Journal of
Plant Ecology

VOLUME 10, NUMBER 1,
PAGES 91–110

FEBRUARY 2017

doi:10.1093/jpe/rtw107

available online at
www.jpe.oxfordjournals.org

A guide to analyzing biodiversity experiments

Bernhard Schmid[1,*], *Martin Baruffol*[1], *Zhiheng Wang*[1,2] *and Pascal A. Niklaus*[1]

[1] *Department of Evolutionary Biology and Environmental Studies and Zürich–Basel Plant Science Center, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland*
[2] *Department of Ecology and Key Laboratory for Earth Surface Processes of the Ministry of Education, College of Urban and Environmental Sciences, Peking University, Beijing 100871, China*
*Correspondence address. Department of Evolutionary Biology and Environmental Studies, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland. Tel: +41 79 681 99 36; E-mail: bernhard.schmid@ieu.uzh.ch

Abstract

Aims

The aim of this guide is to provide practical help for ecologists who analyze data from biodiversity–ecosystem functioning experiments. Our approach differs from others in the use of least squares-based linear models (LMs) together with restricted maximum likelihood-based mixed models (MMs) for the analysis of hierarchical data. An original data set containing diameter and height of young trees grown in monocultures, 2- or 4-species mixtures under ambient light or shade is used as an example.

Methods

Starting with a simple LM, basic features of model fitting and the subsequent analysis of variance (ANOVA) for significance tests are summarized. From this, more complex models are developed. We use the statistical software R for model fitting and to demonstrate similarities and complementarities between LMs and MMs. The formation of contrasts and the use of error (LMs) or random-effects (MMs) terms to account for hierarchical data structure in ANOVAs are explained.

Important Findings

Data from biodiversity experiments can be analyzed at the level of entire plant communities (plots) and plant individuals. The basic explanatory term is species composition, which can be divided into contrasts in many ways depending on specific biological hypotheses. Typically, these contrasts code for aspects of species richness or the presence of particular species. For significance tests in ANOVAs, contrast terms generally are compared with remaining variation of the explanatory terms from which they have been 'carved out'. Once a final model has been selected, parameters (e.g. means or slopes for fixed-effects terms and variance components for error or random-effects terms) can be estimated to indicate the direction and size of effects.

Keywords: analysis of variance, BEF-China, contrasts, linear models, mixed models, non-orthogonality, repeated measures, variance components

Received: 12 June 2016, Revised: 17 September 2016, Accepted: 26 September 2016

INTRODUCTION

In this article, we provide a guide for ecologists to analyze data from biodiversity–ecosystem functioning (BEF) experiments. We found that general textbooks or manuals are often too theoretical or technical to be of practical help. Furthermore, some of these guides suggest approaches such as maximum likelihood-based mixed models (MMs) or Bayesian analysis for BEF experiments without taking full account of the complexity of hierarchical data structures. Many pitfalls can be avoided if such data are first inspected with least squares (LS)-based linear models (LMs) where user-specified tests of biologically relevant hypotheses have to be constructed explicitly by comparing model terms in summary analysis of variance (ANOVA) tables. We focus on ANOVA in both LMs and MMs because of its flexibility to test such hypotheses with contrast terms (Rosenthal and Rosnow 1985). We use an original data set as practical example illustrating the different steps that we discuss and that interested readers may want to develop further on their own. These data were obtained in a pilot experiment of the BEF-China project (Bruelheide *et al*. 2011, 2014). In this experiment, we combined biodiversity with light treatments and measured tree growth over time, as described below.

After an introduction to typical BEF experiments and to the example data set, a subset of plot-level data of a single time point will be used to introduce LMs and ANOVAs with a single error term, which can be used for nonhierarchical data sets. From this, more complex LMs and ANOVAs with

multiple error terms, applicable for hierarchical data sets, will be developed. We then show how such data sets can be analyzed by MMs, where the hierarchical structure is reflected by random-effects terms instead of multiple error terms. Under 'Further model development', we will focus on the use of contrasts in ANOVA, perhaps our most important recommendation for the analysis of BEF experiments. Moving to the analysis of individual-level data, we will discuss repeated-measures analyses and non-orthogonality. Finally, we list recommendations in the 'Discussion' and in a final table.

Several aspects of the analysis of BEF experiments could not be included in this article because of space constraints. These included simulations to study data-generating mechanisms, data transformations and the use of covariates in ANOVA. We recommend Zuur *et al.* (2010) and Zuur and Ieno (2016) as useful guides to some of these and other aspects of the statistical analysis of ecological data. All analyses presented in this article can be repeated with the statistical software R (http://www.r-project.org, 10 October 2016, date last accessed). The corresponding code together with the data themselves can be found at https://github.com/pascal-niklaus/jpe, 10 October 2016, date last accessed, and in the Supplementary Data.

BEF EXPERIMENTS

General design of BEF experiments

Loss of biodiversity, in particular loss of species richness due to extinctions, is considered one of the major threats of global change affecting ecosystems (Rockström *et al.* 2009). To analyze this threat, it is necessary to use experiments, first because we want to find out how ecosystems may respond to continued future loss at a severity that has not yet occurred in the real world and second because naturally occurring variation in biodiversity among ecosystems is also present for reasons other than extinction. Thus, only if species richness is deliberately manipulated it can be treated as independent variable that causally explains variation in dependent variables such as biomass production (Schmid and Hector 2004). Such experiments are commonly called biodiversity–ecosystem functioning or in short BEF experiments.

The characteristic feature of all BEF experiments is that different species (or varieties, genotypes etc.) are combined in varying densities in experimental communities. These communities can be constructed synthetically, e.g. by sowing or planting individuals, or by removing individuals from existing communities. The design space of all BEF experiments can thus be drawn with a density axis for each species (variety, genotype etc.). The ecosystem function of interest, e.g. community productivity, can then be regarded as a function of the densities of all species and corresponding interaction terms between these densities. Obviously, filling this overall design space would result in a very large number of treatments, already with only two species (e.g. see van Kleunen *et al.* 2006). Therefore, convenient subsections of the entire design space are used in typical BEF experiments.

In the case of the most commonly used substitutive designs, subsections of the design space are constrained in such a way that all treatments have the same total density of individuals at the start of the experiment. As a consequence, only the mixing ratio among species varies. This is desirable, because mixing ratios reflect variation in species evenness and in species composition (if the density of some species is zero). While some substitutive designs such as the 'simplex design' (Kirwan *et al.* 2009) focus on varying evenness but not species composition, most designs used in larger BEF experiments only vary on in species composition (Balvanera *et al.* 2006). Only few BEF experiments use additive designs, where species composition is varied and total community density is proportional to the number of species, whose density is the same in monoculture and mixture. Not surprisingly, additive BEF experiments in the short term show significantly stronger species richness effects than substitutive BEF experiments, but at the expense that these might in part be caused by density (Balvanera *et al.* 2006).

The reason that most large BEF experiments focus on substitutive design and only manipulate species composition is that it is much easier to control species presence and absence than species density during the course of an experiment. In fact, although experimental communities are usually set up with initially equal numbers of individuals per species in mixture, i.e. maximum evenness, skewed rank–abundance distributions often result with time due to differential survival and vegetative or sexual reproduction of species (Hector *et al.* 2002). As a consequence, varying total density and species evenness, in addition to species composition, seems to have mainly short-term effects, as has been shown in a corresponding 2-year grassland experiment by Schmitz *et al.* (2013).

Within the subsection of the design space of substitutive and additive experiments with initially equal mixing ratio (species either present at a given initial density or absent), there are again a large number of further subsections that can be made according to the selection of species compositions, some of which are shown in supplementary Table S1. In principle, species compositions in all BEF experiments with different levels of species richness can be described by the combination of the presence/absence of the species in the pool from which the communities in the experiment are constructed. Designs differ in the way in which the set of all possible species combinations is reduced to address specific questions and to achieve an economic way to implement the experiment.

Initial BEF experiments applied a single extinction scenario to a complete ecosystem; i.e. each level of species richness was represented by a single, but replicated, species composition (Naeem *et al.* 1994; Niklaus *et al.* 2001). In subsequent BEF experiments in grassland, several different, randomly selected species compositions were sown at each species richness level, concomitantly varying the number of plant functional groups (e.g. Hector *et al.* 1999; Tilman *et al.* 1996). Number of species and functional group compositions were for the first time independently varied in the so-called Jena Experiment (Le Roux *et al.* 2013; Roscher *et al.* 2004). At the same site a further

experiment was added later that chose species compositions reflecting different levels of trait variation (Ebeling *et al.* 2014). With these experiments, BEF relationships could be generalized beyond the particular species pool investigated. More recently, BEF experiments have been set up with more systematically selected permutations of species compositions, e.g. ensuring that all species occur in communities at all richness levels or that all species compositions at lower richness levels also occur as subsets of species compositions at gradually higher richness levels (Bruelheide *et al.* 2014). All these designs use either complete or random subsets of species compositions within the given constraints. However, there are also BEF experiments where species compositions are deliberately selected along non-random extinction scenarios (Bruelheide *et al.* 2014), attempting to mimic more realistic extinction drivers such as eutrophication (Schläpfer *et al.* 2005). Typical features of BEF experiments are described in more detail in Schmid *et al.* (2002) and Bruelheide *et al.* (2014).

