. Each dot on this figure represents a species. A robust regression line is added to the graph.

This graph indicates that there is little support for a correlation between diameter and survival. So larger species do not seem to be those species that show higher survival. Note that here we are not testing whether fast growing species are those that have lower survival (i.e. a growth-survival trade-off). To test this hypothesis, we must look at the the other coefficients estimated in this analysis. In particular, we must look at the species-specific growth rates. (see next paragraph).

Correlating mean survival and mean growth rate

We can now formally test whether fast growing species have lower survival rates. Given the parametrisation of our model (time is centred at five years), we are testing this hypothesis five years after planting[7].

[7]When fitting a model with a quadratic term for time, growth at time 0 (here five years) is the value estimated value for the linear term.

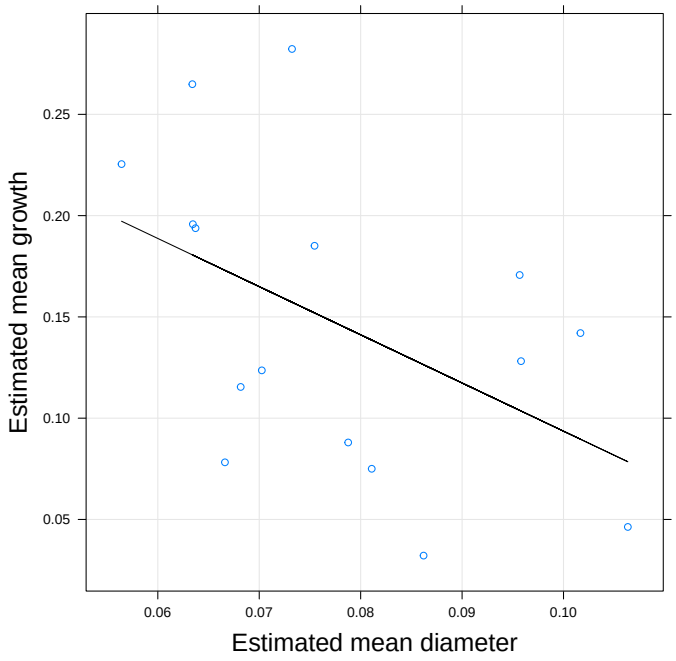


Figure 22: Scatterplot of the estimated mean survival rates and mean growth rates five years after planting. Each dot on this figure represents a particular species. A robust regression line is added to the graph.

This graph shows clearly that there is a trade-off between growth and survival. Faster growing species have lower survivals. This finding corroborates with the results of the larger Sabah Biodiversity Experiment (Tuck et al., 2016).

Correlating species by plot interactions for survival and growth (***)
Finally, we may want to test a more involved hypothesis that combines what done above. The hypothesis that we are going to test here is whether, for a given species, plots showing high survival and also those with high growth rates. This hypothesis is tested with the conditional modes of the species by plot interactions of the survival model (i.e. species-specific local survival) and with the species by plot growth rates (i.e. species-specific local growth rates).

If for a given species a plot that has high survival has also high growth rates, than we will see a positive correlation. We have already seen that this positive correlation was not present at plot level. It is important to note that if we were to correlate the mean survival rates with the mean growth rates for each species-plot combinations, we would observe a negative correlation given by the growth-survival trade-off. Here we are looking at these effect once the effect of this trade-off has been removed. We are thus using a conditional approach. On top of removing the effect of the trade-off, we are also removing the confounding effect of the other variables present in the model.

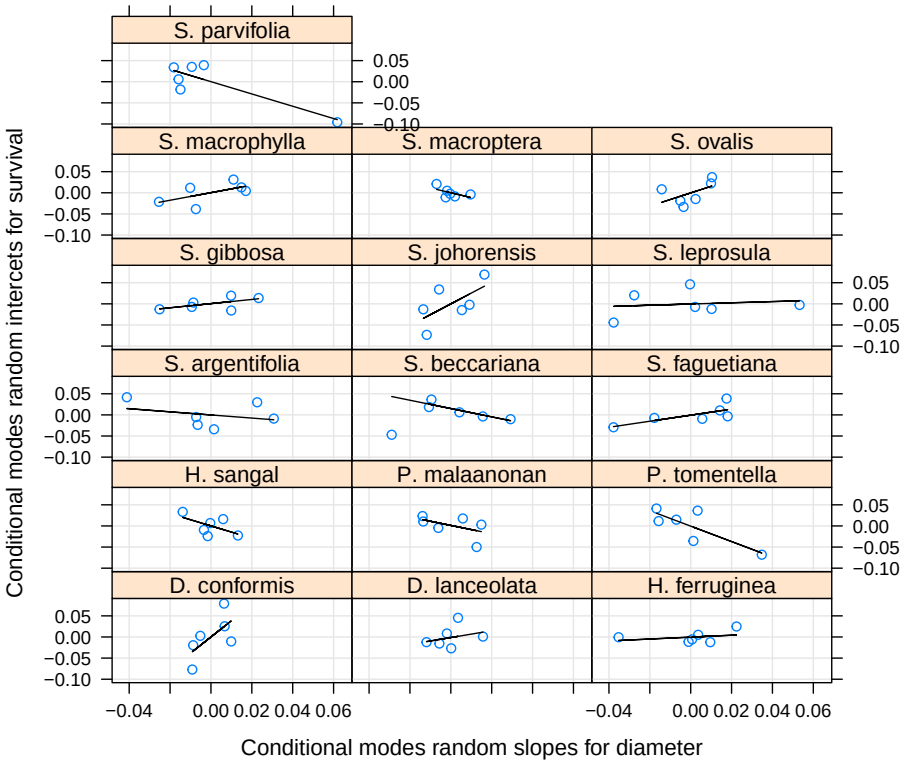


Figure 23: Scatterplot of the estimated local mean survival rates and local mean growth rates at five years after planting. Each dot on this figure represents a species by site combination (e.g. species S.ovalis and plot 3). Robust regression lines are added.

Figure 23 indicates that, for a given species, a good plot to grow is not necessarily at good plot to survive. Indeed, there is no common pattern shared across panels.

Summary

We showed how to use the estimates of two models fitted on a different response variables to inspect very specific questions in a powerful (conditional) way. It worth noting that, to test the growth-survival trade-off we could also estimate the mean survival rates and estimate the mean growth rates for each species. To do that we would first estimate the mean survival rate and then fit a linear model and get the growth estimate (i.e. slopes for time) for each species. Finally, we would correlate them. So the full analysis would have four steps:

1. fit a model to inspect the usual diversity-function questions
2. estimate mean survivals
3. fit a model to estimate growth rates
4. correlate them

What we propose is shorter and more powerful. We fit a single model and use the

estimated parameters to answer the survival-growth trade-off question. The advantage is that we are running a single analysis, but also that we get conditional estimates. This has two main advantages over a marginal analysis i) estimates are more precise as the effect of confounding factors such as plot is removed and ii) we can test more specific questions (such as the last hypothesis tested).

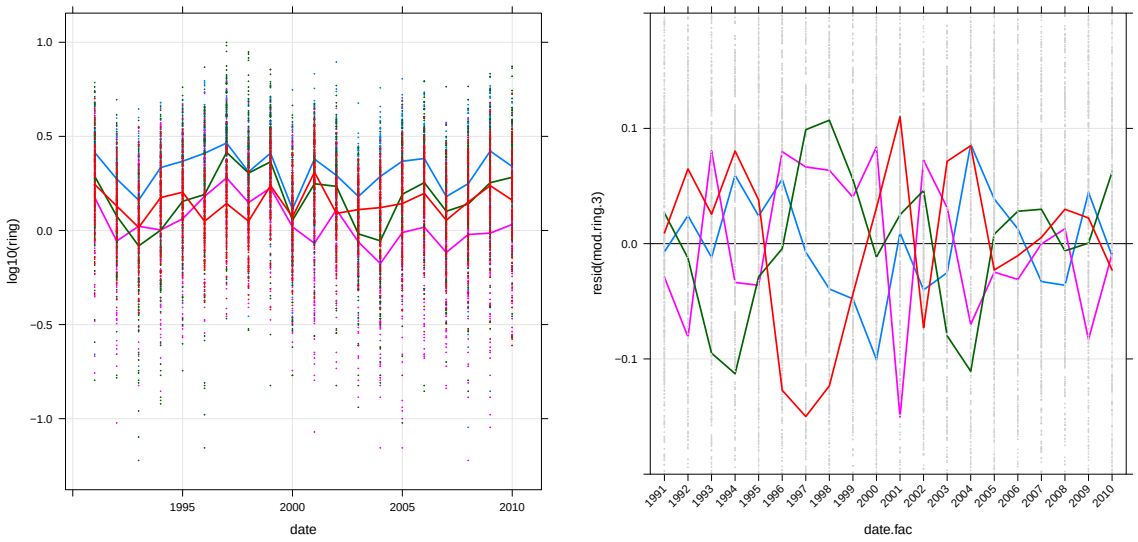
Finally, note that we can extend the approach presented here to correlated estimated effects of different types of data. An an example, survival could also be modelled with a binomial model. From this latter model we can easily get random and fixed effects to correlate with exactly the same procedure.

4 Extending asynchrony analyses: Clustering

In this last section we are using multivariate techniques to find out whether there are groups of observations that clustered together. To make an example, we may want to inspect how the 16 species used in the Sabah Intensive Experiment group together. The are several features that we could use to group the 16 species, species-specific responses to spatial heterogeneity (measured here with the random intercepts `plot:species.short`) is one example. Alternatively, we could also use species-specific temporal fluctuations. It is worth mentioning that although many of the examples presented below involve grouping species, these techniques can be used to find groups with any categorical variables (e.g. species richness).

This section is strongly linked with the concept of asynchrony (see Appendix B). There we tested, with the Czech republic data set, whether species react similarly to temporal fluctuations and spatial heterogeneity. Here we continue this analysis, and inspect how species group together. There are two main reasons why this section is to be found here instead of in the asynchrony appendix. First, the techniques used here are multivariate techniques. Secondly, we want to make clear that this way to continue the asynchrony analysis is one possible option, but that it is not tight to our approach to test for asynchrony.

Finding groups of species based on their temporal fluctuations
In Appendix B, we showed that the four species involved in the Czech Republic Experiment responded in a asynchronous way to temporal fluctuations. To better remember this situation we reproduce two graphs previously used.



(a) Response variable of the Czech Republic data set plotted against date

(b) Species-specific year deviations

Figure 24: Temporal fluctuations and species-specifc deviation for the Czech republic data set. Species identity is higlighted with colours.

Graph (a) in Figure 24 shows that species fluctuate over time. We already discussed that we can decompose these fluctuations in common-year effects (shared among all species) and species-specific year deviations (see graph (b) above). At this stage we may want to look at the pairwise correlations of species to see which are alike and which are not. When the number of species is low, as in this example, we can simply estimate pairwise correlation (of e.g. species-specific year deviations) and display them graphically (see Appendix B). However, we may deal with many species or, alternatively, we may also be interested in how species group together. We may wonder whether species of a certain taxonomic group react more similarly to annual fluctuations than others. To answer this type of questions we can use clustering techniques (Everitt and Hothorn, 2011). We are now going to use the species-specific year deviations to cluster species. To do that we refit the model fitted to the Czech Republic data set and use it to make predictions for each year-species combination. These predicted values will then be used to cluster species.

```
mod.ring.0 <- lmer(log10(ring) ~ species * (density.stand + diversity.stand +
                                            date.fac + age + basal.area) +
                     (1 | tree) + (1 | composition) + (1 | site),
                   data = d.20)
```

The matrix printed below shows the predicted ring width values for each year-species combination (values are in the logarithmic scale). To make an example, European beech in year 1991 had a mean log-ring width of 0.43 (i.e. $10^{0.43} = 2.69$ millimetres in the back-transformed scale).

```
( d.pred.ring.wide )

                 1991   1992   1993   1994  1995 1996 1997 1998 1999
Fagus sylvatica 0.43  0.286  0.175  0.348 0.382 0.42 0.48 0.32 0.42
```

```
Larix decidua   0.10 -0.122 -0.036 -0.050 0.015 0.14 0.24 0.12 0.20
Picea abies     0.19 -0.012 -0.165 -0.076 0.082 0.12 0.35 0.24 0.30
Quercus petraea 0.33  0.212  0.100  0.259 0.287 0.13 0.23 0.14 0.32
                   2000   2001  2002   2003   2004   2005  2006   2007
Fagus sylvatica  0.1243  0.389 0.302  0.191  0.297  0.376 0.389  0.183
Larix decidua   -0.0044 -0.085 0.099 -0.069 -0.177 -0.006 0.025 -0.105
Picea abies     -0.0052  0.192 0.181 -0.066 -0.097  0.154 0.218  0.065
Quercus petraea  0.1578  0.394 0.176  0.196  0.207  0.228 0.281  0.139
                  2008   2009   2010
Fagus sylvatica  0.250  0.420 0.3280
Larix decidua   -0.024 -0.032 0.0049
Picea abies      0.104  0.206 0.2386
Quercus petraea  0.235  0.320 0.2407
```

This matrix is then used to compute a distance matrix with the `dist()` function.

```
( dist.ring.sp <- dist(d.pred.ring.wide) )
```

```
                Fagus sylvatica Larix decidua Picea abies
Larix decidua              1.47
Picea abies                1.06          0.62
Quercus petraea            0.55          1.13        0.77
```

Not unexpectedly European beech and sessile oak are the most similar species (distance is the lowest, i.e. 0.55). It is worth mentioning that this distance matrix is the analogous inverse of the corresponding correlation matrix. Given a distance matrix, we can use multi-dimensional scaling two visualise the differences among species in two dimensions. Put simply, we use this technique to reduce the number of dimensions of our data from 20 (i.e. the number of years) to two, such that we can easily plot that[8].

```
## 1. applying multi-dimensional scaling to the distance matrix
mds.ring.sp <- cmdscale(dist.ring.sp)
##
## 2. plotting the results
plot(mds.ring.sp, type = "n",
     xlim = c(-1, 1))  ## to make names fit
text(mds.ring.sp, rownames(mds.ring.sp),
     cex = 1.5)
```

[8]Note that the plot shown below is equivalent to a biplot produced with a Principal Component Analysis (Venables and Ripley, 2013).

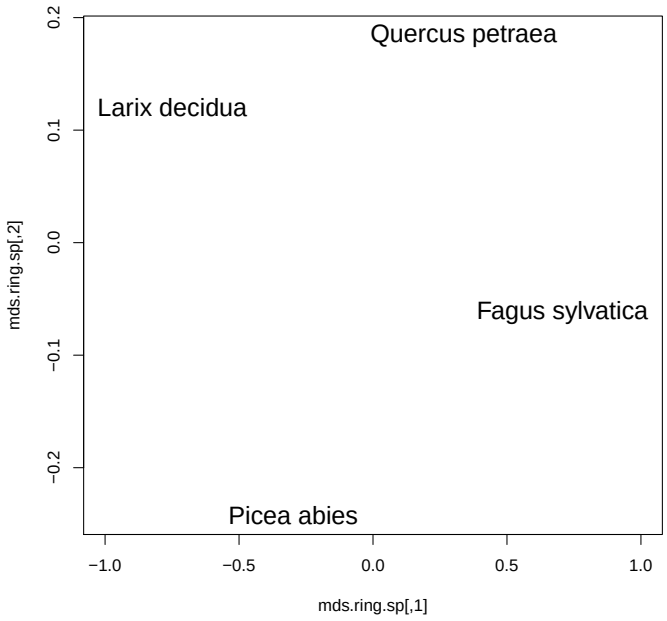


Figure 25: Graphical representation of the results of the multi-dimensional scaling on annual fluctuations (Czech Republic data set).

On this graph there is no evidence for some species to group together. This does not come unexpected as there are only four species. The main purpose of this example is to introduce the technique with a small data set.

Finding groups of years based on species responses

In the previous example, we have been looking at how species group together based on their year responses. We may turn this question and check how years group together based on their effects on species. To achieve that, we simply transpose the matrix of fitted values before computing distances so that we are estimating the distances among years.

```
## 1. transposing the data matrix
t.d.pred.ring.wide <- t(d.pred.ring.wide)
##
## 2. computing the distance matrix for species
dist.ring.year <- dist(t.d.pred.ring.wide)
##
## 3. applying multi-dimensional scaling
mds.ring.year <- cmdscale(dist.ring.year)
##
## 4. plotting the results
plot(mds.ring.year, type = "n")
text(mds.ring.year, rownames(mds.ring.year), cex = 1.5)
```

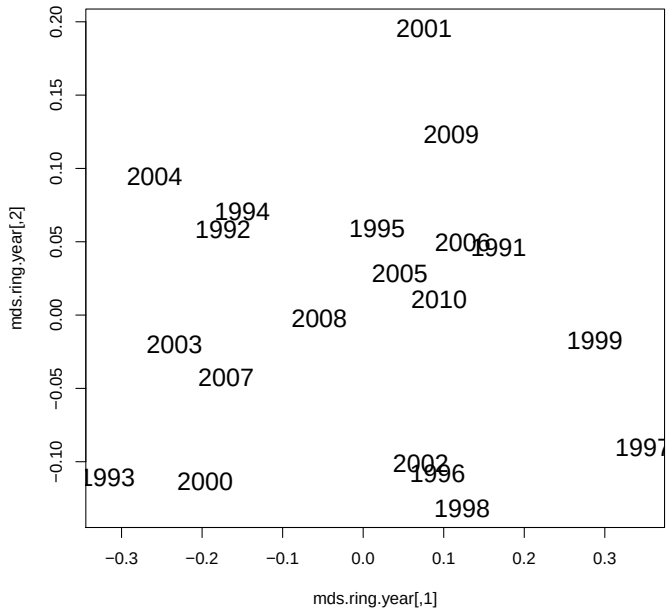


Figure 26: Graphical representation of the results of the multi-dimensional scaling on species (Czech Republic data set).

By looking at Figure 26 we notice that some years group together (e.g. 1996, 1998 and 2002 bottom centre) and that others seems to be quite distinct (e.g. 2001 top centre). This grouping has to be interpreted in terms of species responses to annual fluctuations. So we may ask ourselves why is it 2001 that different from the other years? Was that a drought year? This graph is useful to formulate hypotheses about the climatic variables that explain species responses to annual fluctuations.

Finding groups of species based on their spatial preferences

Another way to cluster species is to look at their responses to heterogeneous environments. By using the conditional modes of the model fitted to diameter above (i.e. `mod.diam.3`), we can group species according to their local preferences for growth. To do that, we are using the random slopes estimated above (i.e. species-specific local growth rates). Indeed, in that model we allowed each species to have a different growth rate in each plot. Here we use hierarchical clustering which is considered to be the best technique to find groups (Venables and Ripley, 2013).

```
hclu.sbe.int.grow <- hclust(d = dist.sbe.plotSp)
plot(hclu.sbe.int.grow, hang = -1)
```

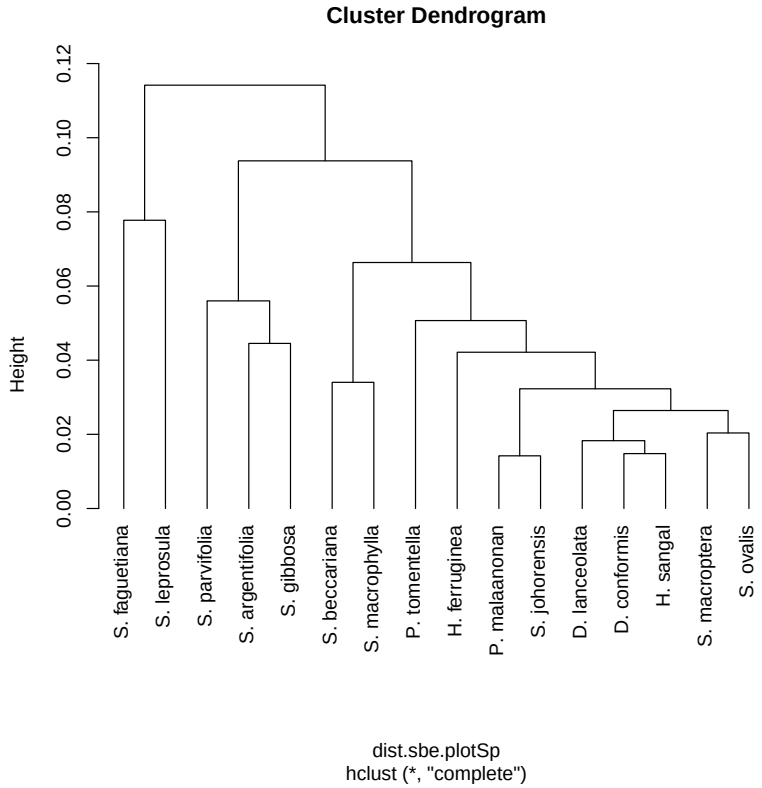


Figure 27: Graphical representation of the results of hierarchical clustering on species-specific local growth rates (Sabah Intensive Experiment).

The dendrogram shown on Figure 27 illustrates how species group together based on their plot-specific growth rates. We may wan to compare this dendrogram with others (e.g. taxonomy). Other clustering techniques exist. For example, Model-Based Clustering enables to formally test the number of groups present in the data (Everitt and Hothorn, 2011).

Final remarks about clustering

As mentioned above, the first example shown in this section could be done without fitting a model to the data. We could simply take the mean value of each species-year combination and use a clustering technique on it. However, It is clear that the last example cannot be done by using the raw data

Reproducibility

Checkpoint date was set to 2016-05-01.

```
R version 3.3.2 (2016-10-31)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Debian GNU/Linux 8 (jessie)

locale:
 [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
 [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
 [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] robustbase_0.92-6   lme4_1.1-12         Matrix_1.2-7.1
[4] MASS_7.3-45         latticeExtra_0.6-28 RColorBrewer_1.1-2
[7] lattice_0.20-34     knitr_1.14

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.7    digest_0.6.10  grid_3.3.2     nlme_3.1-128
 [5] formatR_1.4    magrittr_1.5   evaluate_0.10  highr_0.6
 [9] stringi_1.1.2  minqa_1.2.4    nloptr_1.0.4   splines_3.3.2
[13] tools_3.3.2    stringr_1.1.0  DEoptimR_1.0-6
```

References

D. M. Bates. lme4: Mixed-Effects Modeling with R. *URL http://lme4. R-forge. R-project. org/book*, 2010.

B. Everitt and T. Hothorn. *An Introduction to Applied Multivariate Analysis with R*. Springer Science & Business Media, 2011.

A. Hector, C. Philipson, P. Saner, J. Chamagne, D. Dzulkifli, M. O'Brien, J. L. Snaddon, P. Ulok, M. Weilenmann, G. Reynolds, et al. The Sabah Biodiversity Experiment: A long-term test of the role of tree diversity in restoring tropical forest structure and functioning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1582):3303–3315, 2011.

L. Held and D. Sabanés Bové. *Applied Statistical Inference*. Springer, Berlin Heidelberg, 2014.

M. Koller. *robustlmm: Robust Linear Mixed Effects Models*, 2016. URL `https://CRAN.R-project.org/package=robustlmm`. R package version 1.8.

M. Loreau and A. Hector. Partitioning selection and complementarity in biodiversity experiments. *Nature*, 412(6842):72–76, 2001.

V. Todorov, A. Ruckstuhl, M. Salibian-Barrera, T. Verbeke, M. Koller, and M. Maechler. *robustbase: Basic Robust Statistics*, 2016. URL `https://CRAN.R-project.org/package=robustbase`. R package version 0.92-6.

S. L. Tuck, M. J. O'Brien, C. D. Philipson, P. Saner, M. Tanadini, D. Dzulkifli, H. C. J. Godfray, E. Godoong, R. Nilus, R. C. Ong, B. Schmid, W. Sinun, J. L. Snaddon, M. Snoep, H. Tangki, J. Tay, P. Ulok, Y. S. Wai, M. Weilenmann, R. Glen, and A. Hector. The value of biodiversity for the functioning of tropical forests: Insurance effects during the first decade of the Sabah Biodiversity Experiment. *Proceedings of the Royal Society B*, 283 (1844):20161451, 2016.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-PLUS*. Springer Science & Business Media, 2013.

Chapter 4

Does no-till agriculture influence crop yields?

4.1 Preface

Authors: Matteo Tanadini[1] & Zia Mehrabi[2]
1. Department of Plant Sciences, University of Oxford, South Parks Road, OX1 3RB.
2. Liu Institute For Global Issues, Institute for Resources, Environment and Sustainability, & Centre for Sustainable Food Systems, University of British Columbia, Vancouver, British Colombia, Canada, V6T 1Z2.

Author contributions: MT and ZM designed the project, MT conducted the analysis, ZM and MT wrote the paper.

Target Readership: Ecologists and biologists.

Context: This chapter illustrates the importance of a correct data visualisation and, at the same time, the importance of a correct practical and biological interpretation of the results of a statistical analysis. This is particularly relevant for those fields of research, such as biodiversity ecosystem-functioning experiments, that influence policy-makers decisions.

Article status: This publication is in the second stage of the review process in Nature. After the first review process the paper re-directed its focus on reproducibility. The first version of the paper, shown below, better fits into this thesis.

4.2 Main text

No-till is an agricultural practice widely promoted by governments, development agencies, and agricultural organisations worldwide. However, the costs and benefits to farmers adopting no-till are hotly debated (Friedrich et al., 1941; Derpsch et al., 2014; Powlson et al., 2014; Stevenson et al., 2014; Giller et al., 2015). Using a meta-analysis of unprecedented study size, Pittelkow et al. (2015a) reported that adopting no-till results in average yield losses of -5.7%. They claimed that farmers can avoid this yield loss in dry environments with the added effort of using crop rotation and crop residue retention. Their conclusion was that yield losses are to be expected for many of the smallholders currently adopting no-till practices. In a re-evaluation of their analysis, we found that they overly biased their results towards showing that no-till negatively impacts yields, and that they overlooked the practical significance of their findings. Strikingly, we find that all of the variables they used to predict variation in crop yields under no-till actually perform no better than random. Our results suggest that their meta-analysis cannot be used as the basis for evidence-based decision-making in the agricultural community.

There is never a perfect analysis, and every analysis involves a large number of decisions, some of which are often arbitrary. However, our re-evaluation found four major issues with Pittelkow et al.'s meta-analysis, which are fundamental, based on standard statistical theory and best practice. We show how methodologically accounting for these issues produces results that strongly challenge the claims and conclusions of Pittelkow et al.

First, we found that Pittelkow et al.'s estimates were biased to showing no-till leads to a yield loss. To obtain their -5.7% yield loss figure, they log transformed the ratio of crop yield under no till and conventional till for each experiment, calculated a weighted mean effect of no-till for all these experiments and then backtransformed the result using the exponential function. However, due to Jensen's Inequality, using the exponential function to backtransform a log-ratio is both biased and inconsistent (Neyman and Scott, 1960). In practical terms, this means that the exponentially

back-transformed estimates are underestimated (i.e. too small), even when the sample size is very large. The magnitude of the bias depends on the variance of the random variable itself, and in many cases is likely to be negligible. However, for the case of Pittelkow et al., there was massive variation in the ratio of yields under no-till and till, which forced large bias in their results towards showing a negative effect of no-till on crop yields. Simply calculating the weighted mean yield ratio in the untransformed scale produces a value of -2.4%, which is around half of the crop yield loss reported by Pittelkow et al. (-5.7%).

Second, Pittelkow et al. arbitrarily chose to discard observations from their data. They removed all data points further away than five standard deviations from the weighted mean, representing about 0.5% of the observations, from their analysis. They did not state any reason for how the threshold of five standard deviations was chosen (we are also not aware of any such rule). Because of the long-tailedness of the data, the removal of these observations had an important effect on the results. Including the observations that had been arbitrarily removed from the data leads to a yield ratio of only -1.2%, 80% less severe than the figure originally reported.

Third, we found that the statistical significance testing reported by Pittelkow et al. was invalid. To calculate the confidence intervals around, and statistical significance of the effect of no-till on crop yields Pittelkow et al. used randomization and bootstrapping techniques. A central assumption of the methods they used is the independence of observations (Davison and Hinkley, 1997). However, the 5,493 experiments in their dataset arose from 609 different studies and so are unlikely to be independent. Our analysis clearly shows that observations gathered in the same study are more similar than observations between studies, which in turn invalidates Pittelkow et al.'s statistical tests. There are established ways to account for this non-independence (see Appendix A). Put simply, Pittelkow et al.'s statistical significance testing should be disregarded. In fact, the power of this data set is so large, due to the large sample size that its proper treatment of practical significance is critical.

Finally, we found that Pittelkow et al. overlooked the practical significance of their

findings. Forty-three percent of all experiments in the data set show the same or higher yields under no-till. The majority of yield ratios (e.g. 95% quantiles), lay between a halving in yields, to a three quarter increase in yields under no-till (Figure 4.1). Whilst Pittelkow et al. did qualitatively acknowledge that the effect of no-till on crop yield was variable, they did not plot this variation, nor discuss its consequences for their analyses and results. Instead, they committed to the claim that no-till had a negative effect on crop yields, and that the experiments showing yield gains under no-till could be explained by whether or not farmers chose to include crop rotation and crop residue management alongside no-till practices, or whether this was done in arid environments. However, we found that crop rotation, residue management, and indeed even aridity, are actually of very little practical significance for crop yields under no-till. In fact, all these factors taken together are no better than random at predicting yields under no-till (Figure 4.2).

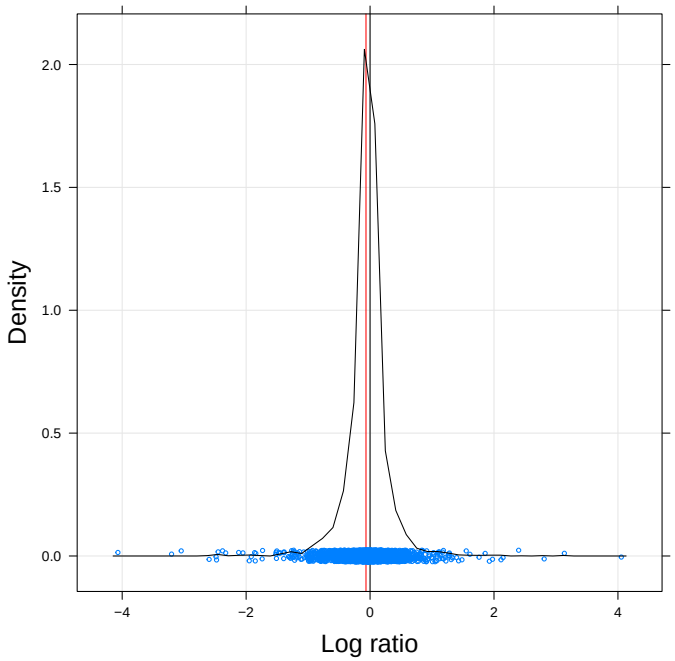


Figure 4.1: This smoothed histogram (also called density plot) shows the distribution of of the log(ratio) (i.e. the logarithm of the ratio between no-till and till yields). Pittelkow's statement that no-till negatively affects crop yields is practically irrelevant. The weighted mean of the data (red line) can hardly be distinguished from the null hypothesis of no difference between yields under no-till and till (black line on zero). The underlying variation in the data is so large that the estimated mean effect is actually a poor representation of experiments in the dataset, and bears little practical relevance to the farming community.

In summary our analysis suggests that Pittelkow et al.'s claim that no-till on average reduces crop yields by 5.7% is incorrect. We found that mean yields under no-till are actually up to 80% less negative than those they reported (-1.2%). This in addition to the large variation in experimental outcomes, and the poor explanatory power of the variables coded in of Pittelkow et al.'s analysis strongly challenges their claims. Our re-evaluation does not corroborate their claim that farmers will experience yield deficits under no-till, or that rotation and residue management will increase yields under no-till. With the large number of predictors that could possibly drive differences in yields under no-till, we suggest that careful attention must be put towards better variable coding, model building, and model interpretation if comparisons between experiments and practically relevant recommendations can be made in future. Even once this has been done, farmers should be cautious of the data sources underlying recommendations.

For example, studies from sub-Saharan Africa contribute less than 3% of the current dataset. The applicability of these data is largely limited in its regional and crop level representation to two crops: maize and wheat, and the USA.

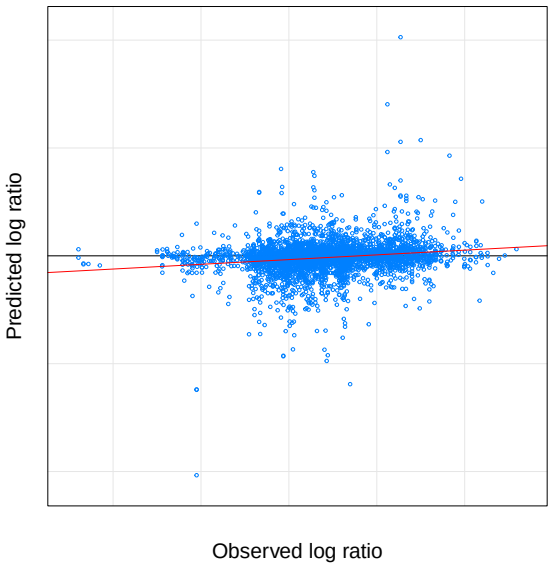


Figure 4.2: This scatterplot illustrates the relationship between predicted and observed values of a linear model where all predictors used by Pittelkow are used. These explanatory variables do not explain any of the variation in their data set. The crop rotation, crop residue management, and aridity variables used as the basis for Pittelkow et al.'s analysis are actually extremely poor predictors of crop yields under no-till. A multiple regression mixed model accounting for all of these variables is no better at predicting experimental outcomes than random In other words, rotation and residue management are unable to predict crop yields under no-till.

There are of course, other, important reasons why farmers may wish to employ crop residue management and crop rotation alongside no-till in dry environments (e.g. to control pests, soil erosion and soil moisture retention). Similarly, there are other, potential limiting factors to the uptake of no-till agriculture by smallholders in some regions of the world (e.g. access to herbicides, implements; Stevenson et al. (2014); Giller et al. (2015)). Our re-evaluation has highlighted some of the key risks involved in using agricultural meta-analyses to make recommendations to farmers, and hopefully sets a road map for people using large meta-analyses to avoid making the same mistakes in future.

Methods

In line with common standards in meta-analytical practice we have made all our analysis fully reproducible (Koricheva and Gurevitch, 2014). We have also included detailed methods to deal with the all of the four major issues we have outlined here, and additional others, which did not fit in this response, in the Supplementary Information. We note that newer publications following the original Pittelkow et al. study do not overcome the problems we have raised here (Pittelkow et al., 2015b; Lundy et al., 2015).

4.3 Appendices

Appendix A, starting at page 269, reiterates and supports the claims made in the main text. Appendix B, starting at page 291, shows how the data set used in the analysis was built. Appendix B is present to enable the full reproducibility of our analysis, but is not fundamental to the chapter.