BEF experiments with initially equal mixing ratio have the advantage that at the analysis stage the focus can be on modeling the influence of species presence and absence on ecosystem functions, which is the focus of the present article. The more general approach of using species densities, mentioned at the beginning of this section, is practically never used and may be inappropriate if density effects are not linear. For example, Le Roux *et al.* (2013) found that the contribution of legumes to the studied soil functions could be best modeled by a term for legume presence and an additional term for sowing density of legumes, modulating the effect of their presence. Another possibility is to use the so-called realized densities and realized richness or diversity measures to analyze BEF experiments. We will not discuss this in the present article, because we focus on design variables and do this for the following reason. While it can be useful to add measured variables as explanatory covariates to search for potential mechanisms underpinning effects of design variables, it also means that the advantages of a manipulative experiment are given up in favor of a correlative observational study. This can be illustrated with the following thought experiment: two experimental communities with 16 and 4 species are set up with replication. At the end of the experiment, all communities have 4 species due to extinctions in the 16-species communities. However, the effect of the initial species richness is still highly significant, because the surviving species in the 16-species treatment make better 4-species realized communities than those of the 4-species treatment without extinctions. Obviously, this effect would be hidden if the realized species richness of four would be used as explanatory variable instead. This thought experiment also explains why design variables such as species richness are used in the analysis of ecosystem functions such as harvest yield, even if not all species of experimental communities are included in the harvest sample.

A pilot experiment of BEF-China as working example

The pilot experiment of the BEF-China project was set up to compare the early growth of subtropical forest tree species under different environmental conditions, in particular different settings of intra- and interspecific competition. The broader objective of this study was to develop an understanding of competitive interactions between species to later interpret the effects occurring during the early establishment of subtropical forest communities in a larger long-term biodiversity experiment of the BEF-China project (the 'main experiment', described in Bruelheide *et al.* 2014; Bu *et al.* 2017; Hahn *et al.* 2017; Li *et al.* 2017; Peng *et al.* 2017; Sun *et al.* 2017). Special features of the presented pilot experiment are the use of three instead of only one species pool and the use of two environments (light and shade). The experiment could thus be considered as consisting of six BEF sub-experiments, each carried out with a different species pool or under different environmental conditions.

Plots 1 × 1 m in size were planted with experimental communities comprising 16 young trees arranged in a 4 × 4 grid. These communities were assembled using 12 common tree species of natural forests of the area, belonging to 3 functional groups, i.e. evergreen conifers, evergreen angiosperms and deciduous angiosperms. These 12 species were first grouped into three pools (X, Y and Z), each containing evergreen and deciduous functional groups (supplementary Table S2 in Supplementary Data). For each pool, we set up all monocultures and 2-species mixtures and the 4-species mixture, yielding 11 community compositions. Because the 3 pools did not overlap with respect to species, there were 33 distinct species compositions altogether. All species compositions were grown under full light and under shade (5% of full light), resulting in 66 treatment combinations. These were replicated in 4 blocks, resulting in a total of 264 plots or 88 per species pool (Fig. 1). Seven plots were not correctly established. As a consequence, pool Y had only 85 and pool Z only 84 plots.

Tree growth was monitored from April 2009 to September 2010. Our hypothesis was that, on average, trees would grow better in plots with 4 than with 2 species and better in 2-species plots than in monoculture. We further expected slower growth and smaller biodiversity effects under shade than in full light. Growth was followed nondestructively through time by measuring tree height and basal stem diameter (5 cm above ground). To obtain plot-level measures of ecosystem functioning, related to productivity, we summed the individual-level measures; dead individuals were included with zero value in these sums.

ANOVA FOR LM AND MM

Analysis of plot-level data using LMs and ANOVAs with single error terms

It is often convenient to start the analysis of BEF experiments with a subset of data and a visual inspection of data plotted as a function of biodiversity, typically species richness. Figure 2 displays the plot-level data (n = 88) of total basal area, 14 months after planting, of all communities assembled from species pool X. Visual inspection of the figure suggests that the total basal area per plot (`ba`, Courier font here and

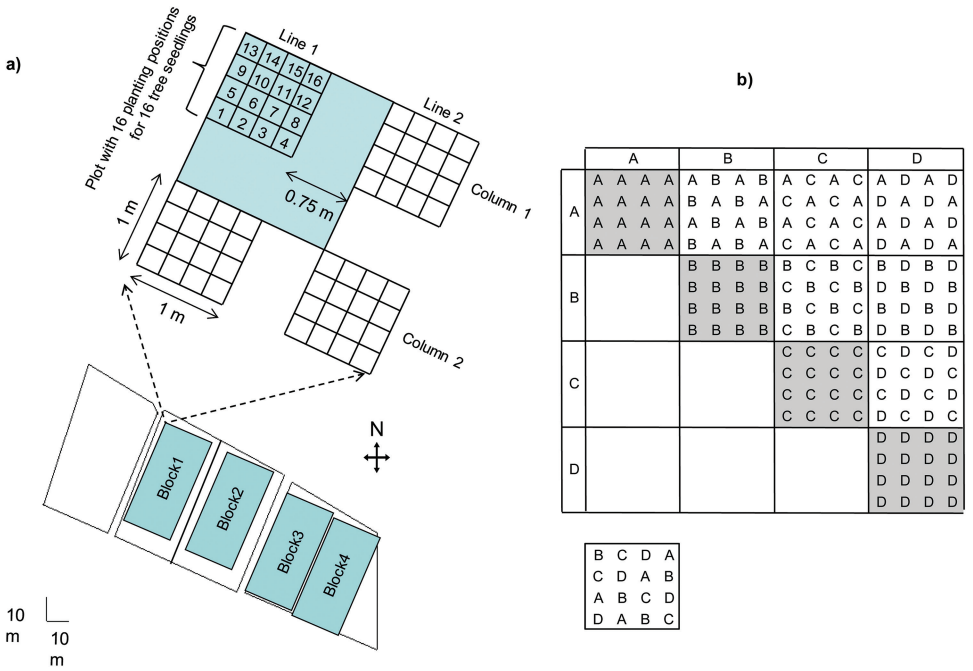

Figure 1: layout of BEF experiment in Xingangshan, Jiangxi, China. **a**) Four blocks were marked in the field. Each block contained 66 plots of the experiment presented in this article (4 are enlarged to explain distances) and within each plot there were 16 planting positions for tree seedlings. Note that in addition to the experiment presented in this article, there were similar plots from other experiments randomly interspersed with the ones of this experiment. **b**) Principle of experiment showing the four monocultures, six 2-species mixtures and, below the bottom left corner, the 4-species mixture for a single species pool with species A, B, C and D. Note that the scheme does not show the random arrangement of plots and plant species within plots as used in the concrete planting of the experiment.

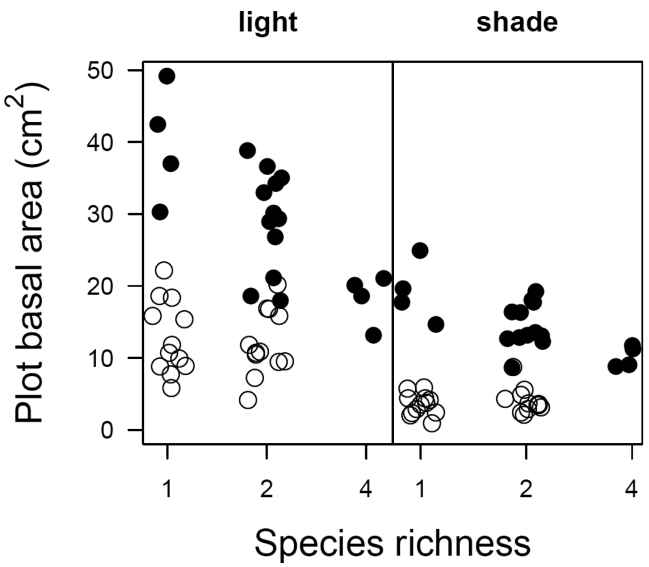

Figure 2: total basal stem area per plot of surviving trees 14 months after planting in plots of species pool X. Left panel represents plots in light, right panel plots in shade. The four species *Schima superba*, *Elaeocarpus decipiens*, *Castanea henryi* and *Quercus serrata* were grown in four monocultures, all six possible 2-species combinations and the single 4-species combination. Plots containing *Elaeocarpus decipiens* have filled symbols; these have particularly large total basal stem area.

in the subsequent occurrences indicates the use of a name in the statistics script) is larger in light than in shade (treatment `light`). Furthermore, there is little indication of species richness effects (`fdiv`, the `f` indicating that `fdiv` represents a factor with levels rather than a continuous variable), except that plots with the dominant species *Eleocarpus decipiens* tend to have a lower basal area the more other species they contain (filled symbols in Fig. 2). Furthermore, there is a large scatter of values in particular among monocultures and 2-species mixtures; this may be due to differences among species compositions (`com`) and plots (`plot`) within species compositions. Because there is only one species composition at the highest diversity level, the scatter of values at this level should only be due to environmental differences among plots.