Appendix A: Supporting elements

Chapter 4

Contents

1 Aims of this document

This document provides the supporting evidence for the claims made in our *response* to the *paper* of Pittelkow et al. (Pittelkow et al., 2015). It was created with `knitr` (Xie, 2016), a software that combines the typesetting system LaTeX and the statistical computing language **R** (R Core Team, 2015). This format increases the readability of of our analysis, and conveniently includes the images, **R**-codes and associated outputs inset within the verbal logic of our approach.

This document is split into seven sections. In Section 2 we illustrate how we make our analysis fully reproducible. Section 3 imports data into **R** and checks that it is in the format required for reproducing the analysis of Pittelkow et al. Section 4 re-states our claims. Section 5 illustrates how our criticisms of Pittelow et al. can be easily addressed through an alternative analysis. Section 6 discusses whether the alternative analysis is needed and whether considering all alternative analyses as equivalent is sensible. Finally, Section 7 draws conclusions about this analysis and the importance of the methodological considerations of our work for future analysis of large data sets.

2 Making everything fully reproducible

The analysis presented by Pittelkow et al. is not directly reproducible. Essential methodological information on their statistical procedure is missing from their *paper*, and the data set accompanying their *paper* on-line is not provided in the format needed to reproduce their analysis. Often, it is very challenging to provide all of the details needed to make an analysis fully reproducible. Indeed, journal text length requirements mean it is often not feasible to describe the statistical analysis in great detail. In addition, it may be very challenging to describe complex analyses with the appropriate scientific wording, whilst still maintaining a broad readership.

To overcome this problem, we make use of the **R** packages `knitr` and `checkpoint` (Xie, 2016; Corporation, 2016). As mentioned above, the package `knitr` makes it possible to produce a dynamic document (here a pdf) that contains all the steps required to analyse the data. Comments are interspersed with **R** commands and relative output. The reader is provided with data and software versions to fully reproduce the analysis.

Sometimes, even if data are provided with **R**-code, re-running an analyses can yield different results to those reported in primary manuscripts. In these cases, it is difficult to understand why results differ: It may not be clear if the data provided are not the ones used for the analysis, whether the code provided contains errors, or whether different versions of software are to blame. A very easy way to overcome this problem is to use the **R** package `checkpoint`. As the help page of the `checkpoint()` function states, it "creates a local library into which it installs a copy of the packages required by your project as they existed on CRAN on the specified snapshot date. Your **R** session is updated to use only these packages". This is tremendously useful as it enables any user to fully reproduce our analysis by simply running the available **R** script. Indeed, those willing to reproduce our analysis do not need to bother about whether the version of packages installed on their computers matches the ones here. `checkpoint()` will install all packages versions for you that we used in our analysis.

```
require(checkpoint)
checkpoint(snapshotDate = "2016-01-01")
```

Note that the **R** version used here is 3.3.2 (2016-10-31). Using other versions of **R** should not have any influence on the results obtained. The function call just above installed all packages used in this document, as available on January first 2016 in the "home" directory of our computer.

3 Importing data into R

As we noted in Section 2, the data provided by Pittelkow et al. with their *paper*, is not in the proper format to reproduce their analysis. For example, the response variable, as well as the weights used in their analysis need to be created. In addition to that, the data is not provided a format directly usable in **R**. Supporting Information B describes the procedure we carried out to re-create the data set needed for reproducing their analysis. This data set is stored in the file named "DataNatureReady.rds". This file is available at `NatureWebsite`. The format we use for the data set (`.rds`) is very convenient for sharing data, especially across operating systems (see Supporting Information B for more details).

```
d.nature <- readRDS("DataNatureReady.rds")
str(d.nature)
```

```
'data.frame': 5492 obs. of  26 variables:
 $ study.new                : Factor w/ 608 levels "Zhang et al..2007.Acta Agriculturae Scandinavi
 $ Author                   : Factor w/ 515 levels "Aase  et al.",..: 1 1 1 1 2 2 2 2 2 2 ...
 $ Year                     : Factor w/ 44 levels "1981","1983",..: 20 20 20 20 29 29 29 29 29 29
 $ Journal                  : Factor w/ 132 levels "Acta Agriculturae Scandinavica Section B",..:
 $ Country                  : Factor w/ 62 levels "Algeria","Argentina",..: 60 60 60 60 33 33 33 3
 $ Location                 : Factor w/ 534 levels "Adana, Cukorova University Ag Experiment Stati
 $ Latitude                 : num  48.3 48.3 48.3 48.3 32.5 ...
 $ Longitude                : num  -105 -105 -105 -105 36 ...
 $ Reps...conv..till.       : int  3 3 3 3 2 2 2 2 2 2 ...
 $ Reps...no.till.          : int  3 3 3 3 2 2 2 2 2 2 ...
 $ Crop                     : Factor w/ 48 levels "alfalfa","barley",..: 47 47 47 47 3 3 3 2 2 2 .
 $ Study.Duration           : int  1 2 3 4 1 1 1 1 1 1 ...
 $ Irrigation.              : Factor w/ 3 levels "","no","yes": 2 2 2 2 3 3 3 3 3 3 ...
 $ Rotation.                : Factor w/ 3 levels "cover","no","yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ Residue.Mgmt             : Factor w/ 4 levels "","burned","removed",..: 1 1 1 1 4 4 4 4 4 4 ...
 $ Aridity.Index            : num  0.334 0.334 0.334 0.334 0.236 0.236 0.236 0.236 0.236 0.236 ...
 $ Yield...conv..till..kg.ha: int  2755 2567 2405 2439 678 696 999 5854 6754 9665 ...
 $ Yield...no.till..kg.ha   : int  2782 2694 2439 2788 704 843 1109 6382 7646 10531 ...
 $ yield.ratio.new          : num  1.01 1.05 1.01 1.14 1.04 ...
 $ log.ratio.new            : num  0.00975 0.04829 0.01404 0.13374 0.03763 ...
 $ weights.new              : num  1.5 1.5 1.5 1.5 1 1 1 1 1 1 ...
 $ no.obs.new               : int  4 4 4 4 6 6 6 6 6 6 ...
 $ weights.study.new        : num  0.375 0.375 0.375 0.375 0.167 ...
 $ rotation.new             : Factor w/ 2 levels "not rotated",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ resid.man.new            : Factor w/ 2 levels "removed","retained": NA NA NA NA 2 2 2 2 2 2 ...
 $ outlier.new              : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

```
head(d.nature, n = 2)
```

```
                                study.new       Author Year
```

```
1 Aase  et al..1997.Soil & Tillage Research  Aase  et al. 1997
2 Aase  et al..1997.Soil & Tillage Research  Aase  et al. 1997
                  Journal Country       Location Latitude Longitude
1 Soil & Tillage Research      USA Montana, Froid       48      -105
2 Soil & Tillage Research      USA Montana, Froid       48      -105
  Reps...conv..till. Reps...no.till.  Crop Study.Duration Irrigation.
1                  3               3 wheat              1          no
2                  3               3 wheat              2          no
  Rotation. Residue.Mgmt Aridity.Index Yield...conv..till..kg.ha
1        no                      0.33                      2755
2        no                      0.33                      2567
  Yield...no.till..kg.ha yield.ratio.new log.ratio.new weights.new
1                   2782               1        0.0098         1.5
2                   2694               1        0.0483         1.5
  no.obs.new weights.study.new rotation.new resid.man.new outlier.new
1          4              0.38  not rotated          <NA>       FALSE
2          4              0.38  not rotated          <NA>       FALSE
```

The data set `d.nature` contains all the variables present in the Excel file provided by Pittelkow et al. with their *paper*, along with several other variables necessary for the analysis. Variables created by us are named with a suffix `.new`.

Pittelkow et al. discarded a small portion of the data before running their analysis. This is a point of discussion in our *response*, and is discussed in detail at page 8. The subset of data used by Pittelkow et al. is stored in a data set named `d.subset`.

```
d.subset <- subset(d.nature, subset = !d.nature$outlier.new)
```

The response variable used in the analysis of Pittelkow et al. is defined here. A more detailed visualization of the data and a discussion of the consequences of the variation in the response variable for making inferences from this data is given in Section 5. The response variable is the logarithm of the crop yield ratio between "no-till" and "till" agriculture (hereafter "log ratio"). We have displayed data with a density plot below [1]. The package `lattice` is used to visualize data (Sarkar, 2015).

```
require(lattice)
dp.sub <- densityplot(~ log.ratio.new, data = d.subset,
                      type = c("p", "g"), cex = 0.5, col.line = "black",
                      xlab = list(label = "Log ratio", cex = 1.5),
                      ylab = list(cex = 1.5))
dp.sub
```

[1] Readers not familiar with this graphical tool, can view it as an histogram with a very large number of bins.

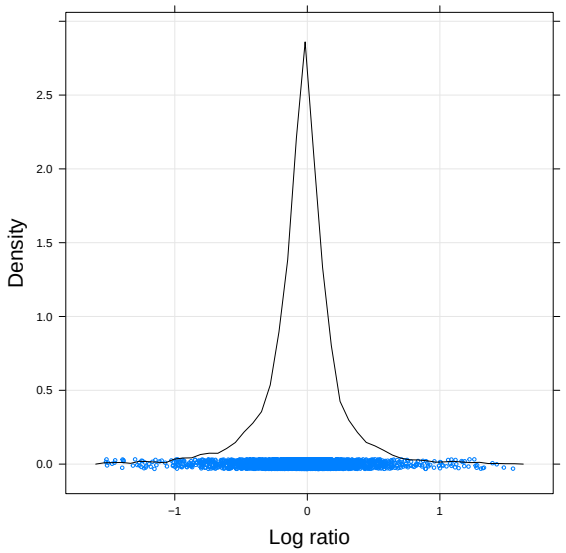


Figure 1: Density plot of the log ratio (subset data).

This graph shows the distribution of the response variable. A negative value indicates lower yield in no-till, while zero means same yield between no-till and conventional till practices (i.e. a yield ratio of one). The "log ratio" does not seem to follow a normal distribution. This is confirmed by the quantile-quantile plot shown at page 8.

4 Criticisms of the analysis published by Pittelkow et al.

Here, we re-iterate the claims made in our *response*, namely that:

1. The analysis was biased towards showing that no-till had a negative effect on yields. This is due to the use of a naive and biased back transform.
2. The data was arbitrarily subsetted. This exacerbated the bias towards showing no-till had a negative effect on yield, and as we will show below, was not justified nor sufficient for dealing with these data.
3. The statistical inference presented is invalid because the assumption of independence of the observations required for the analysis was violated.
4. Practical and statistical significance were mismatched. Given the large sample size the power is very high and all tests carried out on this data set are likely to be statistically significant, even though the models that were proposed as scientifically useful actually do a very poor job of explaining the underlying variation in the data.

We address each of these criticisms in turn. We will focus on the estimation of the overall effect of no-till on crop yields. However, all of our criticisms are equally held against every subgroup analysis reported by Pittelkow et al.. In Section 5, we show

how it is possible to take all of these issues into account by running an alternative analysis with a robust multiple linear regression framework.

4.1 Biasedness of the backtransformation

4.1.1 Jensen's inequality

Pittelkow et al. estimated the mean value of the "log ratio" and then backtransformed it by using the exponential function. This backtransformation is biased (Neyman and Scott, 1960; Duan, 1983). In particular the backtransformed value is too small. This was proven in 1906 and is known as Jensen's inequality (Jensen, 1906).

The magnitude of the bias depends on the variance of the random variable itself (here this is the "log.ratio"). As we will see on page 7 the variance here is very large, and thus the bias is very large as well. Because of this issue, analysing the log ratio of no-till/till is biased towards showing that no-till negatively effects crop yields.

4.1.2 Possible solutions

Fortunately unbiased backtransformations exist. If the data follows a normal distribution the proper backtransformation is simply $exp(\hat{y} + \hat{\sigma_\varepsilon})$. Where $\hat{y}$ is the estimated mean value and $\hat{\sigma_\varepsilon}$ is the estimated variance of the random variable (i.e. the log ratio here). This procedure yields unbiased values and is very handy as it applies to linear regressions as well (see **R**-code below) (Neyman and Scott, 1960). If data does not follow a normal distribution, other methods have been developed (again see **R**-code below) (Duan, 1983).

```
lm.0 <- lm(log(y) ~ x1 + x2 + x3)
# 1) Data following a normal distribution:
sigma.sq <- summary(lm.0)$sigma^2
exp(fitted(lm.0) + sigma.sq / 2)
##
# 2) Data that does not follow normal distribution (the Smearing estimate):
exp(fitted(lm.0) * (1 / nobs(mod.2) * sum(exp(resid(lm.0)))))
```

For the analysis of Pittelkow et al. the solution is trivial. Because just a mean is being estimated, there is no need to first log-transform the yield ratio, and then back transform it. The mean value can simply be directly estimated. It is obvious that given the skewness of the distribution of the untransformed variable (i.e. the yield ratio) the mean value may not be a meaningful location parameter. This is true regardless of whether or not data is transformed prior to estimation.

So the unbiased weighted mean value of yield ratio is[2]:

```
100 * weighted.mean(d.subset$Yield...no.till..kg.ha / d.subset$Yield...conv..till..kg.ha,
            w = d.subset$weights.study.new) - 100

[1] -2.4
```

[2]Note that the results are reported in %.

We thus obtain a loss of -2.4% instead of a biased loss of -5.7%. In this very simple case, we can quantify the bias as being 3.3%. In this case the bias is very large and strongly affects the conclusions of the study.

We believe that a publication by Hedges et al. (Hedges et al., 1999), which Pittelkow et al. cites, may have been partially responsible for their use of a naive back transform. Hedges et al. wrongly state that when analysing log ratio the bias can be ignored if the sample size not small:

> "However, since the magnitude of the bias depends upon the variance of the weighted mean (which will typically be small if the number of studies and the precision of their estimates are not both small), the bias will usually be slight."

This statement is incorrect as the magnitude of the bias depends on the variance the random variable itself and not on the variance of its (weighted) mean (Neyman and Scott, 1960). Thus, even is the sample size is large the problem remains. As a matter of fact, naively backtransforming (i.e. using the exponential function) is not only biased, but also inconsistent[3] (Duan, 1983).

4.1.3 The utility of a log-transform

The reader may wonder why log-transforming is ever applied when it creates the problems we have outlined above. There are in fact several reasons why log-transforming the response variable can be good practice. Whilst not the central focus of this document, for completeness, here we show why this is the case.

Variables that only take positive values, such as weights or heights, often appear to follow a log-normal distribution (Limpert et al., 2001; Limpert and Stahel, 2011), which means that they follow a normal distribution after log-transformation. These kinds of variables were called "amounts" by the famous statistician John Tukey. He introduced the "first aid transformation" to visualize and analyse this kind of data (Mosteller and Tukey, 1977). For "amounts" he suggested the use of the log-transformation (Mosteller and Tukey, 1977).

It is simple to show that if the yields were to follow a log-normal distribution here, then their log ratio would follow a normal distribution, and we could proceed with a classical linear model (or any of its extensions, e.g. a mixed effects model). This would make the statistical inference procedure very convenient. Unfortunately however, the data analysed here does not follow a normal distribution after log-transformation. This is best seen by looking at a normal quantile-quantile plot.

```
qqnorm(d.subset$log.ratio.new)
qqline(d.subset$log.ratio.new)
```

[3] This practically means that the bias is not corrected for even when sample size tends to infinity (i.e. very large sample sizes).

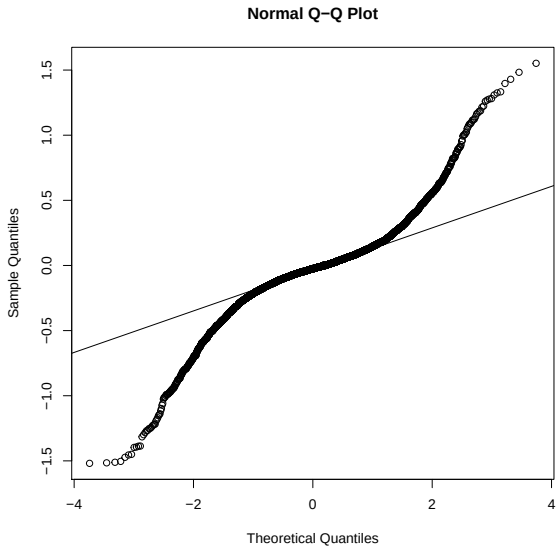


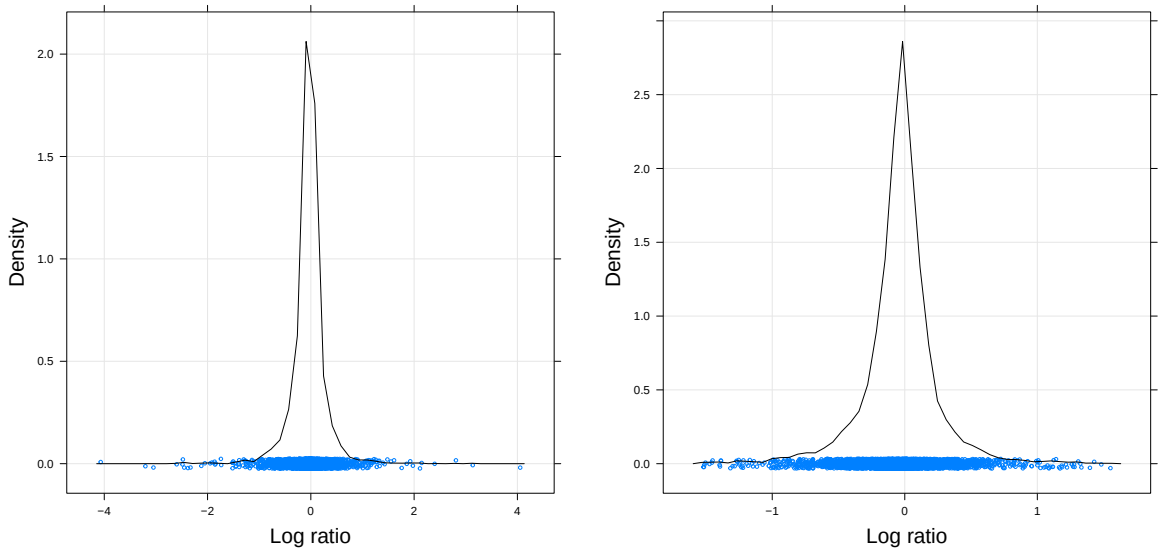
Figure 2: Quantile-quantile plot of the log ratio (subset data).

It is difficult to understand why the data does not follow a normal distribution after log-transformation. Possible reasons may lay on the fact that data comes from different studies which differ in their reporting procedures.

4.2 Arbitrary subsetting of the data

Pittelkow et al. stated they removed a small amount of the data from their analysis. They decided to remove all observations greater than 5 standard deviations from the weighted mean, representing about 0.5% of the observations in their data set. Pittelkow et al. indicate no reason why this was done and how the threshold of 5 standard deviations was chosen. We are not aware of any such rule. Because of the long-tailedness of the data, the removal of these observations has an important effect on the results. Below we visualize the density plot for the subset of data used by Pittelkow et al. and for the complete data set.