To test whether there is indeed systematic variation in the visualized plot-level data, a sequence of LMs, which account for this systematic variation, can be fitted using an LS approach, which minimizes nonsystematic variation. This nonsystematic variation can be considered as random variation or noise, caused by various influences that cannot be controlled by the experimenter. The goal of the analysis is to find an LM that can assign a large part of the total variation in the data to systematic variation, leaving a small amount of unexplained random variation. The statistical model is formulated in such a way that a data value is a linear function of systematic effects and a residual, or error, pertaining to the particular value. Thus, the

simplest model would assume that the dependent variable `ba` is only measurement error and does not exist at all, i.e. has zero mean value. However, more commonly the modeling process starts with the fitting of an overall mean, such that the errors are the differences between a data point and the overall mean. In this model, the unexplained random variation is the variance of the dependent variable, which is then referred to as the total variation, to be partitioned into systematic and (residual) random variation by adding terms to the statistical model.

In a first analysis, the total variation in `ba` is partitioned into systematic variation due to light and diversity treatments and residual random variation, unexplained by these explanatory factors. The statistical model could be formulated as LM in the following way:

$$\mathrm{ba}_{ijk} = m + a_{\mathrm{i}} + b_{\mathrm{j}} + e_{ijk}$$

where m represents the overall mean, a_i the effect of the i-th light treatment, b_j the effect of the j-th diversity treatment and e_{ijk} the error or deviation of ba_{ijk} from the value predicted by the previously fitted effects, i.e.:

$$e_{ijk} = \mathrm{ba}_{ijk} - \left(m + a_i + b_j\right)$$

There are alternative formulations for this LM where the overall mean is replaced, e.g. by the mean of the first group ($i = 1$ and $j = 1$), but instead of writing out the LM with parameters as above, it is more convenient to only specify the terms contributing to systematic variation. In the statistical software R (http://www.r-project.org), which will be used for all analysis presented in this article, this is done in the following way:

```
ba ~ light + fdiv        (LM1)
```

If this LM1 is fitted with the LS approach (R-functions `aov` or `lm`), parameters for the effects of the terms `light` and `fdiv` will be estimated in such a way that the residual variation (`residuals`) will be minimized. The systematic variation explained by the two so-called 'fixed-effects' terms `light` and `fdiv` can be listed together with the `residuals` in an ANOVA table (Table 1a). The `residuals` represent the so-called 'random-effects' term. If there are no other random-effects terms present, all fixed-effects terms can be compared with the `residuals` using variance ratios as explained below. LMs or ANOVAs where all random variation is contained in the `residuals` are called nonhierarchical, in contrast to hierarchical LMs, MMs and ANOVAs, which are characterized by more than one random-effects term and will be discussed further below.

The first column in Table 1a lists the explanatory terms or 'sources of variation' in the dependent variable `ba`. We have added an additional row for the total variation in `ba`. The second column (Df) lists the degrees of freedom. Degrees of freedom indicate the number of 'independent pieces of information' contained in a data set or accounted for by a model term. The Dfs in Table 1a are one less than the number of different levels, or groups, specified by the explanatory terms (2–1 for `light` and 3–1 for `fdiv`). Although Dfs are typically one less than the number of levels of an explanatory term, they can be further reduced for terms where some of the effects for which they encode have already been 'eliminated' by other terms (e.g. if a contrast between groups of levels, say monocultures vs. mixtures in the case of `fdiv`, is fitted first—see below). The Df of the `residuals` in Table 1a is what remains after fitting the overall mean and the explanatory terms, i.e. 88–1–1–2 = 84. Here, the overall mean is fitted implicitly but omitted from the ANOVA output unless this is specifically requested (`summary.aov(m, intercept=TRUE)`). Note that in the output from R, the term for the overall mean in the ANOVA will be called '`intercept`'.

Table 1: ANOVA table (a) and variance components (b) for LM1 in the text, modeling plot-level basal area (`ba`) as a function of the fixed-effects terms `light` (control vs. shade) and `fdiv` (monocultures, 2-species mixtures, 4-species mixture)

a)

Source	Df	SS	%SS	MS	EMS	*F*	*P*
light	1	2824	29.0	2824	$44\ \sigma^2_{\mathrm{light}}+\sigma^2_{\mathrm{res.}}$	34.44	< 0.001
fdiv	2	39	0.4	20	$16.94\ \sigma^2_{\mathrm{fdiv}}+\sigma^2_{\mathrm{res.}}$	0.24	0.790
residuals	84	6888	70.6	82	$\sigma^2_{\mathrm{res.}}$		
total	87	9751	100.0	11			

b)

$\mathrm{VC}_{\mathrm{light}}$ (estimate of $\sigma^2_{\mathrm{light}}$) =	$(\mathrm{MS}_{\mathrm{light}} - \mathrm{MS}_{\mathrm{res.}})/44$ =	62.32
$\mathrm{VC}_{\mathrm{fdiv}}$ (estimate of σ^2_{fdiv}) =	$(\mathrm{MS}_{\mathrm{fdiv}} - \mathrm{MS}_{\mathrm{res.}})/16.94$ =	–3.66
$\mathrm{VC}_{\mathrm{res.}}$ (estimate of $\sigma^2_{\mathrm{res.}}$) =	$\mathrm{MS}_{\mathrm{res.}}$ =	82.00

Abbreviations: Df = degrees of freedom, SS = sum of squares, %SS = SS in percent of total (corresponding to increments in multiple R^2), MS = mean squares (SS/Df), EMS = expected mean squares (linear combination of variance components), *F* = variance ratio (MS of term/MS of `residuals`), *P* = probability of type-I error, VC = variance component (estimated from the MS using the equations given in the EMS column). LM1 was fitted with the function `aov`. In this nonhierarchical analysis there is only one error term, i.e. the `residuals`. Note that the row 'total' and the columns %SS and EMS in a) and all items shown in b) are not part of the standard output obtained after the `aov` function in R.

The third column in Table 1a contains the sum of squares (SSs) contributed by a term to the total sum of squared deviations of observations from the overall mean. That is, in an initial model in which only the overall mean is fitted, the total SS is the same as the SS of the `residuals`. With each row added to the ANOVA, the latter decreases by the amount listed as SS for the added term. In simple cases, the SS of a term can also be viewed as the sum of squared deviations of the group means of the term from the overall mean, but in cases with the so-called non-orthogonal terms (see below) this no longer holds. Often it is useful to add a column in which SSs are expressed as percentage of the total (fourth column in Table 1a; not part of the standard output of R), because these values can be used as measures of effect sizes (Cohen 2002; Rosenthal and Rosnow 1985). Furthermore, because terms are sequentially fitted, starting with the first row and moving down the ANOVA table, %SS values represent increments in multiple $R^2 \cdot 100\%$. For LM1, the multiple R^2 is 0.294 (= (29.0 + 0.4)/100); in other words, the systematic variation explained by LM1 is 29.4% of the total variation in `ba`.

Although SSs are additive and can be used to calculate effects sizes and 'explained' variation, they are not corrected for their Dfs (which are also additive). To obtain variances, the SSs are divided by their Dfs. These mean squares (MSs), which are no longer additive (i.e. they do not add up to the MS of the total), are listed in the fifth column in Table 1a. The MSs are used to derive variance ratios for significance tests (F values in the seventh column of Table 1a). The MS for a certain term will exceed the MS of the denominator term if the first term explains more variance than one would expect by chance. That is, an F value >> 1 indicates a statistically significant effect of the term, which leads to a difference between the two variances that are used in the calculation of the F value. The last column in Table 1a provides the P values for the significance tests, which are based on the F values and the Dfs from the nominator and denominator MSs used to calculate the F values. With $P < 0.001$, the corresponding F value would only be expected in < 1 out of 1000 replications of an experiment *if* there was no 'true' effect. Therefore, the null hypothesis of no effect can be rejected. In the example of Table 1a, the significant effect of `light` indicates that `ba` was larger in light than in shade, whereas the nonsignificant effect of `fdiv` confirms the suspicion from the visual inspection that different levels of `fdiv` had similar `ba`. In fact, $F < 1$ for `fdiv` even suggests that on average two communities randomly selected from different levels would be more similar than two communities randomly selected from the same level.