```
dp.all <- densityplot(~ log.ratio.new, data = d.nature,
                      type = c("p", "g"), cex = 0.5, col.line = "black",
                      xlab = list(label = "Log ratio", cex = 1.5),
                      ylab = list(cex = 1.5))

dp.all
dp.sub
```



(a) All observations

(b) Subset used by Pittelkow et al.

Figure 3: Density plots of the log ratio.

By looking at the two plots in Figure 3, it is clear that the problem of long-tailedness is not solved by arbitrarily subsetting the data after Pittelkow et al. As the quantile-quantile plot on page 8 undoubtedly shows, the subset of data does not follow a normal distribution.

As we will show below, even though the "extreme" observations that Pittelkow et al. chose to arbitrarily subset out of the data set only represent less than 1% of all available observations, their removal strongly affects the results. The classical way to deal with long-tailed distributed data is to use robust methods. Section 5 page 13 implements this more appropriate, approach. The **R**-code below is used to compare the weighted means of the complete data set and that of the subset[4].

```
mean.sub <- weighted.mean(d.subset$Yield...no.till..kg.ha / d.subset$Yield...conv..till..
                          w = d.subset$weights.study.new)
mean.all <- weighted.mean(d.nature$Yield...no.till..kg.ha / d.nature$Yield...conv..till..
                          w = d.nature$weights.study.new)
##
100 * mean.sub  - 100

[1] -2.4

100 * mean.all - 100

[1] -1.2
```

It is quite impressive to see that although the size of the two data sets is almost the same (i.e. 5492 vs. 5463) the differences in estimates are substantial. By arbitrarily removing a subset of the observations the loss in no-till was doubled!

To reiterate, if all available data is used and no bias through back-transformation is

[4]Note that to exclude the influence of the bias of backtransformations, we are directly estimating the mean of yield ratio.

included, we obtain a loss of -1.25%. The difference between this value, and the yield loss under no-till reported by the authors, -5.7% is considerable.

We note that the mean value is heavily affected by a small portion of the data here. It is thus questionable whether the mean is a sensible parameter to describe the distribution of the yield ratio. The mean value is in fact sensitive to outlying observations. The classical way to deal with these situations is to use robust methods. As an example, the median is a more sensible location parameter than the empirical mean in these cases. Using robust methods makes it possible to avoid arbitrary decisions on the subset of data to analyse.

4.3 Independence of observations

4.3.1 Accounting for the non-independence of the observations within studies

Pittelkow et al. used bootstrapping techniques to construct confidence intervals around the estimated means. Possibly, the most fundamental assumption of bootstrap methods they used, is that all observations are independent[5] (Efron and Tibshirani, 1994). This assumption also underlies the randomization tests used in the *paper* to construct P-values, and thus our criticisms against the confidence interval calculation, holds for all the methods of inference presented in their paper. Put simply, Pittelkow et al.'s statistical tests were invalid.

It is legitimate to believe that observations stemming from the same study cannot be considered fully independent. Indeed, within study observations share very many features in terms of study design, location and authors, and they were also made public in the same publication (i.e. same journal, same review process). Pittelkow et al. when estimating the confidence intervals by bootstrap, did not account for the non-independence of the observations. This invalidates the confidence intervals they report in their paper.

The analysis in Section 5 shows clearly that observations of the same study are more similar than observations between studies[6]. This, implies that the confidence intervals obtained by Pittelkow et al. are not only incorrect but too narrow.

Pittelkow et al. included "study" in their analysis by giving all studies, assumed to have the same number of replicates, the same weight in the analysis. In particular, they stated in their *paper*: "In situations where more than one observation from a study was included in a category, weights were divided by the total number of observations from that study." Unfortunately, this action does not account for non-independence. It simply downweights observations from larger studies. This is counter intuitive and unfortunate. We don't see any reason why these observations are less valuable. According to the weighting used by Pittelkow et al., the larger the study is, the less important its observations it will be. As an example, assuming the same number of

[5] Extensions to this requirement exist, but were not employed by Pittelkow et al., and so not further discussed here.

[6] Indeed, the variance component of the study random effect is considerable compared to the residual variance.

replicates, an observation of the largest study, which yielded 144 observations, counts 144 times less than an observation stemming from the 58 studies that have just one single observation.

Forcing this equivalence is strange, and was left unjustified by Pittelkow et al. Indeed, it could be argued that rather than being downweighted, larger studies should actually count more than the small studies.

4.3.2 One comprehensive model instead of several disconnected models

Pittelkow et al. carried out several disconnected analyses. They analysed all data and some subgroups of data (e.g. to compare residue or rotation management, or to compare these practices between arid or humid environments). This approach is problematic because when computing the overall mean and its confidence intervals they assumed all observations to be independent. Nevertheless, when further analysing their data, they tested and found differences in different subcategories of the data. This in turn invalidated the assumption of independence made for the analysis on the overall mean.

One solution to this problem may be a single analysis where all questions are addressed simultaneously. This is the underlying idea of linear model with several predictors (i.e. a multiple linear regression model). This approach is superior to several disconnect models also because provides better estimates. Indeed, all effects are estimated simultaneously, and therefore take into account all other predictors present in the model.

In a publication following Pittelkow et al.'s paper, Lundy et al. attempted to account for multiple testing and conditionality issues by analysing Pittelkow et al.'s data set with linear mixed effects models (Lundy et al., 2015). However, Lundy et al. violated the assumptions underlying their methods, and our criticisms of Pittelkow et al.'s work were not addressed by them.

4.4 Statistical significance is not practical significance

Statistical significance does not always imply biological or practical significance (Cox, 1982; Wasserstein and Lazar, 2016). In particular the variation of the data must be taken into account when significance is interpreted. The most sensible way to interpret statistical significance of an effect of no-till on on crop yields is to visualize the observations along with the obtained estimates and the estimates expected under the null hypothesis of no yield difference between no-till and conventional till practices.

```
mean.log.ratio.all <- weighted.mean(d.nature$log.ratio.new,
                                    w = d.nature$weights.study.new)
update(dp.all, abline = list(v = c(mean.log.ratio.all, 0), col = c("red", "black")))
```

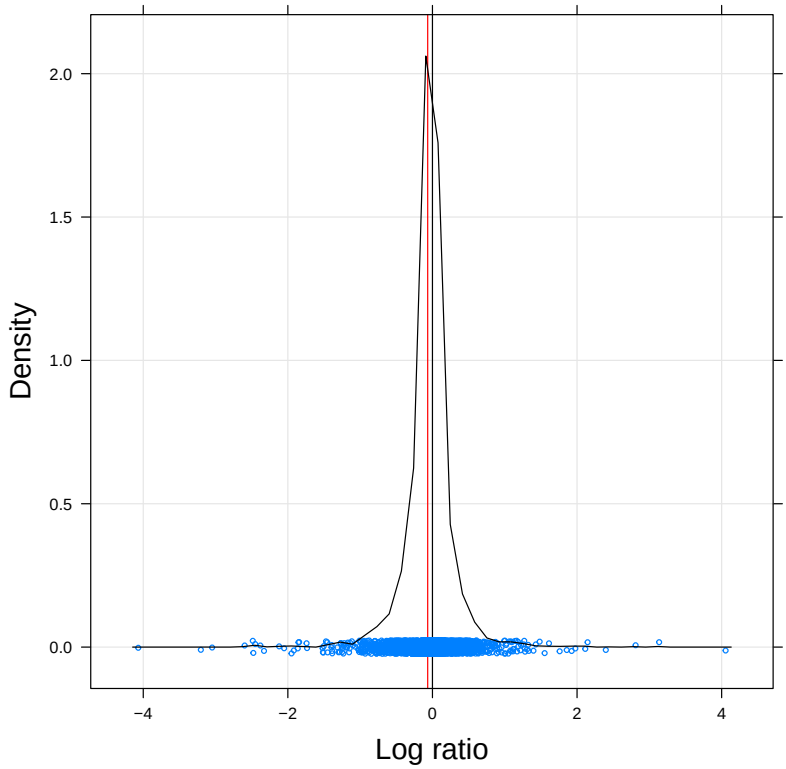


Figure 4: Density plot of the log ratio along with the observed mean and that expected under the null hypothesis of no difference between no-till and till (complete data set).

It is quite striking that the weighted mean of the data (red vertical line) can be hardly distinguished from the black vertical on zero (i.e. no difference in yield between the two practices)[7]. As mentioned in our response, the practical and biological significance of that difference has to be questioned. Even if the estimated difference in yields may be of a magnitude of practical relevance, we must note that the underlying variation in the data may mean this estimate poorly represents the range of experiments in the data set, and it of little relevance to the farming community. We address this issue formally in Section 5.

To get a better feeling of the range of effect size we can look at other summary statistics. We start by looking at the percentage of observations that show equal or higher yield in no-till practice.

```
prop.table(table(d.nature$log.ratio.new >= 0))


FALSE  TRUE
 0.57  0.43
```

Is interesting to see that actually 43% of the observations showed same or higher yield in no-till conditions. We can look at quantiles to inspect this further.

```
q.log.ratio <- quantile(d.nature$log.ratio.new, probs = c(0.025, 0.975))
100 * exp(q.log.ratio) - 100
```

[7]Note that in case of equality of yields the ratio would be one and thus its logarithm 0.

```
2.5%  98%
 -51   73
```

When looking at the 95% quantiles of the yield ratio, we see that the most of the data lies between −51% and +73%. So roughly speaking we go from half of the yield under no-till to a gain of two thirds under no-till! Indeed there is, without doubt, considerable and variation around the mean effect.

5 An alternative analysis

One possible way to take into account all of the issues we have raised above is to first display data appropriately and then use robust mixed effects models. We have already seen that the package `lattice` is a very powerful tool to inspect multivariate data, and how plotting the data is critical to understanding the practical relevance of point estimates and statistical significance tests.

The package `robustlmm` "provides functions for estimating linear mixed effects models in a robust way. The main workhorse is the function rlmer; it is implemented as direct robust analogue of the popular lmer function of the lme4 package." (Koller, 2016; Bates et al., 2015).

Using the tools of data visualization, and robust models, enable the analyst to:

1. Produce meaningful plots where the practical relevance of the effects can be interpreted.
2. Overcome biasedness: As robust methods estimate median values, which, as all other quantiles are invariant to monotonic transformations. So backtransforming, when needed, does not yield biased estimates.
3. Gain meaningful estimates for skewed variables.
4. Deal with long-tailedness: These are appropriately dealt with and no arbitrary subsetting of the data is needed.
5. Account for independence assumptions: "Study" can be included in the analysis as random effect.
6. Account for independence assumptions/ and gain better estimates: All subgroups can be analysed simultaneously.

In addition to that, the relevance of the estimated effects can be interpreted better, as it is possible to partition the variation explained between fixed effects and random effects (see page 17).

Here, we present a model that tries to stay as close to the analysis of Pittelkow et al., as much as possible, whilst accounting for criticisms we have raised above and in our *response*. We note that when using a linear model we don't need to discretise continuous variables such as "aridity" or "study duration". This is convenient because it avoids making arbitrary choices on the number of groups to use, and the break points used in categorization. Note, for our analysis we were not able to use the weights as these are internally estimated by the `rlmer()` function.

It is important to note that the alternative analysis presented here does not aim at being the ultimate analysis. Neither is this a general solution for meta-analyses. We present this analysis in order to show that:

- It is relatively easy to account for all our criticisms.
- Even when all predictors are considered simultaneously, the explanatory power of the models presented by Pittelkow et al. remain extremely low.
- Considering observations within studies independent is inappropriate for this data set.

5.1 Visualizing data

As standard practice, we begin our analysis by plotting the response variable against the predictors. Pittelkow et al. assumed that the effect of some predictors was not additive. As an example the effects of the two additional principles of conservation agriculture (rotation management and residue management) on crop yields under no-till are analysed separately in "dry" and "humid" environments. In a linear model this is achieved by letting the predictors "aridity" and "rotation and residue management" interact.

Here we use an extension to `lattice` package `latticeExtra` (Sarkar and Andrews, 2013). We display relationship between the response variable and "aridity" in separate panels to inspect whether this effect is influenced by the these two additional principles of conservation agriculture. This is consistent with the analysis by Pittelkow et al.

```
require(latticeExtra)
xy.ar <- xyplot(log.ratio.new ~ Aridity.Index | rotation.new * resid.man.new,
                data = d.nature,
                xlab = list(label = "Aridity index", cex = 1.5),
                ylab = list(label = "Log ratio", cex = 1.5),
                scales = "free",
                type = c("p","smooth","g"),
                col.line = "red", cex = 0.5)
useOuterStrips(xy.ar)
```

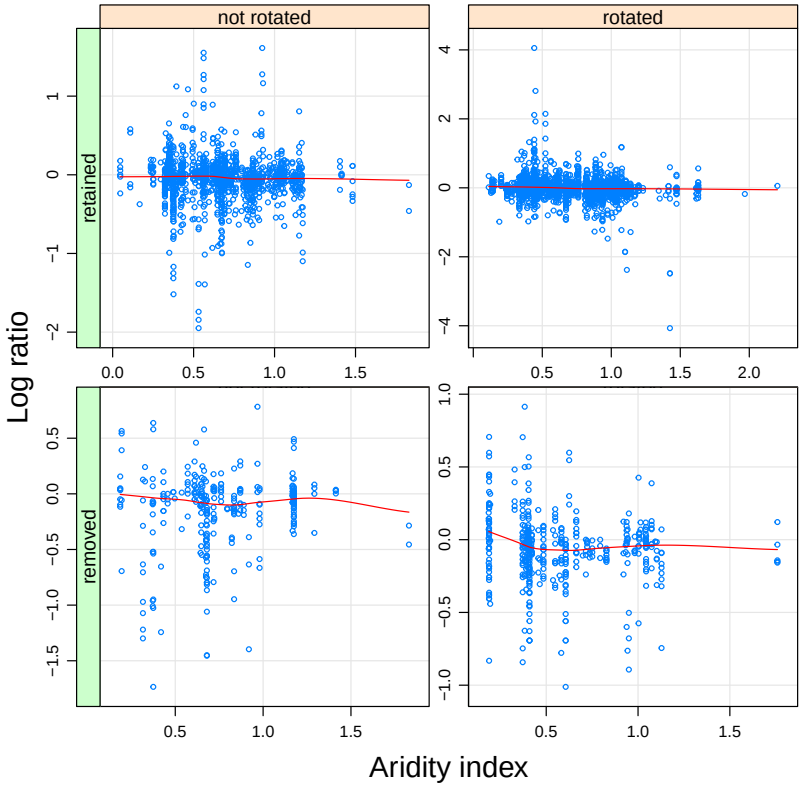


Figure 5: Logarithm of yield ratio plotted against aridity index (complete data set).

We don't see any striking patterns that need to be commented on here.

We now inspect the effect of "Study duration".

```
xy.dur <- xyplot(log.ratio.new ~ Study.Duration | rotation.new * resid.man.new,
                 data = d.nature,
                 xlab = list(label = "Study duration [years]", cex = 1.5),
                 ylab = list(label = "Log ratio", cex = 1.5),
                 scales = "free",
                 type = c("p", "smooth", "g"),
                 col.line = "red", cex = 0.5)
useOuterStrips(xy.dur)
```

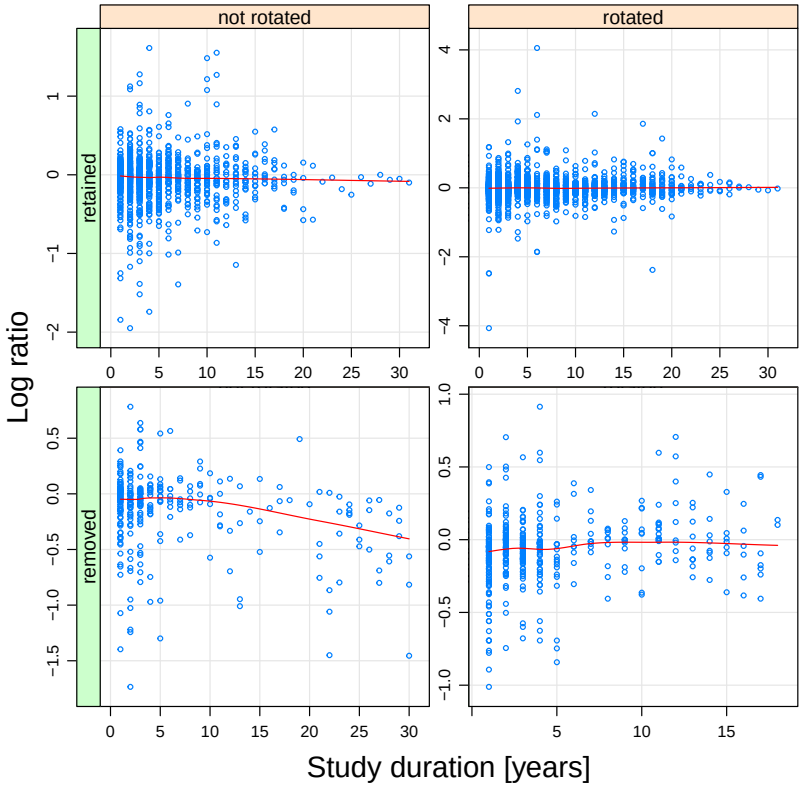


Figure 6: Logarithm of yield ratio plotted against study duration (complete data set).

The smoother visible in red enables us to decided whether assuming a linear effect is sensible. There is very little evidence that these variables have an effect at all, except for the lower left panel (i.e. not rotated and residue removed). We note here that for this specific combination of practices the study duration seems to have a negative effect on the response variable. There is no clear evidence that assuming linearity is inappropriate here.

5.2 Fitting a robust mixed effects model with `rlmer()`

As explained above we let some variables to interact. Pittelkow et al. assumed the effect "rotation" and "residue management" to be non additive. Indeed, their analysis looks at all four combinations of these treatments separately. In a linear model this corresponds to an interaction between the two factors. We create a new factor variable with these four categories.

```
d.nature$rot.res.new <- interaction(d.nature$rotation.new, d.nature$resid.man.new)
levels(d.nature$rot.res.new)
```

```
[1] "not rotated.removed"  "rotated.removed"      "not rotated.retained"
[4] "rotated.retained"
```

We now fit the robust mixed effects model. Note that this action is computationally intensive and takes several minutes.

```
require(robustlmm)
rob.mod <- rlmer(log.ratio.new ~ rot.res.new + Study.Duration + Aridity.Index +
                   rot.res.new:Study.Duration + rot.res.new:Aridity.Index +
                   (1 | study.new),
                 data = d.nature)
```

Our two main results obtained by fitting the robust mixed effect model are that the variance component for the random effect "study" is not negligible and that the predictive power of the model is extremely low. Both results are discussed below, in relation to our criticisms of the analysis of Pittelkow et al.

Independence of the observations in the same study

As stated above, we can use the variance components of the fitted model to understand whether observations of the same study can be considered to be independent or not. To do that, we use the `VarCorr()` method for robust mixed models. This method is based on the corresponding function in `lme4` (Bates et al., 2015).

```
require(lme4)
VarCorr(rob.mod)
```

```
 Groups    Name        Std.Dev.
 study.new (Intercept) 0.104
 Residual              0.169
```

The variance component of the random effect "study", which is estimated in a robust way, is considerable[8]. Based on these estimates we can estimate the intraclass correlation coefficient (i.e. the correlation between observations from the same study), to be as high as 0.27. Therefore, assuming independence of observations is inadequate.

Goodness of fit

It is debated whether the goodness of fit of a linear mixed effects model can be quantified with an extension of the classical R^2. We respect the opinion of both the authors of `lme4` and `robustlmm` packages and do not compute it. However, to get a feeling on how good our model is, we plotted the fitted values of the model against the observed values[9]. In Figure 7 we plot the relationship obtained for (1) a hypothetical perfect model (left panel); for (2) the robust model where the conditional modes of "study" are included in the prediction (second panel from left); for (3) the robust model without conditional modes (third from left); and finally for (4) a model that has no predictive power (right panel).

```
obs.y <- model.frame(rob.mod)[ ,"log.ratio.new"]
hat.y <- fitted(rob.mod)
hat.y.pop <- predict(rob.mod, re.form = ~ 0)
set.seed(2)
hat.y.random <- sample(hat.y.pop)
##
d.predictions <- make.groups(obs.y, hat.y, hat.y.pop, hat.y.random)
```

[8] This must be compared to the residuals standard deviation. Here there is roughly a 2:3 ratio, which implies that the correlation within study is important and not negligible.

[9] This can be seen as the graphical analogue of R^2. Indeed, this measure is the squared correlation coefficient obtained by regressing the observed values against the predicted ones.

```
d.predictions$obs.y <- obs.y
str(d.predictions)

'data.frame': 15012 obs. of  3 variables:
 $ data : num  0.0376 0.1916 0.1045 0.0864 0.124 ...
 $ which: Factor w/ 4 levels "obs.y","hat.y",..: 1 1 1 1 1 1 1 1 1 ...
 $ obs.y: num  0.0376 0.1916 0.1045 0.0864 0.124 ...

##
xy.fig2 <- xyplot(obs.y ~ data | which, data = d.predictions,
                  type = c("p", "r", "g"), col.line = "red",
                  xlab = "Observed log ratio", ylab = "Predicted log ratio",
                  strip = strip.custom(
                    factor.levels = c("Perfect fit",
                                      expression(paste(hat(y), " with study")),
                                      expression(paste(hat(y), " withOUT study")),
                                      "No relationship")),
                  scales = list(relation = "free", x = list(at = NULL),
                                y = list(at = NULL)),
                  asp = 1, cex = 0.5,
                  between = list(x = -1),
                  abline = 0, layout = c(4,1))
xy.fig2
```

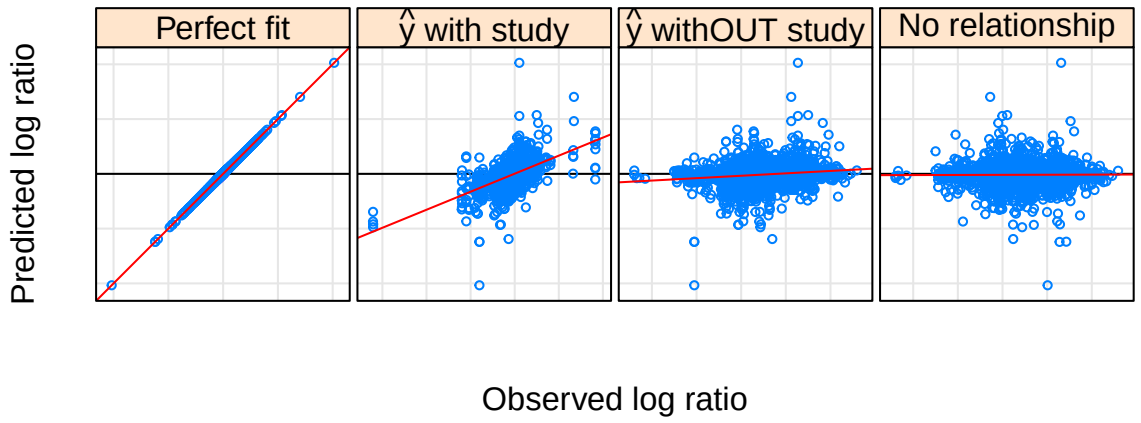


Figure 7: Goodness of fit evaluated graphically.

These graphs show that the when the explanatory power of "study" is taken into account (second panel from left), the model explains a fair proportion of the variance. However, when predictions are made without taking "study" into account (third panel) the predictive power drops to essentially nothing. This a debatable way to quantify the predictive power of a mixed model. Nevertheless, it is not questionable to claim that the fixed effects in this analysis (i.e. all predictors but "study", and the central models forwarded by Pittelkow et al. for explaining the effects of no-till on crop yields) are very weak at explaining the large variation of the data.

As mentioned above, the scope of displaying data with `lattice` plots and fitting the robust model was to demonstrate that the issues raised above can be dealt with, in a relatively simple manner. This is not the ultimate analysis, nor a general solution for meta-analyses. Several other options exist, as an example see the `robumeta` package (Fisher et al., 2016).

No inference is carried out here as we saw above that making inference with such a large sample size can be misleading. For those interested in carrying out formal statistical inference using the techniques we used here, the recommended way is to use bootstrapping techniques (Manuel Koller personal communication). Note this procedure is computationally very intensive and may require parallelization.

As a side note, when fitting a robust mixed effects model, only observations with information will be used. In other words, any observation that contains at least one missing value will be dropped.

```
nrow(d.nature)
```

```
[1] 5492
```

```
nrow(na.omit(d.nature[ ,c("Study.Duration", "Aridity.Index", "rot.res.new")]))
```

```
[1] 3753
```

Whilst this reduced the overall sample size in our analysis of the Pittelkow et al. data set, the sub sample of the data forced by this analyses highlights the number of 'information deficient' observations that did not stand up to the rigorous criteria of reporting/coding in primary studies that would be needed to unambiguously compare the effects of the suite of conservation agriculture practices (e.g. residue management vs. rotation), on crop yield under no-till. In other words the subsetting actually makes logical sense. As we are estimating as few as 12 parameters with more than 3700 observations, the power to detect any effects, if present, is still very large. Importantly, this subsetting does not affect our critique of the poor explanatory power of the moderators used by Pittelkow et al. or the non-independence resulting from high correlation between observations within studies.

6 There is always a "better" model

The reader may argue that by spending additional time there is always room for improvement and a better model can always be found. This is very true. As an example here we stuck faithfully to the analysis of Pittelkow et al. and did not consider all predictors available (e.g. crop type, geographic location), neither did we consider that observations may be correlated not just within study, but also in time and space.

Any statistical analysis will unavoidably involve decisions that are, to some extent, arbitrary (e.g. here we assumed linearity of the effects). As a result an analysis will always be somehow debatable. Nevertheless, the analysis presented by Pittelkow et al. presents four important faults that can hardly be ignored. These faults invalidate their conclusions and the core message of their paper.

7 Conclusions

General methodological conclusions of this study are:

- An appropriate visualization of the analysed data is fundamental. Deepayan Sarkar, the author of the `lattice` package, wrote an excellent companion book to introduce his package (Sarkar, 2008).
- Statistical significance does not imply biological or practical significance.
- Tools to make analyses fully reproducible and thus transparent exist. These require relatively little effort to be used.

General agricultural conclusions of this study are:

- Pittelkow's analysis was biased towards showing no-till had a negative effect on crop yields. Their analysis and 5.7 percent yield loss figure should not be used in decision making processes as evidence for farmers not to adopt no-till.
- The variation in no-till's effects on crop yields is so large that any broad brush average value is likely to be practically irrelevant to farmers.
- The explanatory power of conservation agriculture practices, such as rotation and residue management, for explaining variation in crop yields under no-till is no better than random. This completely invalidates the claims of Pittelkow et al. that smallholder farmers would need to use rotation and residue management to maximize yields under no-till.
- Study level differences, do appear to help explain the effects or differences in no-till practices. The reasons for these effects are unknown, and the explanatory power of other, study specific, covariates and moderators need to be explored, and taken into account before scientists using this meta-data set can make any practical recommendations to farmers based on the effect of no-till on crop yields.

Session information

```
sessionInfo()

R version 3.3.2 (2016-10-31)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Debian GNU/Linux 8 (jessie)

locale:
 [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
 [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
 [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
```

```
[1] robustlmm_1.8       lme4_1.1-10         Matrix_1.2-3
[4] latticeExtra_0.6-26 RColorBrewer_1.1-2  lattice_0.20-33
[7] checkpoint_0.3.16   knitr_1.14

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.2       magrittr_1.5      splines_3.3.2
 [4] MASS_7.3-45       munsell_0.4.2     xtable_1.8-0
 [7] colorspace_1.2-6  minqa_1.2.4       stringr_1.1.0
[10] highr_0.6         plyr_1.8.3        tools_3.3.2
[13] grid_3.3.2        nlme_3.1-122      gtable_0.1.2
[16] nloptr_1.0.4      ggplot2_2.0.0     formatR_1.4
[19] robustbase_0.92-5 evaluate_0.10     stringi_1.1.2
[22] DEoptimR_1.0-4    scales_0.3.0
```

References

D. Bates, M. Maechler, B. Bolker, and S. Walker. *lme4: Linear Mixed-Effects Models using 'Eigen' and S4*, 2015. URL `https://CRAN.R-project.org/package=lme4`. R package version 1.1-10.

M. Corporation. *checkpoint: Install Packages from Snapshots on the Checkpoint Server for Reproducibility*, 2016. URL `https://CRAN.R-project.org/package=checkpoint`. R package version 0.3.16.

D. Cox. Statistical significance tests. *British Journal of Clinical Pharmacology*, 14(3):325–331, 1982.

N. Duan. Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association*, 78(383):605–610, 1983.

B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Z. Fisher, E. Tipton, and Z. Hou. *robumeta: Robust Variance Meta-Regression*, 2016. URL `https://CRAN.R-project.org/package=robumeta`. R package version 1.8.

L. V. Hedges, J. Gurevitch, and P. S. Curtis. The meta-analysis of response ratios in experimental ecology. 1999.

J. L. W. V. Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30(1):175–193, 1906.

M. Koller. *robustlmm: Robust Linear Mixed Effects Models*, 2016. URL `https://CRAN.R-project.org/package=robustlmm`. R package version 1.8.

E. Limpert and W. A. Stahel. Problems with using the normal distribution–and ways to improve quality and efficiency of data analysis. *PLoS One*, 6(7):e21403, 2011.

E. Limpert, W. A. Stahel, and M. Abbt. Log-normal distributions across the sciences: Keys and clues on the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probabilitynormal or log-normal: That is the question. *BioScience*, 51(5):341–352, 2001.

M. E. Lundy, C. M. Pittelkow, B. A. Linquist, X. Liang, K. J. Van Groenigen, J. Lee, J. Six, R. T. Venterea, and C. Van Kessel. Nitrogen fertilization reduces yield declines following no-till adoption. *Field Crops Research*, 183:204–210, 2015.

F. Mosteller and J. W. Tukey. Data analysis and regression: a second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*, 1977.

J. Neyman and E. L. Scott. Correction for bias introduced by a transformation of variables. *The Annals of Mathematical Statistics*, 31(3):643–655, 1960.

C. M. Pittelkow, X. Liang, B. A. Linquist, K. J. Van Groenigen, J. Lee, M. E. Lundy, N. van Gestel, J. Six, R. T. Venterea, and C. van Kessel. Productivity limits and potentials of the principles of conservation agriculture. *Nature*, 517(7534):365–368, 2015.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `https://www.R-project.org/`.

D. Sarkar. *Lattice: multivariate data visualization with R*. Springer Science & Business Media, 2008.

D. Sarkar. *lattice: Trellis Graphics for R*, 2015. URL `https://CRAN.R-project.org/package=lattice`. R package version 0.20-33.

D. Sarkar and F. Andrews. *latticeExtra: Extra Graphical Utilities Based on Lattice*, 2013. URL `https://CRAN.R-project.org/package=latticeExtra`. R package version 0.6-26.

R. L. Wasserstein and N. A. Lazar. The asa's statement on p-values: context, process, and purpose. *The American Statistician*, (just-accepted):00–00, 2016.

Y. Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2016. URL `https://CRAN.R-project.org/package=knitr`. R package version 1.14.

Appendix B: Data preparation

Chapter 4

Contents

1 Aim of this document

The data set accompanying the paper of Pittelkow et al. (Pittelkow et al., 2015) is not directly usable to reproduce the analysis presented in their publication. Several important variables are missing from the data set. Below, we carry out all steps required to reproduce their results and perform the analysis given in the Supplementary Information A document. We take this opportunity to demonstrate how easy it is to make statistical analysis fully reproducible in **R**.

2 Making everything fully reproducible

We make use of the package `checkpoint` (Corporation, 2016) to make all of our analysis fully reproducible. Note, that it is very unlikely that updates in the packages since the publication of Pittelkow et al. would lead to any change in the data set preparation. Our use of `checkpoint` allows any future researcher to directly reproduce our results.

```
require(checkpoint)
checkpoint(snapshotDate = "2016-01-01")
```

Note that the **R** version used here is 3.3.2 (2016-10-31). Using other versions of **R** should not have any influence on the data set obtained here. The checkpoint() function called above, installs all packages used in this document as they were available on the 1st of January 2016 in your 'home' directory.

This dynamic document was created with package `knitr` (Xie, 2016).

3 Data preparation

The data set built by Pittelkow et al. is freely available at `http://www.nature.com/nature/journal/v517/n7534/full/nature13809.html`. We downloaded it on March 10 2016. The first two heading rows of this document (empty lines that do not contain any data), were omitted so that it can be imported into **R**. This file was then saved with the 'Microsoft Excel 2007/2010/2013 XML (.xlsx)' file extension. Because special characters and quotation marks are present in the original data file (e.g. author "Angás"), we used the `read.xls()` function found in package `gdata` (Warnes et al., 2015) to import data into **R**[1].

```
require(gdata)
d.nature <- read.xls("nature13809-s1.xlsx")
```

```
str(d.nature)
```

```
'data.frame': 5513 obs. of  17 variables:
 $ Author                  : Factor w/ 515 levels "Aase  et al.",..: 1 1 1 1 2 2 2 2 2 2 ...
 $ Year                    : Factor w/ 44 levels "1981","1983",..: 20 20 20 20 29 29 29 29 29 29 ..
```

[1]Note that importing '.xlsx' files into **R** may lead to difficulties across different operating systems. Other file extensions are more suited to make data publicly available.