In nonhierarchical ANOVA, with `residuals` as the only random-effects term, the denominator of the F values is always the MS of the `residuals`. However, with the so-called hierarchical data the MS of the residuals does not provide the correct denominator for all terms to be tested. To construct appropriate F values, it is necessary to know which variance *components* (VCs) are contained in an MS, and therefore, we introduce the important concept of expected mean squares (EMSs) here and add them in a separate column between MSs and F values to Table 1a. It has been shown that EMSs can be considered as linear combinations of VCs (e.g. see Green and Tukey 1960) and rules to calculate coefficients of VCs are, e.g., given in Snedecor and Cochran (1989). However, to find out which VCs are contained in an EMS requires both statistical and biological understanding of the data and can vary with the questions being asked (Hector *et al.* 2011; Nelder 1994; Nelder and Lane 1995; Searle 1971). The EMS of the `residuals` contains one VC only, $\sigma^2_{res.}$, which measures the residual random variance in the data that cannot be explained by the model. The EMS of each other term equals σ^2_{res} plus a VC that accounts for the differences between groups, here between light treatments (σ^2_{light}) or between diversity treatments (σ^2_{fdiv}). These variance components are multiplied with coefficients reflecting the number of replicates in each treatment group. In the case of unequal group sizes, a good approximation is their harmonic mean (Snedecor and Cochran 1989). For `fdiv`, this is $16.94 = 3/(1/(4 \times 8)+1/(6 \times 8) + 1/8)$, where the different components in the dominator indicate the number of replicates for monocultures ($4 \times 8 = 4$ species × 2 light levels × 4 blocks), 2-species mixtures ($6 \times 8 = 6$ 2-species compositions × 2 light levels × 4 blocks) and the 4-species mixture (8 = 1 species composition × 2 light levels × 4 blocks).

Given the EMS equations in Table 1a, it becomes clear that F values are always calculated in such a way that the two variances to be compared only differ in one VC. Therefore, $F > 1$ indicates that the additional VC in the EMS of the nominator is >0. That $F < 1$ for `fdiv` in the example shows an important specific case where the corresponding estimate for σ^2_{fdiv} is negative (Table 1b). Values of $F < 1$ always indicate negative VCs. Although variances are never negative, VCs as components of variances need not be constrained to positive values (Nelder 1977). Nevertheless, in MMs using maximum likelihood (ML) or restricted/residual maximum likelihood (REML) methods, negative VCs are often not allowed and therefore restricted to zero or very small positive values in the fitting process. This can be dangerous because negative VCs can indicate that the EMS in the denominator contains more VCs than specified in the EMS, i.e. that the residual 'noise' is not just random but contains systematic components or correlated data. Negative VCs are often found when explanatory terms that are not included in the model contribute to the MS of the `residuals` but not to that of the tested term. (For example, this commonly occurs with plot residuals in ANOVAs for hierarchical data from split-plot experiments, when the split-plot treatment is omitted from an LM or MM.) As long as a final model has not been found, we recommend to allow values of $F < 1$ and VCs < 0. Not doing so, which is particularly tempting in MMs, can lead to F tests that are too liberal (see below: section 'Individual-level data and repeated-measures analyses'). Beyond guiding the appropriate construction of F values for significance tests in ANOVA, VCs can also be used

as an alternative to %SSs for measuring the size of effects of explanatory terms (Hector *et al.* 2011). This measure gives more importance to terms with fewer Dfs than is the case for %SSs; thus the effect of `light` is almost as large as that of the `residuals` when measured in terms of VCs, but less than half of that of the `residuals` when measured in terms of SSs (Table 1).

The ANOVA table of LM1 contains information about SSs, MSs, *F* values, *P* values and VCs. Another goal of fitting LMs is to obtain parameter estimates, in the present case, group means and their standard errors or differences between groups and standard errors of these differences. Because these estimates reflect the fitted model and the significance tests derived from it, these estimates are generally preferred over means and standard errors calculated form the raw data. In the case of a model with orthogonal treatments, the LS means are equal to the raw means. However, the standard errors calculated during the model fit are normally different from those calculated directly from the data because the model assumes that the true variation is the same for all measures and thus standard errors of estimates can be calculated from the standard deviation of the `residuals`, which is the square root of the MS of the `residuals` or *s*. The standard error of a group mean equals $s/\sqrt{n}$, where *n* is the number of data points in the group. If two groups have the same size, the standard error of the difference between them is simply the standard error of the group means multiplied by $\sqrt{2}$. Parameter estimates can be obtained with the R-function `summary.lm`. How estimated parameters are combined to describe the modeled data depends on how the model matrix relating these to the data is constructed. By default `aov` uses the so-called 'treatment contrasts' in which the first group (in alphabetical order) is given as `Intercept`; the additional parameters estimate the other group means as difference to the first group (which type of contrast coding is used can be changed using `options(contrasts=…)`).

Analysis of plot-level data using LMs and ANOVAs with multiple error terms (hierarchical models)

The above LM1 did not contain a term for the different species compositions (`com`), and the associated variance was thus included in the `residuals`. Including `com` leads to a hierarchical statistical model with multiple random-effects or error terms. If such models are analyzed as LMs, it is useful to assemble the error terms in a so-called error model (Payne *et al.* 1993):

`ba ~ com + light × com + plot`

This model is comprehensive, because it fits a mean for each of the 11 levels of `com`, for each of the 22 combinations of `light × com` (the interaction modeling the combination of the 2 light treatments with the 11 species compositions) and for each of the 88 plots. Fitting this model leaves no `residuals` because these are fully captured by the term `plot`. To obtain significance tests (*F*- and *P* values), `plot` thus has to be omitted:

`ba ~ com + light × com` (LM2)

Note that `residuals` now only has Df = 66 (Table 2a), because Df = 10 have already been consumed by `com` and Df = 11 by `light × com` (if `com` would have been omitted, `light × com` would have Df = 21). In the ANOVA (Table 2a), these two explanatory terms are treated as fixed-effects terms and tested 'against' `residuals`. For `com`, this is not appropriate because its EMS has two VCs more than `residuals`. It thus remains unclear whether the significance of `com` implies that VC_{com} or $VC_{light \times com}$ or both are >0. A corrected *F* value for `com` could be obtained by dividing MS_{com} by $MS_{light \times com}$.

The error model LM2 includes the systematic variation previously explained by LM1, which is also called treatment model (Payne *et al.* 1993). This is because `fdiv` is included in `com` and `light` is included in `light × com` (as is `light × com`, but we do not consider this yet). Here, the terms of the treatment model represent fixed-effects contrasts that can be 'carved out' from the corresponding terms of the error model. Note that these error terms correspond to random-effects terms in the MMs discussed further below. Using LMs for the analysis of hierarchical data requires combining treatment and error models, i.e. LM1 and LM2, respectively, in a new LM3, where all terms are fitted sequentially, making sure that contrasts are always fitted before the terms from which they have been carved out:

`ba ~ light + fdiv + com + light × com` (LM3)

From the EMS inserted in the ANOVA (Table 2c), VCs can be calculated for explanatory terms (Table 2d). Comparing the VCs in Table 2b and d shows that $VC_{light \times com}$ has decreased in LM3 because it no longer contains the very large fixed effects of `light`. In contrast, VC_{com} has increased because $MS_{light \times com}$, which is used in the calculation of VC_{com}, has decreased from 297 to 45 (Tables 2a and c, respectively). VC_{fdiv} is still negative although MS_{fdiv} is now slightly larger than $MS_{res.}$ (Table 2c). However, to construct appropriate significance tests, *F*- and *P* values must be calculated with those error terms in the denominator that differ by only one VC from the tested terms. These corrected *F* values test whether the respective VCs are >0 (Table 2c): $F_{light} = MS_{light}/MS_{light \times com}$, $F_{fdiv} = MS_{fdiv}/MS_{com}$ and $F_{com} = MS_{com}/MS_{light \times com}$. The new ANOVA in Table 2c confirms that `ba` is significantly larger in plots with full light than under shade (`light`) and that the differences between the three diversity levels are not statistically significant (`fdiv`). However, there are significant differences among particular species compositions within diversity levels (significant effects of `com`)—against which the differences between diversity levels are tested—and among particular species compositions in their response to shading (interaction `light × com`). This significant interaction, like all 2-way interactions, can also be biologically interpreted in a second way, namely that differences between species compositions are not the same in full light and under shade.