```
 $ Journal                  : Factor w/ 132 levels "Acta Agriculturae Scandinavica Section B",..:
 $ Country                  : Factor w/ 62 levels "Algeria","Argentina",..: 60 60 60 60 33 33 33 3
 $ Location                 : Factor w/ 534 levels "Ihinger Hof, University of Hohenheim",..: 293
 $ Latitude                 : num  48.3 48.3 48.3 48.3 32.5 ...
 $ Longitude                : num  -105 -105 -105 -105 36 ...
 $ Reps...conv..till.       : int  3 3 3 3 2 2 2 2 2 2 ...
 $ Reps...no.till.          : int  3 3 3 3 2 2 2 2 2 2 ...
 $ Crop                     : Factor w/ 48 levels "alfalfa","barley",..: 47 47 47 47 3 3 3 2 2 2 .
 $ Study.Duration           : int  1 2 3 4 1 1 1 1 1 1 ...
 $ Irrigation.              : Factor w/ 3 levels "","no","yes": 2 2 2 2 3 3 3 3 3 3 ...
 $ Rotation.                : Factor w/ 3 levels "cover","no","yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ Residue.Mgmt             : Factor w/ 4 levels "","burned","removed",..: 1 1 1 1 4 4 4 4 4 4 ...
 $ Aridity.Index            : num  0.334 0.334 0.334 0.334 0.236 0.236 0.236 0.236 0.236 0.236 ...
 $ Yield...conv..till..kg.ha: int  2755 2567 2405 2439 678 696 999 5854 6754 9665 ...
 $ Yield...no.till..kg.ha   : int  2782 2694 2439 2788 704 843 1109 6382 7646 10531 ...
```

```
head(d.nature, n = 3)
```

```
        Author Year                  Journal Country       Location
1 Aase  et al. 1997 Soil & Tillage Research      USA Montana, Froid
2 Aase  et al. 1997 Soil & Tillage Research      USA Montana, Froid
3 Aase  et al. 1997 Soil & Tillage Research      USA Montana, Froid
  Latitude Longitude Reps...conv..till. Reps...no.till.  Crop
1       48      -105                  3               3 wheat
2       48      -105                  3               3 wheat
3       48      -105                  3               3 wheat
  Study.Duration Irrigation. Rotation. Residue.Mgmt Aridity.Index
1              1          no        no                      0.33
2              2          no        no                      0.33
3              3          no        no                      0.33
  Yield...conv..till..kg.ha Yield...no.till..kg.ha
1                      2755                   2782
2                      2567                   2694
3                      2405                   2439
```

The data set provided contains 17 variables and 5513 observations. Unfortunately, not all variables used in their analysis are provided, which prevents immediate reproduction of the results in their paper. As an example, the "yield ratio" response variable and the "study" variable are missing. Thus, the data set needs to be extended to include these missing variables. All new variables created by us in this document will have the suffix `.new`.

Due to the fact that the original data set provided by Pittelkow et al. uses a naming that is not convenient for **R**, the naming of the columns may appear somewhat odd. However we do not modify the original column names here to keep things simple.

3.1 Creating the "yield ratio" response variable

We first compute the ratio between the two paired yield observations of no-till and conventional till. This variable is called "yield ratio".

```
d.nature$yield.ratio.new <- d.nature$Yield...no.till..kg.ha /
  d.nature$Yield...conv..till..kg.ha
summary(d.nature$yield.ratio.new)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0     0.9     1.0     Inf     1.1     Inf      19
```

The newly created variable contains "infinite" observations and missing values. We will address this problem later in section 3.6.

We now compute the natural logarithm of the yield ratio. This variable is called "log ratio".

```
d.nature$log.ratio.new <- log(d.nature$yield.ratio.new)
summary(d.nature$log.ratio.new)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   -Inf    -0.1     0.0     NaN     0.1     Inf      19
```

Note that positive values of "log.ratio" indicate an higher yield under no-till, while negative values indicate that yield is lower in no-till relative to tilled systems. There are a couple of `Inf/-Inf` and `NaN` values. This likely results from the fact that we have been dividing zero by zero, or taking the log of zero. Let's check that.

```
d.nature[!is.finite(d.nature$log.ratio.new),
         c("Yield...conv..till..kg.ha", "Yield...no.till..kg.ha", "log.ratio.new")]

     Yield...conv..till..kg.ha Yield...no.till..kg.ha log.ratio.new
2214                         0                     0           NaN
2216                         0                   480           Inf
2224                         0                     0           NaN
2479                     10600                     0          -Inf
4014                         0                     0           NaN
4029                         0                     0           NaN
4035                         0                     0           NaN
4037                         0                     0           NaN
4039                         0                     0           NaN
4040                         0                     0           NaN
4050                         0                     0           NaN
4065                         0                     0           NaN
4071                         0                     0           NaN
4073                         0                     0           NaN
4075                         0                     0           NaN
4076                         0                     0           NaN
4104                         0                     0           NaN
4119                         0                     0           NaN
4127                         0                     0           NaN
4129                         0                     0           NaN
4130                         0                     0           NaN
```

Indeed, these observations will be dropped from the analysis, consistent with the procedure of Pittelkow et al.

3.2 Creating the new "study identifier" variable

The data set does not provide a study level variable. Pittelkow et al. do not indicate how to create this predictor in their paper. Therefore, we try to create it by intersecting "author", "year" and "journal". This assumes that an author has not published more than one paper in the same journal within one year.

```
d.nature$study.new <- with(d.nature,
                           interaction(Author, Year, Journal, drop = TRUE))
nlevels(d.nature$study.new)
```

```
[1] 608
```

We have 608 different studies. Pittelkow report having 610 studies. To explore this mismatch further, we considered "country", "study duration" and "location", but we were not able to find 610 studies in this dataset. However, this difference in study number (608 vs. 610) is extremely unlikely to have any influence our results (see Supplementary Information A for the reproduction of Pittelkow et al's findings using 608 studies).

3.3 Creating the new "weights" variables

We create a vector with the weights that takes the number of replicates into account.

```
d.nature$weights.new <- (d.nature$Reps...conv..till. * d.nature$Reps...no.till.) /
  (d.nature$Reps...conv..till. + d.nature$Reps...no.till.)
```

We check how many observations each study has.

```
table(table(d.nature$study.new), dnn = "No. of observations")
```

```
No. of observations
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
 58  82  69  71  36  60  12  39  15  18   4  26   7  13  10   8   6   7
 19  20  21  22  23  24  25  28  30  32  34  36  40  42  46  47  50  55
  4   3   2   5   1  16   2   5   4   1   1   6   1   1   1   1   1   1
 56  64  72  74  80  81  84 144
  1   4   1   1   1   1   1   1
```

There are 58 studies with just one observation. On the other extreme, we have one single study with 144 observations.

We now create a new vector of weights that takes study size into account too.

```
d.nobs.study <- as.data.frame(table(d.nature$study.new))
names(d.nobs.study)
```

```
[1] "Var1" "Freq"
```

```
names(d.nobs.study) <- c("study.new", "no.obs.new")
head(d.nobs.study, n = 3)
```

```
                                                                    study.new
1                        Zhang et al..2007.Acta Agriculturae Scandinavica Section B
2                Lotjonen and Isolahti.2010.Acta Agriculturae Scandinavica Section B
3 Taser and Metingoglu.2005.Acta Agriculturae Scandinavica Section B-Soil and Plant
  no.obs.new
1          2
2         12
```

```
3          1
```

```
d.nature <- merge(x = d.nature, y = d.nobs.study, by = "study.new")
d.nature$weights.study.new <- d.nature$weights.new / d.nature$no.obs.new
```

3.4 Creating the new "rotation" variable with two levels

In their paper Pittelkow et al. use "rotation" as a categorical variable with two levels. However, in the dataset provided with their paper, three levels are specified. We therefore merge two of them. To do that we will use the `recode()` function found in package `car` (Fox and Weisberg, 2016).

```
require(car)
levels(d.nature$Rotation.)
```

```
[1] "cover" "no"    "yes"
```

```
d.nature$rotation.new <- recode(d.nature$Rotation.,
                                "c('cover','yes') = 'rotated'; else = 'not rotated'")
```

We now check that everything worked well.

```
levels(d.nature$rotation.new)
```

```
[1] "not rotated" "rotated"
```

```
table(d.nature[ ,c("Rotation.", "rotation.new")],
      useNA = "ifany", dnn = c("old", "new"))
```

```
       new
old     not rotated rotated
  cover           0     214
  no           2074       0
  yes             0    3225
```

3.5 Creating the new "residue management" variable with two levels

As for the variable "rotation", we need to merge some categories to obtain the levels used by Pittelkow et al. in their analysis. We keep "removed" and "retained" as stated by Pittelkow et al. We assume that empty entries mean that the residue management treatment is unknown. We set these entries to `NA`.

```
levels(d.nature$Residue.Mgmt)
```

```
[1] ""         "burned"   "removed"  "retained"
```

```
d.nature$resid.man.new <- recode(d.nature$Residue.Mgmt,
                                 "c('burned', 'removed') = 'removed';
                                 '' = NA")
```

Again we check that everything worked properly.

```
levels(d.nature$resid.man.new)
```

```
[1] "removed"  "retained"
```

```
table(d.nature[ ,c("Residue.Mgmt", "resid.man.new")],
      useNA = "ifany", dnn = c("old", "new"))
```

```
          new
old        removed retained <NA>
                 0        0 1462
  burned       122        0    0
  removed      607        0    0
  retained       0     3322    0
```

3.6 Removing the "non-finite" observations

```
nrow(d.nature)
```

```
[1] 5513
```

```
d.nature <- d.nature[is.finite(d.nature$log.ratio.new), ]
nrow(d.nature)
```

```
[1] 5492
```

The sample size does not change drastically after removing non-finite observations.

3.7 Removing the "extreme" observations

The authors arbitrarily removed observations more than 5 standard deviation away from the weighted mean of the log ratio in their analysis. We re-create a response variable with "extreme" observations removed, so that we can reproduce the analysis of Pittelkow et al. in the case where the overall mean is estimated.

```
w.mean.all <- weighted.mean(d.nature$log.ratio.new, w = d.nature$weights.study.new)
sd.es <- sd(d.nature$log.ratio.new)
d.nature$outlier.new <- abs(d.nature$log.ratio.new - w.mean.all) > (sd.es * 5)
prop.table(table(d.nature$outlier.new)) * 100
```

```
FALSE  TRUE
99.47  0.53
```

```
##
d.subset <- subset(d.nature, subset = !d.nature$outlier.new)
```

In the case, where "extreme" outliers are removed, about 0.5% of the observations are dropped.

4 Checking everything again and store data

```
str(d.subset)
```

```
'data.frame': 5463 obs. of  26 variables:
 $ study.new                : Factor w/ 608 levels "Zhang et al..2007.Acta Agriculturae Scandinavica
 $ Author                   : Factor w/ 515 levels "Aase  et al.",..: 1 1 1 1 2 2 2 2 2 2 ...
 $ Year                     : Factor w/ 44 levels "1981","1983",..: 20 20 20 20 29 29 29 29 29 29 ..
 $ Journal                  : Factor w/ 132 levels "Acta Agriculturae Scandinavica Section B",..: 114
 $ Country                  : Factor w/ 62 levels "Algeria","Argentina",..: 60 60 60 60 33 33 33 33 3
 $ Location                 : Factor w/ 534 levels "Ihinger Hof, University of Hohenheim",..: 293 293
 $ Latitude                 : num  48.3 48.3 48.3 48.3 32.5 ...
 $ Longitude                : num  -105 -105 -105 -105 36 ...
 $ Reps...conv..till.       : int  3 3 3 3 2 2 2 2 2 2 ...
 $ Reps...no.till.          : int  3 3 3 3 2 2 2 2 2 2 ...
 $ Crop                     : Factor w/ 48 levels "alfalfa","barley",..: 47 47 47 47 3 3 3 2 2 2 ...
 $ Study.Duration           : int  1 2 3 4 1 1 1 1 1 1 ...
 $ Irrigation.              : Factor w/ 3 levels "","no","yes": 2 2 2 2 3 3 3 3 3 3 ...
 $ Rotation.                : Factor w/ 3 levels "cover","no","yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ Residue.Mgmt             : Factor w/ 4 levels "","burned","removed",..: 1 1 1 1 4 4 4 4 4 4 ...
 $ Aridity.Index            : num  0.334 0.334 0.334 0.334 0.236 0.236 0.236 0.236 0.236 0.236 ...
 $ Yield...conv..till..kg.ha: int  2755 2567 2405 2439 678 696 999 5854 6754 9665 ...
 $ Yield...no.till..kg.ha   : int  2782 2694 2439 2788 704 843 1109 6382 7646 10531 ...
 $ yield.ratio.new          : num  1.01 1.05 1.01 1.14 1.04 ...
 $ log.ratio.new            : num  0.00975 0.04829 0.01404 0.13374 0.03763 ...
 $ weights.new              : num  1.5 1.5 1.5 1.5 1 1 1 1 1 1 ...
 $ no.obs.new               : int  4 4 4 4 6 6 6 6 6 6 ...
 $ weights.study.new        : num  0.375 0.375 0.375 0.375 0.167 ...
 $ rotation.new             : Factor w/ 2 levels "not rotated",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ resid.man.new            : Factor w/ 2 levels "removed","retained": NA NA NA NA 2 2 2 2 2 2 ...
 $ outlier.new              : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

The number of observations matches with that reported by Pittelkow et al. in their the article. We have also been able to recreate all the variables needed except for "study". However, the difference is in the "study" variable is minimal (i.e. we had 608 different studies while Pittelkow et al. report 610 studies). This difference is extremely unlikely to have any influence on the results obtained.

We store this data is a format that is very easy to handle across multiple operating systems (in contrast to Excel files which can cause compatibility problems for sharing data).

```
saveRDS(object = d.nature, file = "DataNatureReady.rds")
```

The data is saved as a `.rds` file that can be read on any computer regardless of their operating system. This feature of cross-compatibility is also present in other data formats such as text files (i.e. `.txt`). However, when reading rds files no action needs to be taken to modify the data so that is ready for the analysis. As an example, when a text file is read with the `read.table()` function, we still need to declare the class of columns (e.g. factors), this is not needed with `.rds` format.

Session Information

```
sessionInfo()
```

```
R version 3.3.2 (2016-10-31)
```

```
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Debian GNU/Linux 8 (jessie)

locale:
 [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
 [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
 [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] car_2.1-3         gdata_2.17.0      checkpoint_0.3.16 knitr_1.14

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.2        magrittr_1.5       splines_3.3.2
 [4] MASS_7.3-45        lattice_0.20-33    minqa_1.2.4
 [7] stringr_1.1.0      highr_0.6          tools_3.3.2
[10] nnet_7.3-12        pbkrtest_0.4-6     parallel_3.3.2
[13] grid_3.3.2         nlme_3.1-122       mgcv_1.8-16
[16] quantreg_5.29      MatrixModels_0.4-1 gtools_3.5.0
[19] lme4_1.1-10        Matrix_1.2-3       nloptr_1.0.4
[22] formatR_1.4        evaluate_0.10      stringi_1.1.2
[25] SparseM_1.74
```

References

M. Corporation. *checkpoint: Install Packages from Snapshots on the Checkpoint Server for Reproducibility*, 2016. URL `https://CRAN.R-project.org/package=checkpoint`. R package version 0.3.16.

J. Fox and S. Weisberg. *car: Companion to Applied Regression*, 2016. URL `https://CRAN.R-project.org/package=car`. R package version 2.1-3.

C. M. Pittelkow, X. Liang, B. A. Linquist, K. J. Van Groenigen, J. Lee, M. E. Lundy, N. van Gestel, J. Six, R. T. Venterea, and C. van Kessel. Productivity limits and potentials of the principles of conservation agriculture. *Nature*, 517(7534):365–368, 2015.

G. R. Warnes, B. Bolker, G. Gorjanc, G. Grothendieck, A. Korosec, T. Lumley, D. MacQueen, A. Magnusson, J. Rogers, and others. *gdata: Various R Programming Tools for Data Manipulation*, 2015. URL `https://CRAN.R-project.org/package=gdata`. R package version 2.17.0.

Y. Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2016. URL `https://CRAN.R-project.org/package=knitr`. R package version 1.14.

Chapter 5

Forest diversity promotes individual tree growth in central European forest stands

5.1 Preface

Authors: Juliette Chamagne[1,2], Matteo Tanadini[3], David Frank[4], Radim Matula[5], Timothy CE Paine[6], Christopher Philipson[7], Martin Svátek[5], Lindsay A Turnbull[3], Daniel Volařík[5] & Andy Hector[3]
1. Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
2. Forest Management and Development (ForDev) Group, Department of Environmental System Sciences, Swiss Federal Institute of Technology, Zurich, Universitätstrasse 16, 8092 Zurich, Switzerland
3. Department of Plant Sciences, University of Oxford, OX1 3RB Oxford, UK
4. Swiss Federal Research Institute WSL, Zürcherstrasse 111, 8903 Birmensdorf, Switzerland
5. Department of Forest Botany, Dendrology and Geobiocoenology, Faculty of Forestry and Wood Technology, Mendel University, Zemědělská 3, 613 00 Brno, Czech Republic
6. Biological and Environmental Sciences, University of Stirling, FK9 4LA Stirling, UK
7. Ecosystem Management Group, Department of Environmental System Sciences, Swiss Federal Institute of Technology, Zurich, Universitätstrasse 16, 8092 Zurich, Switzerland

Author contributions (as stated on the paper): JC, AH, RM, CETP, MS and LT conceived and designed the research; JC, AH, RM, CP, MS and DV conducted the research; JC, MT, DF, AH, CDP, CETP and LAT designed and conducted the analyses and interpretation; JC, AH and MT led the manuscript writing with substantial contributions from DF, CDP, CETP and LAT and input from all authors.

Detailed personal contributions: I carried out the data preparation required for the analysis. I designed and performed the statistical analysis using the method developed in Chapter 2, 3 and 4 of my thesis. Myself and Dr Juliette Chamagne interpreted

the results with Prof Andrew Hector (thesis supervisor). Dr Juliette Chamagne, myself and Prof Andrew Hector wrote the paper. I actively contributed to the revision of the paper in the reviewing phase. I wrote the dynamic document that describes the analysis (Appendix of the paper found in the electronic supporting material).

Target Readership: Ecologists and biologists. Especially researchers working in the field of biodiverstiy ecosystem functioning experiments.

Context: This chapter shows an application of the method presented in the early chapters of this thesis. In particular, the new method was used to quantify spatial asynchrony, and partial residual plots were used to judge the biological relevance of the predictors used in the modelling phase.

Article status: This chapter has been published in the *Journal of Ecology* (Chamagne et al., 2016).

Copyright: This chapter was omitted from the Bodleian version of the thesis as the publisher does not allow for the reproduction of the published paper.

5.2 Appendix

The Appendix of this chapter is to be found online (http://onlinelibrary.wiley.com/doi/10.1111/1365-2664.12783/full).

Chapter 6

The value of biodiversity for the functioning of tropical forests: Insurance effects during the first decade of the Sabah Biodiversity Experiment

6.1 Preface

Authors: Sean L Tuck[1], Michael J O'Brien[2,3], Christopher D Philipson[4], Philippe Saner[5], Matteo Tanadini[1], Dzaeman Dzulkifli[6], H Charles J Godfray[7], Elia Godoong[8], Reuben Nilus[9], Robert C Ong[9], Bernhard Schmid[5], Waidi Sinun[10], Jake L Snaddon[11], Martijn Snoep[12], Hamzah Tangki[10], John Tay[13], Philip Ulok[3], Yap Sau Wai[10], Maja Weilenmann[5], Glen Reynolds[3] & Andy Hector[1]

1 Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK

2. Consejo Superior de Investigaciones Científicas, Estacíon Experimental des Zonas Aridas, Carretera de Sacramento s/n, 04120 La Cañada, Almería, Spain

3. Danum Valley Field Centre, The SE Asia Rainforest Research Partnership (SEARRP), PO Box 60282, 91112 Lahad Datu, Sabah, Malaysia

4. Ecosystem Management, Institute of Terrestrial Ecosystems, ETH Zurich, Switzerland

5. Department of Evolutionary Biology and Environmental Studies, University of Zurich, 8057 Zurich, Switzerland

6. Tropical Rainforest Conservation and Research Centre, Lot 2900 and 2901, Jalan 7/71B Pinggiran Taman Tun, 60000 Kuala Lumpur, Malaysia

7. Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

8. Institute for Tropical Biology and Conservation, Universiti Malaysia Sabah, 88400 Sabah, Kota Kinabalu, Malaysia

9. Sabah Forestry Department Forest Research Centre, Mile 14 Jalan Sepilok, 90000 Sandakan, Sabah, Malaysia

10. Yayasan Sabah (Conservation and Environmental Management Division), 12th Floor, Menara Tun Mustapha, Yayasan Sabah, Likas Bay, PO Box 11622, 88813 Kota

Kinabalu, Sabah
11. Centre for Biological Sciences, University of Southampton, Southampton, UK
12. Face the Future, Utrechtseweg 95, 3702 AA, Zeist, The Netherlands
13. School of International Tropical Forestry, Universiti Malaysia Sabah, Kota Kinabalu, 88400 Sabah, Malaysia

Author contributions (as stated on the paper): Sabah Biodiversity Experiment conceived and designed by H.C.J.G., A.H. and G.R.; overall coordination by A.H. and G.R. with field coordination by M.J.O., C.D.P., P.S. and J.L.S. with advice and support from E.G., R.N., R.C.O., H.T., J.T., M.S., W.S., Y.S.W. and J.L.S.; data protocols, collection and processing by D.D., A.H., M.J.O., C.D.P., J.L.S., P.S., B.S., S.L.T., P.U., M.W.; data analysed by S.L.T., M.T. and A.H. with input from M.J.O. and C.D.P.; paper writing led by S.L.T. and A.H. with M.J.O., C.D.P. and M.T. and input from all authors.

Detailed personal contributions: I prepared the data required for the analysis (data aggregation described in the paper), while Dr Sean Tuck created the data set used in the analysis. I designed the analysis of both response variables analysed in the publication. I analysed the response variable "growth" and guided the analysis of the other response variable analysed, "survival". This latter analysis was carried out by Dr Sean Tuck. Myself and Dr Sean Tuck interpreted the results with input from Prof Andrew Hector. I wrote those paper sections concerning the statistical analysis of the response variable growth. I wrote a dynamic document regarding the analysis of growth that was then slightly modified by Dr Sean Tuck to comply with journal requirements.

Target Readership: Ecologists and biologists. Especially researchers working in the field of biodiversity ecosystem functioning experiments.

Context: This chapter shows an application of the method presented in the early chapters of this thesis. In particular, the new method was used to quantify spatial asynchrony and stability.

Article status: This chapter has been published in the *Proceedings of the Royal Society B: Biological Sciences* (Tuck et al., 2016).

Copyright: This chapter was omitted from the Bodleian version of the thesis as the publisher does not allow for the reproduction of the published paper.

6.2 Appendix

The Appendix and the supplementary tables of this chapter is to be found online (http://dx.doi.org/10.5061/dryad.1tc43).

Chapter 7

General Discussion

7.1 Summary of the current situation

In the last decades our understanding of ecosystem functioning and how biodiversity modulates it has greatly improved. This has yielded new and more specific research questions. To answer these questions and to be more relevant to real ecosystems, diversity-function experiments were ran in more natural and complex settings. The more realistic conditions brought in new complexities, but also the opportunity to gather new interesting biological information. In particular, spatial heterogeneity and temporal fluctuations allowed researchers to further explore the consequences of spatial and temporal asynchrony for ecosystem stability.

At the same time, the tools used to analyse these experiments followed two distinct paths. The primary analysis is mostly carried out with a modern mainstream statistical method: Linear Mixed Effects Models. This allows the user to incorporate the complexity of the new generation of experiments. On the other hand, secondary analyses are performed with field-specific metrics that did not entirely account for the fact that these experiments are nowadays run into more realistic settings. In addition, there is no formal connection among these secondary analyses nor with the primary analysis. As a consequence, the interpretation and communication of the results of a biodiversity-ecosystem functioning experiment can be difficult. In particular, it can be hard to interpret their biological relevance.

7.2 Summary of my research

In my doctorate I developed a new and comprehensive approach to analyse adequately modern diversity-function experiments. In particular, I merged the the two parts of the analysis of these experiments, by using the model-based approach of the primary analysis and the specificity of the metrics currently-used in the secondary analyses. This allows us to simultaneously address all relevant questions within a single framework, while taking into account the complexity of modern diversity-function experiments.

Put simply, I proposed to extend the primary Linear Mixed Effect Model so that many other relevant questions can be addressed at the same time. For spatial and temporal asynchrony this involves adding an interaction term between the space or time component and the grouping factor of interest (e.g. species identity). To quantify stability I proposed to graphically inspect the mean-variance relationship as seen on the diagnostics plots. For each group of interest a separate mean-variance relationship is estimated. This approach can be used at different organisational levels (e.g. population or community). I also proposed to use post-hoc contrasts to compare ecosystem functioning (e.g. to quantify overyielding). Finally, I showed how to interlink the results of different response variables, measured on the same experiment, to answer new biological questions.

The approach presented in this thesis has been successfully applied to several data sets. Some of the results have been used to illustrate how to use the method in the extensive appendices associated with the methodological publications (Chapters 2 and 3). Other results have been published as case-study publications (Chapters 5 and 6). I also highlighted the importance of a correct biological interpretation of the results stemming from ecological studies that have real-world impacts (Chapter 4).

7.3 Advantages of the proposed approach

The advantages of my comprehensive approach over a disconnected analysis are manifold. The model-based framework enables us to estimate all processes of interest

simultaneously. This implies that we can estimate, for example, stability at several organisational levels (e.g. population and community) while disentangling the contributions of each. Furthermore, this conditional approach enables us to estimate, for example, temporal asynchrony when the biasing effect of tree growth as well as the confounding effect of tree size are controlled for.

A concurrent estimation of all processes of interest, rather then examining them in isolation, is also essential to understand their biological relevance. As a matter of fact, a process may be present, but so weak compared to the other effects, to be considered irrelevant. My approach makes it possible to compare the results of different ecological aspects (e.g. overyielding and ecosystem stability). All questions are addressed in such a way that the answer is quantified on the same scale as the response variable and can easily displayed graphically.

The tailored use of a mainstream statistical technique allows us to compute p-values for each hypothesis as well as estimating confidence intervals for the values that are estimated by the analysis. Interestingly, statistical inference can be carried out with essentially any kind of data (continuous, count, binary etc.). This is feasible through the use of the extensions of Linear Mixed Effects Models, such as Generalised Linear Mixed Effects Models or by using bootstrapping techniques. To get a complete picture of ecosystem functioning, it is fundamental to analyse a wider variety of response variables. Virtually all currently-used metrics can only deal with continuous traits such as diameter or biomass.

This modern approach drops many of the design requirements made by secondary metrics. For example, to compute temporal asynchrony, species do not need to be measured at the same discrete time points. Generalised Additive Models further relax the design assumptions implicitly made by secondary metrics, such as a known scale at which spatial asynchrony works.

In general, the new approach represents an advancement of the field as it makes it possible to biologically interpret, compare and communicate the results of modern diversity-function experiments.

7.4 Implications of my research

Implications for past research

In Chapters 2 and 3 I showed that some of the implicit assumptions made by secondary metrics are not met in present-day diversity-function studies and that this can have important consequences. In particular:

- Species asynchrony may often be underestimated. This is due to, for example, the presence of trends (e.g. growth) or human-induced fluctuations (e.g. harvesting scheme) that can obscure asynchrony. As a consequence, the stabilising effect of species diversity may be underestimated.

- The scale at which spatial asynchrony is measured, is currently defined by the design of the experiment. The actual scale a which this process works may be different from the assumed one. By looking at the wrong spatial scale we may incorrectly assume that asynchrony is absent. Both, positive and negative spatial asynchrony are likely to be underestimated.

- Many asynchrony metrics condense the whole biological information of the ecological process under study into a single value. Unfortunately, biologically different situations can lead to the same value.

- The results obtained when the asynchrony metric proposed by Gross (2014) is used, i) depend on the number of species involved, ii) depend on the variability of each species (species with larger variances contribute more to the estimated asynchrony value) and iii) are inappropriate to study spatial asynchrony (as only a small subset of the data is used).

- Many currently-used stability metrics cannot discriminate between the effects of temporal and spatial stability. In particular, when spatial stability is quantified using these metrics, we are rather quantifying the sum of spatial and temporal stability.

- Human-induced fluctuations add variability to the response variable and thus decrease the estimated ecosystem stability. This noise also leads to an underestimation of differences in stability among groups.

- Most stability metrics assume a particular mean-variance relationship. The mean-variance relationship is expected to vary among types of data and is influenced by data transformations. This can lead to misleading results when the wrong mean-variance relationship is assumed.

- Comparing the stability of different ecosystems may be harder than commonly believed because many data sets analysed may not fulfil the strict assumption that their data is lognormally distributed (assumptions made by CV) or normally distributed (assumption made by the method proposed by Carnus (2015)).

- In general, most currently-used secondary metrics can only deal with continuous response variables. Analysing a subset of response variables may give us a biased understanding of ecosystem functioning.

Implications for future research

If the method presented here is used to analyse diversity-function experiments, we expect this to have the following beneficial consequences.

- **Better biological interpretation**

  My approach makes it easier to understand and interpret the results of these analyses. This is expected to improve our understanding of ecosystem functioning and how biodiversity modulates it. This is primarily due to the use of a shared framework to answer all questions addressed in these studies and the ease of interpretation of the obtained results. For example in Chapter 3 I compared, graphically and quantitatively, the values obtained for temporal stability and overyielding (see Figure 3.4).

- **Broader conclusions**

  A clear path to analyse non-continuous response variables has been shown. For

example, in Chapter 6, we analysed spatial asynchrony for survival (i.e. a proportion). Broader conclusions about the effects of biodiversity on ecosystem functioning are expected when additional types of response variables are analysed.

- **Better communication**

  The results are easier to communicate across scientific fields and also to policy makers. This is mainly due to the fact that mainstream methods are used, that all results lie on the same scale as the response variable, and that a strong focus is put on the graphical visualisation of the results. For example, in Chapter 2 I showed that, on average, the two-species mixtures of the Zurich experiment yielded 60 g/m^2 more than the corresponding monocultures. We also highlighted that this gain in productivity was variable among two-species mixtures (with some mixtures having larger gains and some with apparently no gain). These results were clearly illustrated with a simple graph (Figure 2.1) where all then ten post-hoc contrasts (e.g. `Al & An` vs `AlAn`) along with the global hypothesis (i.e. two-species mixtures vs monocultures) were illustrated.

- **Fast adaptability**

  This approach can easily adapt to new designs or research questions. This is convenient as new aspects of ecosystem functioning do not require the development of new metrics.

- **Flexibility**

  In order to drop design requirements (such as discreteness in measurements), or to analyse a particular kind of data, users can easily use extensions of Linear Mixed Effects Models. Generalised Additive Models (extensions of the methods used here) have the great potential to let these experiments evolve (see subchapter 7.6).

- **Benefits for other research fields**

  Other fields of ecology will benefit from using a similar unified approach that involves a shift of paradigm: Moving from the use of field-specific metrics to a

use of general modern techniques that can be tailored to take into account the specifics of each study. For example in Chapter 4, I showed how the specifics of the study (long-tailed data) can be dealt with by using mainstream statistical methods (Robust Regression). The original analysis that used field-specific methods led to an incorrect practical and biological interpretation of the results.

7.5 Evaluation of research

Contribution of the thesis

I developed a comprehensive framework to fully analyse modern diversity-function and I successfully applied the framework to several data sets. This approach aims to bring the focus back onto the biological meaning and relevance of the obtained results. This framework is well-suited to cope with the complexity of modern diversity-function experiments and it is flexible enough to be adapted to future questions and designs.

Limitations of the thesis

The approach presented here is based on the use of modern Linear Mixed Effects Models. This technique, although broadly used in ecology (Bolker et al., 2009), is not straightforward to understand and to use in all its facets. However, most researchers working in the field of diversity-function studies currently use that technique in their primary analyses. Therefore, the step needed to apply the technique to address secondary analysis questions is reasonable.

Some of the examples shown in the appendices, especially the starred ones, may be difficult to understand and implement by the user. This may suggest that the method is actually complex to use. On the contrary, the basic idea behind the new way to analyse these experiments is fairly straightforward and builds on the knowledge required to carry out the primary analysis. Complex examples are shown to highlight the great potential of this new approach. In particular, the approach presented here to quantify stability at different organisational level can get fairly involved. However,

this is mainly due to the fact that this ecological process is more complex than may be currently believed. Using simplistic solutions may lead to misleading results.

My analytical framework is well-suited to being extended by using closely related statistical methods such as Generalised Linear Mixed Effects Models (GLMMs) and Generalised Mixed Effects Models (GAMMs). This may indicate that the user needs to become a statistical expert in order to use this approach. This is not accurate, as to perform the main analyses carried out in diversity-function studies basic Linear Mixed Effects Models knowledge is required.

An important question that is currently addressed in these studies, is the relative importance of "sampling effect" and "complementarity" (Loreau and Hector, 2001), is not addressed in this thesis. The reason is that to address this type of questions the analysis should be carried out at the population level. Unfortunately, the software used here, and the underlying mathematical theory, are not yet fully-developed to analyse this kind of data. But I hope to address this in the future.

7.6 Future work

In the last consensus paper published in (2012) Cardinale and colleagues identified three aspects on which diversity-function studies need to focus: *Relevance*, *realism* and *predictive ability*.

My approach enables the user to judge the biological *relevance* of the results (as all estimated effects are on the same scale of the response variable), yields more *realistic* estimates (as it does not make any unrealistic assumptions) and it is well-suited to make *predictions* (as it is model-based).

Nevertheless, the statement made by Cardinale and colleagues also refers to the experiments themselves and not just the methods used to analyse them. Indeed, most research carried out so far has used simple communities composed of organisms of one trophic level (mostly producers; Cardinale et al. (2012, 2009)). Natural ecosystems usually harbour larger number of species which have complex interactions among them (Cardinale et al., 2012). Moreover, which species are lost from an ecosystem is

not a random process, as is the case in many diversity-function studies, but depends on species traits (Wardle et al., 2011; Suding et al., 2008; Fox, 2006; Larsen et al., 2005). The effects of biodiversity on ecosystem functioning are stronger at larger spatial scales and longer time frames (Cardinale et al., 2012). All these elements imply that in the future we need to design more complex experiments that are run for longer on larger scales.

To analyse such new generation experiments that include such complexities greater flexibility is required. Generalised Additive Mixed Models are the best candidate to cope with this evolution. Their use will allow the user to design and analyse new experiments where many discreteness assumptions are dropped. To mention an important example, the scale at which spatial asynchrony works will be estimated from the data rather than assumed to be known.

7.7 Conclusions

In conclusion, the analysis of biodiversity experiments has made rapid progress through the use of a variety of analytical approaches, but have involved the creation of many new metrics to quantify concepts like ecosystem stability. As experimental designs have become more complex, Linear Mixed Effects Models have become the natural starting point for analysis. I have shown how this Linear Mixed Effects Models framework can be extended to allow a more unified approach to the comprehensive analysis of many different types of response variables recorded from diversity-function experiments. This unified approach has statistical advantages and provides a better understanding of the biological processes that determine the relationship between biodiversity and ecosystem functioning.