Table 2: ANOVA tables (a, c) and variance components (b, d) for LM2 and LM3 in the text, modeling plot-level basal area (`ba`)

	Df	SS	MS	EMS	*F*	*P*
a)						
com	10	5242	524	$8\,\sigma^2_{com} + 4\,\sigma^2_{light \times com} + \sigma^2_{res.}$	27.9	< 0.001
light × com	11	3270	297	$4\,\sigma^2_{light \times com} + \sigma^2_{res.}$	15.8	< 0.001
residuals	66	1239	19	$\sigma^2_{res.}$		
b)						
VC_{com} =				$(MS_{com} - MS_{light \times com})/8$ =	28.38	
$VC_{light \times com}$ =				$(MS_{light \times com} - MS_{res.})/4$ =	69.50	
$VC_{res.}$ = (VC_{plot} =)				$MS_{res.}$ = (MS_{plot} =)	19.00	
c)						
	Df	SS	MS	EMS	*F*	*P*
light	1	2824	2824	$44\,\sigma^2_{light} + 4\,\sigma^2_{light \times com} + \sigma^2_{res.}$	62.756	<0.001
fdiv	2	39	20	$16.94\,\sigma^2_{fdiv} + 8\,\sigma^2_{com} + 4\,\sigma^2_{light \times com} + \sigma^2_{res.}$	0.031	0.970
com	8	5202	650	$8\,\sigma^2_{com} + 4\,\sigma^2_{light \times com} + \sigma^2_{res.}$	14.444	< 0.001
light × com	10	447	45	$4\,\sigma^2_{light \times com} + \sigma^2_{res.}$	2.368	0.018
residuals	66	1239	19	$\sigma^2_{res.}$		
d)						
VC_{light} =				$(MS_{light} - MS_{light \times com})/44$ =	63.2	
VC_{fdiv} =				$(MS_{fdiv} - MS_{com})/16.94$ =	−37.2	
VC_{com} =				$(MS_{com} - MS_{light \times com})/8$ =	75.6	
$VC_{light \times com}$ =				$(MS_{light \times com} - MS_{res.})/4$ =	6.5	
$VC_{res.}$ = (σ^2_{plot} =)				$MS_{res.}$ = (MS_{plot} =)	19.0	

These LMs were fitted with the function `aov`. In contrast to Table 1 these are hierarchical analyses with multiple error terms, i.e. `com` (species composition, 11 groups), `light × com` (interaction term, 22 groups) and `residuals`. F- and *P*-values in c) are recalculated using appropriate error terms, i.e. `com` for `fdiv`, and `light × com` for `light` and `com`. The corrected *F* values form ratios of MS in such a way that the EMS differ in a single VC. See text and legend to Table 1 for further explanations.

The point that the fixed-effects term `fdiv` should be tested against the remaining variation among species compositions within diversity levels (`com` appearing after `fdiv` in the model formula) is a specialty in the analysis of BEF experiments (Schmid *et al.* 2002). Biologically it can be justified because species richness is a property emerging from the specific composition of a community, i.e. community composition is the unit of replication for species richness, thus representing an additional hierarchical level in the ANOVA. Variation between plots of the same composition thus reflects variation unrelated to factors that 'generate' species richness. A true species richness effect will be indicated by a difference between two randomly chosen species compositions of different levels of species richness that on average is larger than a difference between two randomly chosen species compositions of a single level. Statistically it can be justified if we consider `fdiv` a fixed-effects and `com` an error or, in MM terminology, a random-effects term. In this case, the rule according to Snedecor and Cochran (1989) is that the fixed-effects term includes VCs of all random-effects terms nested within it, i.e. the EMS of `fdiv` includes σ^2_{com} (Table 2c). The term `com` is nested within `fdiv`, which in R can be expressed in short using the nesting operator '/':

```
fdiv / com ⇔ fdiv + fdiv : com
```

However, because `com` is coded with different values throughout the levels of `fdiv`, the following formulae are equivalent (Wilkinson and Rogers 1973):

```
fdiv / com ⇔ fdiv + fdiv : com ⇔ fdiv + com
```

If `com` would be considered a fixed-effects term, then the EMS of `fdiv` would not include a VC_{com} and as a consequence MS_{fdiv} should not be compared with MS_{com} in the *F*-test. We further elaborate this point in the following section.

Which terms belong to the treatment and which to the error model?

There are no strict rules as to whether a term should be assigned to the treatment or the error model or, in the terminology of MMs, whether a term should be considered as

fixed- or random-effects term. Sometimes both could reasonably be justified. For example, if inferences about diversity effects are to be made about a larger statistical population such as all possible species compositions that could be assembled from a regional species pool, then the corresponding term `com`, whose levels represent a random sample of these, needs to be treated as error or random-effects term in LMs or MMs, respectively, and `com` used as unit of replication for `fdiv`. If, on the other hand, inferences about diversity effects are restricted to the particular species compositions investigated, then `com` can be treated as a fixed-effects term and `residuals` (= `plot`) be used as unit of replication for `fdiv`. However, in this second case, inferences cannot be generalized beyond the investigated set of species compositions. For fixed-effects terms, we are interested in the individual group means (e.g. for monocultures, 2-species mixtures and 4-species mixtures) or their differences (e.g. between light vs. shade means). For an error or random-effects term, however, we are interested in the VC, because we consider the different levels or groups as a random sample out of many more groups that could have been selected and focus on their degree of variation rather than on specific levels. Green and Tukey (1960) provide a simple rule of thumb: the levels of a fixed-effects term represent 'all out of few' and the levels of an error or random-effects term represent 'few out of many', where the second 'few' is larger than the first.

For the ANOVA of a model with multiple error terms, the basic `aov` (and `lm`) function has its limits because by default all explanatory fixed-effects terms are tested against the `residuals`, which is the only random-effects term. As shown above, it is nevertheless possible to calculate VCs, an item typically assigned to random-effects terms, and to 'manually' calculate appropriate *F* values at multiple error strata.

Often, *F* values for error or random-effects terms are not shown in ANOVA tables, because the significance of VC > 0 for these terms is of little interest (but note their use as measures of effect sizes by Hector *et al.* (2011) mentioned earlier). Nevertheless, our suggestion is to list random-effects terms in the same way as fixed-effects terms or to present their VCs in a second part of a table as shown in Tables 1 and 2. First, this allows others to check which error or random-effects terms have been included in a model and how much they contributed to hierarchical structure. Second, these terms can have biological meaning such as in the case of `com`, where a significantly positive VC can be interpreted as significant variation among species compositions. It may then be reasonable to look for further fixed-effects contrasts that could possibly be carved out from `com`, e.g. contrasts for the presence vs. absence of particular species in an experimental community (see below).

Multiple error terms can be accommodated with an extension of the model formula syntax available in `aov`, which allows to add error strata in `Error(...)`:

```
aov(ba ~ light + fdiv + Error(com + light:com))
```

We could also write `+Error(com/light)`, which would expand to `Error(com + light:com)`. The function `summary` yields separate ANOVA tables for each error stratum; this is a very powerful aspect of the LM approach to derive ANOVA tables (Gelman 2005). A specialty of this approach is that fixed-effects terms will be tested in more than one error stratum, if they are not orthogonal to these (e.g. due to missing values, see below). These multiple tests will consume extra Dfs, which will lead to failure if not enough Dfs are available in the corresponding stratum.

Analysis of plot-level data using MMs and ANOVAs (hierarchical models)

Whereas in LMs all VCs are calculated via the estimation of parameters for every level of their explanatory terms, in MMs, the VCs for random-effects terms are directly estimated from the data using ML-based methods. This is computationally much faster because only one instead of many parameters needs to be estimated per term. Of course, subsequent significance testing using ANOVA still has to respect the fact that there are many levels for random-effects terms and therefore Dfs are adjusted accordingly, which is referred to as the REML method (Butler *et al.* 2009). Because MMs estimate the influence of different explanatory terms differently, i.e. via the estimation of multiple parameters (one for each level) for fixed-effects terms and single variance parameters for random-effects terms, they are called thus, i.e. MMs. They have the disadvantage that no contributions to the total SS can be calculated for random-effects terms and thus these cannot be listed in the same way as fixed-effects terms in a comprehensive ANOVA table. However, an advantage of MMs is that the direct estimation of variances for random-effects terms naturally accounts for unequal sample sizes of different groups. In addition, in MMs, the treatment model with the fixed-effects terms can be written separately from the error model with the random-effects terms and thus the sequence of fitting only matters for the fixed-effects terms.

In R, MMs can be fitted using the functions `lme` (library `nlme`) or `lmer` (library `lme4`):

```
lme(ba ~ light + fdiv, random = ~1|com/light)
lmer(ba ~ light + fdiv + (1|com/light)          (MM3)
```

ANOVA tables can be obtained by applying the function `anova`, but in the case of `lmer`, *P* values are only produced if the R-library `lmerTest` has been loaded first (Table 3a). VCs (in `lme` the corresponding standard deviations, i.e. √VC) are obtained by applying the function `summary`. The above MMs fit the same statistical model as LM3, but the results are not exactly the same (compare Tables 2c and d with Table 3a) because of the different ways of parameter estimation. Furthermore, `lme` and `lmer` do not allow VCs < 0, so that their estimates become zero or near-zero. Fortunately, this is not the case in MM3, because no VCs are calculated

Table 3: ANOVA table and variance components (a) and corresponding parameter estimates with standard errors (SE; b) for MM3 in the text, modeling plot-level basal area (`ba`)

a)

Source	SS	MS	Df	denDf	*F*	*P*
light	1186.5	1186.5	1	10	63.22	<0.001
fdiv	1.1	0.6	2	8	0.03	0.970
Variance components						
VC_{com}	75.71					
$VC_{light \times com}$	6.47					
$VC_{res.}$	18.77					

b)

Parameters	Estimate	SE
Intercept (light, monocultures)	19.19	4.56
Shade (shade–light)	−11.33	1.42
Diversity 2 (2-sp. mixtures–monocultures)	1.43	5.82
Diversity 4 (4-sp. mixtures–monocultures)	0.72	10.08

MM3 was fitted with the function `lmer` and parameter estimates were obtained with the function `summary`. The column 'denDf' in a) lists the denominator Df for the *F* value; this can be compared with the Df of the term used as error for the calculation of the corresponding *F* value in ANOVAs using LMs (here Table 2c). See legend to Table 1 for further explanations.

for fixed-effects terms. Nevertheless, this problem requires particular attention, because VCs 'bound' to zero indicate that the variance structure implied by the model does not fit the data, which may lead to too liberal or too conservative significance tests, depending on the hierarchical level considered. Whether the tests are too liberal or too conservative can only be detected if a better model is found, in which none of the VCs is bound yet all fixed-effects terms have corresponding random-effects terms. For example, in the case of LM3, now replaced by MM3, the random-effects term `com` must be included to obtain a correct test for the fixed-effects term `fdiv`, because `fdiv` is a contrast of `com`. Thus, if `com` would be bound, the modeling process would have to be continued until this 'bounding' can be removed. If this is not possible, one could use special functions such as `asreml` (provided by the commercially available ASReml; Gilmour *et al.* 2009), which does not constrain variance components to positive values (see example below and shown in Table 5b). An alternative solution is to use the LM instead of the MM approach.

In contrast, advantages of the MM approach are not only the better handling of unequal group sizes but also that standard errors of parameter estimates for fixed effects (Table 3b) are corrected for the hierarchical structure of the data, specified by the random-effects terms. Furthermore, MMs are generally easier to specify than LMs, but this can sometimes also be a disadvantage because it is also easier to specify an inappropriate MM than an inappropriate LM. Finally, LMs are usually easier to be fitted in R, because they use LS rather than ML/REML methods. As a consequence, we recommend to beginning the analysis of hierarchical data from BEF experiments with LMs and moving to MMs when the relevant explanatory terms contributing to variation in the dependent variable have been identified.

In addition to LS and ML/REML methods, hierarchical models can also be fitted with Bayesian methods (Hector *et al.* 2011; Ogle and Barber 2008; Qian and Shen 2007; Zuur and Ieno 2016). We demonstrate this in the supplementary Appendix S3. Bayesian methods directly estimate the parameters of the fixed-effects terms and the VCs of the random-effects terms together with credibility intervals (CrI) that respect the hierarchical structure of the data, if this is properly specified. The CrIs can be used instead of *P* values to check whether parameter estimates or VCs deviate from zero. In Bayesian analysis, the model is often specified with directed acyclic graphs (Fig. 3a; Clark and Gelfand 2006). Similar graphs can also be applied to develop LMs and MMs (Fig. 3b); and we will provide further examples of this in Fig. 4. Graphical representations help to better understand complex models because they visualize the hierarchical structure of contrasts that can be developed by starting with a complete model that specifies a mean for each individual data point (see below).

FURTHER MODEL DEVELOPMENT

Splitting up variances by making contrasts

The basic concept of ANOVA is the splitting of variances or more exactly SSs. In the example above with the plot-level data of species pool X, the total SS of Df = 87 can be split into an SS of Df = 21 for the 22 combinations of `light × com` and an SS of Df = 66 for the `residuals`. The former can further be split into contrasts in many possible ways, in the most extreme case into 21 contrasts of Df = 1. Well-chosen

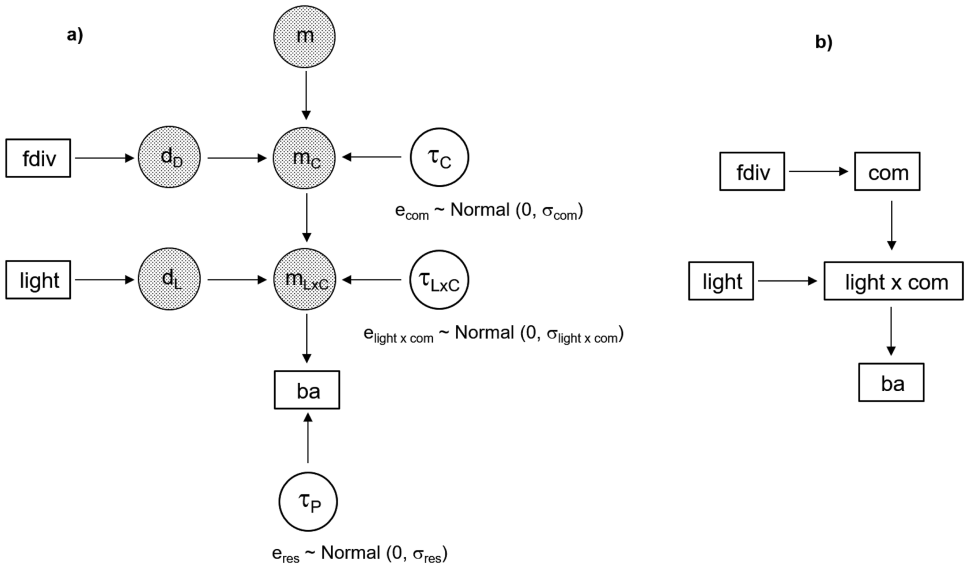

Figure 3: **a**) Directed acyclic graph for the Bayesian approach to analyze the example data of Fig. 2. The horizontal and vertical arrows indicate how different terms (rectangles) and difference (d), mean (m) or precision (τ) parameters (circles) affect each other and the dependent variable `ba`. The 88 values of the dependent variable are predicted from the means (m_{LxC}) of the 2 light treatments · 11 species composition combinations. The m_{LxC} are predicted form the means (m_C) of the 11 species compositions and the effects of the light treatments. The m_C are predicted from the intercept (m, to indicate that this is comparable to the overall mean) and the effects of the diversity treatments. `ba`, m_{LxC} and m_C have associated error variation which is taken from normal distributions with mean zero and the square root of the corresponding VCs, indicated by the subscripted σ values in the equations next to the circles with equally subscripted τ values. These τ values are called precision parameters and are the reciprocals of the VCs. **b**) Simplified graph which shows the same model structure as a). This type of graph is particularly well suited to understand model development with LM approaches such as implemented in the R-functions `aov` and `lm`. Starting with a single explanatory term with a different level for each light treatment-by-species composition combination (`light × com`) one could form contrasts for the `light` and `com` terms and then furthermore a `fdiv`-contrast for the `com` term, in R-syntax, after forming `LxC <- factor(paste(light,com)):ba ~ LxC` ⇔ `ba ~ light + com + LxC` ⇔ `ba ~ light + fdiv + com + LxC` ⇔ LM3 in the text. In these models, the term `fdiv` can not be put after `com` and the terms `light` and `com` can not be put after `LxC` (contrast terms always precede the terms from which they have been carved out).

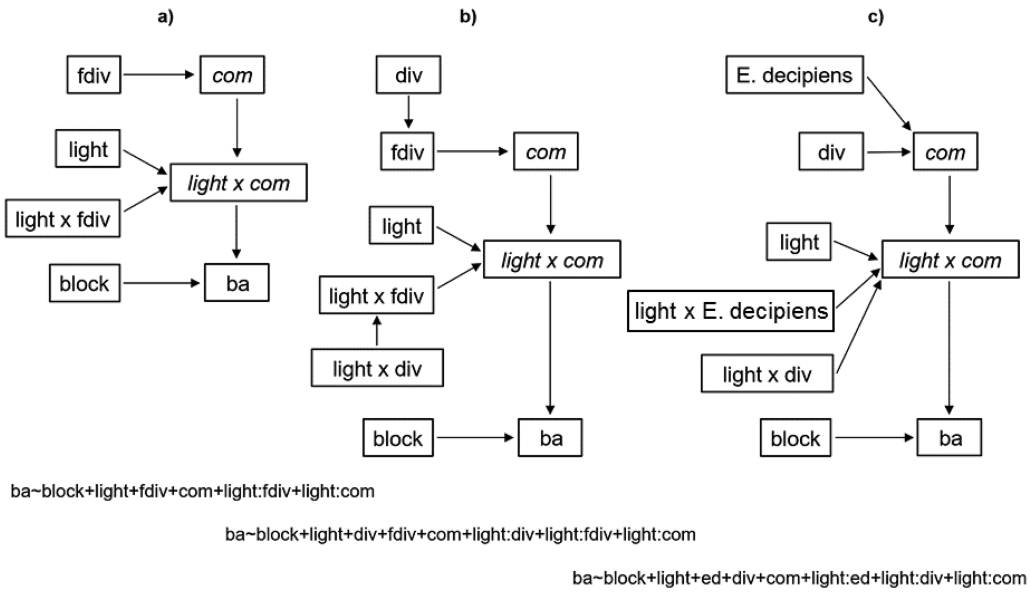

Figure 4: continuing the model development for the example data of Fig. 2 with the approach introduced in Fig. 3b. **a**)–**c**) are models with the same residual variance but show different ways of forming contrasts from the two random terms (in italics). Terms pointing at another term are contrasts of that term and must precede the latter in the model formulae listed below the graphs. Otherwise, terms can appear in different order but if they are correlated, i.e. non-orthogonal, the first term explains both, variation due purely to itself and due to the correlated action of the two terms (i.e. 'ignoring' the second term), whereas the second term only explains variation due purely to itself (i.e. after 'eliminating' the variation due to the first term). See text for further explanations.

contrasts of Df = 1 are particularly useful, because they make focused comparisons and test the most parsimonious hypotheses (Rosenthal and Rosnow 1985).