Bibliography

F. Arenas, I. Sánchez, S. J. Hawkins, and S. R. Jenkins. The invasibility of marine algal assemblages: Role of functional diversity and identity. *Ecology*, 87(11):2851–2861, 2006.

A. D. Barnosky, N. Matzke, S. Tomiya, G. O. Wogan, B. Swartz, T. B. Quental, C. Marshall, J. L. Mc Guire, E. L. Lindsey, K. C. Maguire, et al. Has the earth's sixth mass extinction already arrived? *Nature*, 471(7336):51–57, 2011.

D. M. Bates. lme4: Mixed-Effects Modeling with R. *URL http://lme4. R-forge. R-project. org/book*, 2010.

T. Bell, A. K. Lilley, A. Hector, B. Schmid, L. King, and J. A. Newman. A linear model method for biodiversity-ecosystem functioning experiments. *The American Naturalist*, 174 (6):836–849, 2009.

B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White. Generalized Linear Mixed Models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3):127–135, 2009.

H. Bruelheide, K. Nadrowski, T. Assmann, J. Bauhus, S. Both, F. Buscot, X.-Y. Chen, B. Ding, W. Durka, A. Erfmeier, et al. Designing forest biodiversity experiments: General considerations illustrated by a new large experiment in subtropical China. *Methods in Ecology and Evolution*, 5(1):74–89, 2014.

B. Cardinale, E. Duffy, D. Srivastava, M. Loreau, M. Thomas, and M. Emmerson. *Biodiversity, ecosystem functioning, and human wellbeing: An ecological and economic perspective*, chapter : Towards a food web perspective on biodiversity and ecosystem functioning, pages 105–120. Oxford University Press, 2009.

B. J. Cardinale, J. E. Duffy, A. Gonzalez, D. U. Hooper, C. Perrings, P. Venail, A. Narwani, G. M. Mace, D. Tilman, D. A. Wardle, et al. Biodiversity loss and its impact on humanity. *Nature*, 486(7401):59–67, 2012.

T. Carnus, J. A. Finn, L. Kirwan, and J. Connolly. Assessing the relationship between biodiversity and stability of ecosystem function-is the coefficient of variation always the best metric? *Ideas in Ecology and Evolution*, 7(1), 2015.

G. Ceballos, P. R. Ehrlich, A. D. Barnosky, A. García, R. M. Pringle, and T. M. Palmer. Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Science Advances*, 1(5):e1400253, 2015.

J. Chamagne, M. Tanadini, D. Frank, R. Matula, C. Paine, C. D. Philipson, M. Svátek, L. A. Turnbull, D. Volařík, and A. Hector. Forest diversity promotes individual tree growth in central European forest stands. *Journal of Applied Ecology*, 54:71–79, 2016.

J. Connolly, T. Bell, T. Bolger, C. Brophy, T. Carnus, J. A. Finn, L. Kirwan, F. Isbell, J. Levine, A. Lüscher, et al. An improved model to predict the effects of changing biodiversity levels on ecosystem function. *Journal of Ecology*, 101(2):344–355, 2013.

D. Cox. Statistical significance tests. *British Journal of Clinical Pharmacology*, 14(3):325–331, 1982.

A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*, volume 1. Cambridge University Press, 1997.

R. Derpsch, A. Franzluebbers, S. Duiker, D. Reicosky, K. Koeller, T. Friedrich, W. Sturny, J. Sá, and K. Weiss. Why do we need to standardize no-tillage research? *Soil and Tillage Research*, 137:16–22, 2014.

D. F. Doak, D. Bigger, E. Harding, M. Marvier, R. O'Malley, and D. Thomson. The statistical inevitability of stability-diversity relationships in community ecology. *The American Naturalist*, 151(3):264–276, 1998.

J. E. Duffy. Why biodiversity is important to the functioning of real-world ecosystems. *Frontiers in Ecology and the Environment*, 7(8):437–444, 2009.

B. Everitt and T. Hothorn. *An Introduction to Applied Multivariate Analysis with R*. Springer Science & Business Media, 2011.

J. J. Faraway. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC press, 2005.

J. W. Fox. Using the Price Equation to partition the effects of biodiversity loss on ecosystem function. *Ecology*, 87(11):2687–2696, 2006.

J. D. Fridley. Resource availability dominates and alters the relationship between species diversity and ecosystem productivity in experimental plant communities. *Oecologia*, 132 (2):271–277, 2002.

J. D. Fridley. Diversity effects on production in different light and fertility environments: an experiment with communities of annual plants. *Journal of Ecology*, 91(3):396–406, 2003.

T. Friedrich, R. Derpsch, and A. Kassam. Overview of the global spread of conservation agriculture. *Field Actions Science Reports*, 6, 1941.

A. Gelman. Analysis of variance why it is more important than ever. *The Annals of Statistics*, 33(1):1–53, 2005.

K. E. Giller, J. A. Andersson, M. Corbeels, J. Kirkegaard, D. Mortensen, O. Erenstein, and B. Vanlauwe. Beyond conservation agriculture. *Frontiers in Plant Science*, 6, 2015.

K. Gross, B. J. Cardinale, J. W. Fox, A. Gonzalez, M. Loreau, H. W. Polley, P. B. Reich, and J. Van Ruijven. Species richness and the temporal stability of biomass production: A new analysis of recent biodiversity experiments. *The American Naturalist*, 183(1):1–12, 2014.

Y. Hautier, D. Tilman, F. Isbell, E. W. Seabloom, E. T. Borer, and P. B. Reich. Anthropogenic environmental changes affect ecosystem stability via biodiversity. *Science*, 348 (6232):336–340, 2015.

A. Hector. The effect of diversity on productivity: Detecting the role of species complementarity. *Oikos*, 82(3):597–599, 1998.

A. Hector, B. Schmid, C. Beierkuhnlein, M. Caldeira, M. Diemer, P. Dimitrakopoulos, J. Finn, H. Freitas, P. Giller, J. Good, et al. Plant diversity and productivity experiments in European grasslands. *Science*, 286(5442):1123–1127, 1999.

A. Hector, C. Philipson, P. Saner, J. Chamagne, D. Dzulkifli, M. O'Brien, J. L. Snaddon, P. Ulok, M. Weilenmann, G. Reynolds, et al. The Sabah Biodiversity Experiment: A long-term test of the role of tree diversity in restoring tropical forest structure and functioning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1582):3303–3315, 2011.

W. A. Hendricks and K. W. Robey. The sampling distribution of the coefficient of variation. *The Annals of Mathematical Statistics*, 7(3):129–132, 1936.

D. U. Hooper, E. C. Adair, B. J. Cardinale, J. E. Byrnes, B. A. Hungate, K. L. Matulich, A. Gonzalez, J. E. Duffy, L. Gamfeldt, and M. I. O'Connor. A global synthesis reveals biodiversity loss as a major driver of ecosystem change. *Nature*, 486(7401):105–108, 2012.

T. Hothorn, F. Bretz, and P. Westfall. *multcomp: Simultaneous Inference in General Parametric Models*, 2016. URL `https://CRAN.R-project.org/package=multcomp`. R package version 1.4-5.

A. R. Hughes and J. J. Stachowicz. Genetic diversity enhances the resistance of a seagrass ecosystem to disturbance. *Proceedings of the National Academy of Sciences*, 101(24):8998–9002, 2004.

M. A. Huston. Hidden treatments in ecological experiments: Re-evaluating the ecosystem function of biodiversity. *Oecologia*, 110(4):449–460, 1997.

F. Isbell, A. Gonzalez, M. Loreau, J. Cowles, S. Diaz, A. Hector, G. M. Mace, D. A. Wardle, M. O'Connor, E. J. Duffy, L. A. Turnbull, and P. L. Thompson. Linking the influence and dependence of people on biodiversity. *Nature*, submitted.

J. Koricheva. Sustainable forestry in temperate regions. In L. Bjork, editor, *Proceedings of the SUFOR International Workshop (Lund, Sweden. April 7-9)*, pages 152–153, 2002.

J. Koricheva and J. Gurevitch. Uses and misuses of meta-analysis in plant ecology. *Journal of Ecology*, 102(4):828–844, 2014.

T. H. Larsen, N. M. Williams, and C. Kremen. Extinction order and altered community structure rapidly disrupt ecosystem functioning. *Ecology Letters*, 8(5):538–547, 2005.

J. S. Lefcheck, J. E. Byrnes, F. Isbell, L. Gamfeldt, J. N. Griffin, N. Eisenhauer, M. J. Hensel, A. Hector, B. J. Cardinale, and J. E. Duffy. Biodiversity enhances ecosystem multifunctionality across trophic levels and habitats. *Nature Communications*, 6(3906), 2015.

F. Leisch. Sweave: Dynamic generation of statistical reports using literate data analysis. In W. Härdle and B. Rönz, editors, *Compstat 2002 — Proceedings in Computational Statistics*, pages 575–580. Physica Verlag, Heidelberg, 2002. ISBN 3-7908-1517-9.

E. Limpert and W. A. Stahel. Problems with using the normal distribution-and ways to improve quality and efficiency of data analysis. *PLoS One*, 6(7):e21403, 2011.

M. Loreau. Separating sampling and other effects in biodiversity experiments. *Oikos*, 82(3): 600–602, 1998.

M. Loreau. *From populations to ecosystems: Theoretical foundations for a new ecological synthesis (MPB-46)*. Princeton University Press, 2010.

M. Loreau and C. de Mazancourt. Species synchrony and its drivers: Neutral and nonneutral community dynamics in fluctuating environments. *The American Naturalist*, 172(2):48–66, 2008.

M. Loreau and C. de Mazancourt. Biodiversity and ecosystem stability: A synthesis of underlying mechanisms. *Ecology Letters*, 16(1):106–115, 2013.

M. Loreau and A. Hector. Partitioning selection and complementarity in biodiversity experiments. *Nature*, 412(6842):72–76, 2001.

M. Loreau, S. Naeem, and P. Inchausti. *Biodiversity and Ecosystem Functioning: Synthesis and Perspectives*. Oxford University Press, 2002.

M. E. Lundy, C. M. Pittelkow, B. A. Linquist, X. Liang, K. J. Van Groenigen, J. Lee, J. Six, R. T. Venterea, and C. Van Kessel. Nitrogen fertilization reduces yield declines following no-till adoption. *Field Crops Research*, 183:204–210, 2015.

K. S. McCann. The diversity–stability debate. *Nature*, 405(6783):228–233, 2000.

S. Naeem, L. J. Thompson, S. P. Lawler, J. H. Lawton, R. M. Woodfin, et al. Declining biodiversity can alter the performance of ecosystems. *Nature*, 368(6473):734–737, 1994.

S. Naeem, J. E. Duffy, and E. Zavaleta. The functions of biological diversity in an age of extinction. *Science*, 336(6087):1401–1406, 2012.

J. Neyman and E. L. Scott. Correction for bias introduced by a transformation of variables. *The Annals of Mathematical Statistics*, 31(3):643–655, 1960.

C. D. Philipson, P. Saner, T. R. Marthews, R. Nilus, G. Reynolds, L. A. Turnbull, and A. Hector. Light-based regeneration niches: Evidence from 21 dipterocarp species using size-specific RGRs. *Biotropica*, 44(5):627–636, 2012.

S. L. Pimm, C. N. Jenkins, R. Abell, T. M. Brooks, J. L. Gittleman, L. N. Joppa, P. H. Raven, C. M. Roberts, and J. O. Sexton. The biodiversity of species and their rates of extinction, distribution, and protection. *Science*, 344(6187):1246752, 2014.

J. Pinheiro and D. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer Science & Business Media, 2006.

C. M. Pittelkow, X. Liang, B. A. Linquist, K. J. Van Groenigen, J. Lee, M. E. Lundy, N. van Gestel, J. Six, R. T. Venterea, and C. van Kessel. Productivity limits and potentials of the principles of conservation agriculture. *Nature*, 517(7534):365–368, 2015a.

C. M. Pittelkow, B. A. Linquist, M. E. Lundy, X. Liang, K. J. Van Groenigen, J. Lee, N. Van Gestel, J. Six, R. T. Venterea, and C. Van Kessel. When does no-till yield more? A global meta-analysis. *Field Crops Research*, 183:156–168, 2015b.

C. Potvin and N. J. Gotelli. Biodiversity enhances individual performance but does not affect survivorship in tropical trees. *Ecology Letters*, 11(3):217–223, 2008.

D. S. Powlson, C. M. Stirling, M. Jat, B. G. Gerard, C. A. Palm, P. A. Sanchez, and K. G. Cassman. Limited potential of no-till agriculture for climate change mitigation. *Nature Climate Change*, 4(8):678–683, 2014.

I. Prieto, C. Violle, P. Barre, J.-L. Durand, M. Ghesquiere, and I. Litrico. Complementary effects of species and genetic diversity on productivity and stability of sown grasslands. *Nature Plants*, 1(4), 2015.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org/`.

R. Rosenthal and R. L. Rosnow. *Contrast analysis: Focused comparisons in the analysis of variance*. CUP Archive, 1985.

D. Sarkar. *Lattice: Multivariate data visualization with R*. Springer Science & Business Media, 2008.

B. Schmid, A. Hector, M. Huston, P. Inchausti, I. Nijs, P. Leadley, and D. Tilman. *Biodiversity and Ecosystem Functioning: Synthesis and Perspectives. Oxford University Press*, pages 61–75. 2002.

D. S. Srivastava and M. Vellend. Biodiversity-ecosystem function research: Is it relevant to conservation? *Annual Review of Ecology, Evolution, and Systematics*, pages 267–294, 2005.

J. R. Stevenson, R. Serraj, and K. G. Cassman. Evaluating conservation agriculture for small-scale farmers in Sub-Saharan Africa and South Asia. *Agriculture, Ecosystems & Environment*, 187:1–10, 2014.

K. N. Suding, S. Lavorel, F. Chapin, J. H. Cornelissen, S. Diaz, E. Garnier, D. Goldberg, D. U. Hooper, S. T. Jackson, and M.-l. Navas. Scaling environmental change through the community-level: A trait-based response-and-effect framework for plants. *Global Change Biology*, 14(5):1125–1140, 2008.

M. Tanadini and A. Hector. A modern model-based framework to compare ecosystem functioning. in prep.a.

M. Tanadini and A. Hector. A unified model-based framework for the analysis of biodiversity-ecosystem functioning experiments. in prep.b.

D. Tilman. Biodiversity: Population versus ecosystem stability. *Ecology*, 77(2):350–363, 1996.

D. Tilman. The ecological consequences of changes in biodiversity: A search for general principles. *Ecology*, 80(5):1455–1474, 1999.

D. Tilman and J. A. Downing. *Biodiversity and stability in grasslands*. Springer, 1996.

D. Tilman, D. Wedin, J. Knops, et al. Productivity and sustainability influenced by biodiversity in grassland ecosystems. *Nature*, 379(6567):718–720, 1996.

D. Tilman, C. L. Lehman, and K. T. Thomson. Plant diversity and ecosystem productivity: Theoretical considerations. *Proceedings of the National Academy of Sciences*, 94(5):1857–1861, 1997.

D. Tilman, C. L. Lehman, and C. E. Bristow. Diversity-stability relationships: Statistical inevitability or ecological consequence? *The American Naturalist*, 151(3):277–282, 1998.

D. Tilman, P. B. Reich, and J. M. Knops. Biodiversity and ecosystem stability in a decade-long grassland experiment. *Nature*, 441(7093):629–632, 2006.

D. Tilman, F. Isbell, and J. M. Cowles. Biodiversity and ecosystem functioning. *Annual Review of Ecology, Evolution, and Systematics*, 45(1):471, 2014.

S. L. Tuck, M. J. O'Brien, C. D. Philipson, P. Saner, M. Tanadini, D. Dzulkifli, H. C. J. Godfray, E. Godoong, R. Nilus, R. C. Ong, B. Schmid, W. Sinun, J. L. Snaddon, M. Snoep, H. Tangki, J. Tay, P. Ulok, Y. S. Wai, M. Weilenmann, R. Glen, and A. Hector. The value of biodiversity for the functioning of tropical forests: Insurance effects during the first decade of the Sabah Biodiversity Experiment. *Proceedings of the Royal Society B*, 283 (1844):20161451, 2016.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-PLUS*. Springer Science & Business Media, 2013.

E. Vojtech, M. Loreau, S. Yachi, E. M. Spehn, and A. Hector. Light partitioning in experimental grass communities. *Oikos*, 117(9):1351–1361, 2008.

D. A. Wardle and M. Jonsson. Biodiversity effects in real ecosystems-a response to Duffy. *Frontiers in Ecology and the Environment*, 8(1):10–11, 2010.

D. A. Wardle, R. D. Bardgett, R. M. Callaway, and W. H. Van der Putten. Terrestrial ecosystem responses to species gains and losses. *Science*, 332(6035):1273–1277, 2011.

R. L. Wasserstein and N. A. Lazar. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 2016.

O. R. Wearn, D. C. Reuman, and R. M. Ewers. Extinction debt and windows of conservation opportunity in the Brazilian Amazon. *Science*, 337(6091):228–232, 2012.

A. Weigelt, E. Marquard, V. M. Temperton, C. Roscher, C. Scherber, P. N. Mwangi, S. Felten, N. Buchmann, B. Schmid, E.-D. Schulze, et al. The Jena Experiment: Six years of data from a grassland biodiversity experiment. *Ecology*, 91(3):930–931, 2010.

S. Wood. *Generalized Additive Models: An introduction with R*. CRC press, 2006.

Y. Xie. *Dynamic Documents with R and knitr*, volume 29. CRC Press, 2015.

Y. Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2016. URL `https://CRAN.R-project.org/package=knitr`. R package version 1.13.