Once a list of contrasts has been set up and conveniently coded, those that still have a large Df and are considered as random-effects terms can be moved to the error model. The model-building process can thus be highly flexible and include combining SSs again after a splitting process that has gone too far and led to too many terms that explain only little variation (%SS). Typical for BEF experiments is that, in the process of splitting, fixed-effects terms can be carved out as contrasts from random-effects terms, i.e. 'SS-material' can be moved from the error to the treatment model or, in other words, random variation can be moved to systematic variation, which is the main goal of all statistical modeling. Nevertheless, the opposite is also possible, i.e. that in the process of combining SSs, fixed-effects terms are omitted and thus lumped with random-effects terms (including residuals) and moved from the treatment to the error model or, in other words, systematic variation is moved back to random variation.

Typical contrasts of species compositions can address important biological questions in BEF experiments (Fig. 4). Often, blocking terms are also included in these experiments; they generally are unrelated, i.e. fully orthogonal, to the treatment-related explanatory terms and serve only the purpose to reduce the residual variance. In the examples shown in Fig. 4, `block` is fitted first in the models to remove its potential contributions to the variation explained by other terms; a situation that can arise if orthogonality is lost, e.g. due to missing values. The problem of non-orthogonality will be discussed in detail further below. In brief, whenever two fixed-effects terms A and B are fitted in sequence and correlated (non-orthogonal), SS_A will contain contributions from B, whereas SS_B will not contain any contributions from A because these have already been explained. A is thus fitted 'ignoring' B, whereas B is fitted 'eliminating' A (Hector *et al.* 2010; McCullagh and Nelder 1989). Because one is generally not interested in variation among blocks, it is not critical to ignore that `block` may contain variation contributed by terms succeeding it in the fitting process. However, it is crucial that these later terms do not contain contributions of the variation among blocks, therefore this variation is eliminated by fitting `block` first. In MMs, terms related to blocking can be fitted either as fixed- or as random-effects term. We generally prefer to fit them as fixed effects because (i) their number is usually too low to reliably estimate a variance component and (ii) blocks typically do not fulfill the requirement of a random sample with normally distributed effects because they often are systematically arranged in a linear sequence. Indeed, it can even be useful to make contrasts for blocking terms such as linear or quadratic spatial gradients (e.g. see Le Roux *et al.* 2013), especially if these explain a large amount of the variation contained in the blocking terms.

With contrasts, it is critical to consider the order in which terms are being fitted. If the model of Fig. 4c is fitted as LM4 (Table 4a) or MM4 (Table 4b), the linear contrast (`div`) of the diversity term `fdiv` becomes significant after the presence of the species with largest basal area, *Elaeocarpus decipiens* (`ed`), is eliminated. The common slope of the linear diversity effect, now fitted with separate intercepts for plots with and plots without *E. decipiens*, is negative, indicating that a reduction in the density of this species with increasing diversity leads to a reduction in `ba` (Fig. 2). The interactions of the fixed-effects contrasts `ed` and `div`, carved out from `com`, with the treatment `light` are also significant, supporting the expectation that biodiversity effects, including species presence effects, are stronger in light than under shade (see Fig. 2).

Individual-level data and repeated-measures analyses

The typical statistical analysis for BEF experiments uses plot-level data because both biodiversity (B) and ecosystem functioning (EF) are manipulated and measured, respectively, at plot level. Nevertheless, it can be interesting to also analyze measurements obtained for species or even individuals. Individual-level data analysis follows the same principles as plot-level data analysis, but the data structure is more complicated because of the additional hierarchical level. In particular, explanatory terms may now not only vary between but also within plots.

The statistical model for a comprehensive analysis can be written as an error model with error (LM) or random-effects (MM) terms only:

```
y ~ uind * ftime
```

The term `uind` refers to the 257 plots × 16 trees = 4112 distinct individual trees and `ftime` refers to the 17 time points (month since planting) at which the height (`y`) of all tree individuals was measured. Because several trees died during the course of the experiment and trees in block 4 were harvested after 14 months, approximately one-third of the total of 4112 trees × 17 dates (n = 69 904 for the complete design) were missing, a problem to be discussed below (see section 'Individual-level data and non-orthogonality'). Note that the above model treats the 17 levels of the factor `ftime` as a random selection of many possible time points, even though they are in fact ordered and equally spaced. In the following, we first analyze tree height of all individuals over time (shown in Fig. 5) to introduce approaches for repeated-measures analysis, which has to deal with the problem of potential autocorrelation among measures taken at adjacent time points for the same individual. One solution to this problem is to analyze data at a single point in time; this will be done further below to discuss the earlier mentioned problem of non-orthogonality in more detail.

The above error model fits a separate value for every individual observation, leaving no variation to be estimated as `residuals`. Using the principles explained in the previous section, the term `uind` can be partitioned into several terms

Table 4: ANOVA tables for the model shown in Fig. 4c and referred to as LM4 (a) and MM4 (b) in the text, modeling plot-level basal area (`ba`)

a)

Source	Df	SS	MS	*F*	*P*	Error term
block	3	209.66	69.89	4.3	0.008	plot
light	1	2823.64	2823.64	193.1	<0.001	light × com
Elaeocarpus decipiens (= ed)	1	4137.35	4137.35	78.3	<0.001	com
div	1	681.37	681.37	12.9	0.007	com
com	8	422.95	52.87	3.6	0.044	light × com
light × ed	1	189.40	189.40	12.9	0.007	light × com
light × div	1	140.28	140.28	9.6	0.015	light × com
light × com	8	116.97	14.62	0.9	0.526	plot
residuals (= plot)	63	1029.11	16.34			

b)

Source	Df	dendF	F	*P*
Intercept (overall mean)	1	63	343.77	< 0.001
block	3	63	4.33	0.008
light	1	8	174.93	< 0.001
Elaeocarpus decipiens (= ed)	1	8	78.26	< 0.001
div	1	8	12.89	0.007
light × ed	1	8	11.73	0.009
light × div	1	8	8.69	0.018

LM4 was fitted with the function `aov` and MM4 with the function `lme`. For the ANOVA of LM4 appropriate *F*- and *P*-values were calculated using the error terms listed in the last column (cf. Table 2c). See legends to Tables 1 and 3 for further explanations.

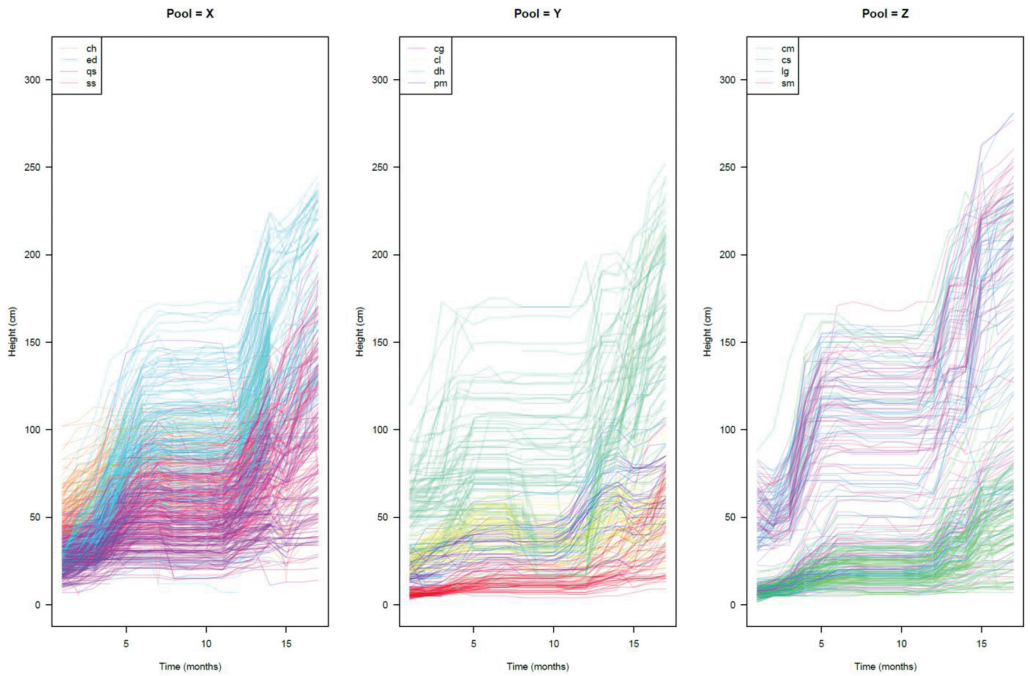

Figure 5: stem height in centimeter for 4089 individuals over the 17 months duration of the experiment. In each pool, there were four species. Several trees died during the experiment, and all trees in block 4 were harvested after month 14. Time 1 is May 2009, and time 17 is September 2010.

to reflect the hierarchical structure of the data set. In condensed form this can be written as:

```
y ~ (com + light:com + plot + uind) * ftime
```

which equals:

```
y ~ com + light:com + plot + uind + ftime + com:ftime
+ light:com:ftime + plot:ftime + uind:ftime
```

Note that in the comprehensive formulation (`y ~ uind * ftime`), `uind` is fitted as the aggregated term, but in the formulations above, `uind` is fitted after the three preceding terms (`com + light:com + plot`) have been carved out from it. Here, `uind` thus only explains the rest of the variation among individuals, when variation due to the preceding terms has already been eliminated.

Excluding the term `uind:ftime` (which corresponds to `residuals`), model MM5 can be fitted with `lmer`:

```
lmer(y ~ (1|com) + (1|light:com) + (1|plot) + (1|uind)
+ (1|ftime) + (1|com:ftime)
+ (1|light:com:ftime) + (1|plot:ftime))          (MM5)
```

Fitting MM5 may take several minutes and applying the methods `summary` or `anova` again may take several minutes. However, fitting this model as LM5 with `aov` is not possible on most currently available desktop or laptop computers, because the more than 4000 levels of `uind` and the over 4000 levels for `plot:ftime` result in a very large model matrix. This example thus illustrates one of the major advantages of MMs over LMs when analyzing hierarchical data.

To develop a treatment model, fixed-effects contrasts can now be carved out in specific ways from MM5 to test hypotheses that are of biological interest. This process, introduced in the previous section, is perhaps the most difficult but also most creative step in the analysis of BEF experiments. Furthermore, there are no easy ways to speed up the process nor is it advisable to use automatic model selection procedures based on purely statistical considerations. Here, we present just one biologically meaningful model (and we encourage the reader to try out further models):

```
lmer(y ~ block + light + div + sp +
light:div + light:sp +
div:sp + time + light:time + div:time + sp:time +
(1|plot) + (1|light:com:sp) +
(time|com) + (time|light:com) + (time|com:sp))
                                                (MM6)
```

In the ANOVA for MM6 (Table 5a), `time` is a linear contrast of the factor `ftime` and can be used to test whether the height growth of plants over the 17 months of the experiment had a significant linear slope, which seems to be approximately the case according to Fig. 5. The term `sp` refers to the 12 species of tree individuals and is a fixed contrast carved out from `uind`. However, there is a slight complication, because, for monoculture plots, all plants belong to the same species, and in this case, `sp` could also be considered as a fixed-effects contrast carved out from `com`. This problem could be solved making two separate contrasts for species, one called `monosp` (referring to species in monocultures) and one called `mixsp` (referring to species in mixtures). However, then the interaction `div × mixsp` only refers to differences between 2- and 4-species communities, because `monosp` fully explains compositional differences among monocultures. Nevertheless, MM6 is not perfect because species effects very likely contain 'contaminations' of σ^2_{com} and σ^2_{plot}. Therefore, tests for the species term and its interactions will tend to be too liberal and therefore should be interpreted with caution.

Unfortunately, MM6 is difficult to fit with `lmer` (Table 5a). Not only does it take very long to run but `lmer` also does not provide the possibility to keep the order of fixed-effect terms as specified but moves interactions after main effects, even if these are unrelated. Furthermore, `lmer` constrains two variance components in MM6 to zero, which as mentioned earlier can be problematic. We therefore present in Table 5b the ANOVA obtained for MM6 using the function `asreml.nvc` (using the commercially available software ASReml for R; Gilmour *et al.* 2009), which allows to keep the order of terms as specified and variance components to be negative. With this analysis, the tests for the fixed-effects terms `light` and `div` are now significant, because they can be compared with the appropriate random-effects terms `light × com` and `com`, respectively. This reflects an important consideration that must be made in all MMs: for every fixed-effects term (first three rows of the formula above) appropriate random-effects terms (last two rows of the formula above) must be specified and estimable to define the correct error strata. This matching between fixed- and random-effects terms is indicated in Table 5b (see also Table 4a). It should be noted, however, that sometimes a fixed-effects term can be considered as a contrast carved out from several random-effects terms, i.e. a linear combination of these, as, e.g. explained in the previous paragraph in the case of `sp`. Furthermore, note that in the model formula above, `(time|com)` represents the terms `com` and `com × time`, `(time|light:com)` represents the terms `light × com` and `light × com × time` and `(time|com:sp)` represents the terms `com × sp` and `com × sp × time` in Table 5b.

From the ANOVA in Table 5b, we can conclude that individual plant height within plots in the example was significantly affected by the plot-level explanatory terms `light` and linear species richness (`div`) and by the individual-level explanatory terms species (`sp`) and its interactions with `light` (`light × sp`) and `div` (`div × sp`). Because all these between-individual differences can be compared with random-effects terms that do not involve time (see Table 5b), the corresponding tests are not affected by potential autocorrelations within individuals over time. Furthermore, carving out the linear contrast `time` from `ftime` allows comparing linear slopes of height growth with appropriate random-effects terms that do not contain `ftime` and are thus still unaffected by potential autocorrelations (Table 5b). The biological interpretations here are that

Table 5: ANOVA tables for MM6 in the text, modeling individual-level height over time

a) `lmer`

Source	SS	MS	Df	denDf	*F*	*P*
block	19 017	6339	3	228.7	14.4	<0.001
light	1060	1060	1	79.9	2.4	0.125
div	443	443	1	78.6	1.0	0.320
sp	12 85 714	116 883	11	56.5	264.8	<0.001
time	732 090	73 2090	1	72.0	1658.6	<0.001
light × div	379	379	1	246.2	0.9	0.355
light × sp	93 917	8538	11	70.7	19.3	<0.001
div × sp	9931	903	11	102.7	2.0	0.031
light × time	13 739	13 739	1	50.7	31.1	<0.001
div × time	0	0	1	73.2	0.0	0.991
sp × time	476 704	43 337	11	47.8	98.2	<0.001
Variance components						
VC_{com}	0.00					
$VC_{light \times com}$	4.52					
VC_{plot}	55.60					
$VC_{com \times sp}$	0.57					
$VC_{light \times com \times sp}$	9.49					
$VC_{com \times time}$	0.00					
$VC_{light \times com \times time}$	0.32					
$VC_{com \times sp \times time}$	0.17					
$VC_{res.}$	441.00					

b) `asreml.nvc`

Source	Df	denDf	F	*P*	Corresponding random effects
block	3	192.6	2.1	0.108	plot
light	1	15.7	33.9	<0.001	light × com
div	1	9.4	100.1	<0.001	com
sp	11	30.6	330.3	<0.001	com; com × sp
light × div	1	6.9	0.6	0.463	light × com
light × sp	11	40.3	9.3	<0.001	light × com; light×com×sp
div × sp	11	32.8	4.1	<0.001	com × sp
time	1	18	4761	<0.001	com × time
light × time	1	32.4	23.6	<0.001	light ×com×time
div × time	1	7.4	1.1	0.325	com × time
sp × time	11	44.0	123.7	<0.001	com × time; com×sp×time
Variance components					
VC_{com}			1.39		
$VC_{light \times com}$			−17.35		
VC_{plot}			61.17		
$VC_{com \times sp}$			−8.16		
$VC_{light \times com \times sp}$			21.75		
$VC_{com \times time}$			−0.23		
$VC_{light \times com \times time}$			0.40		
$VC_{com \times sp \times time}$			0.20		
$VC_{res.}$			441.42		

MM6 was fitted with the function `lmer` (a) or the function `asreml.nvc` (R-script, Supplementary Data). Note that in a) VCs are constrained to zero, whereas they are allowed to be negative in b). The last column in b) indicates from which random-effects terms the fixed-effects terms have been carved out. See legends to Tables 1 and 3 for further explanations.

individuals show significant linear height growth (`time`) but in addition that linear height growth of individuals is significantly affected by `light` (`light × time`) and differs among species (`spec × time`).

Using linear or polynomial contrasts of time is a powerful approach to the problem of repeated measures with potential autocorrelations over time. An alternative would be a two-stage approach where in a first step a linear slope of the regression of height against time would be calculated for each individual. In a second step, the individual slopes could then be analyzed as metadata using LMs and MMs, no longer involving time. Also, nonlinear growth functions could be used (e.g. logistic growth), which is difficult otherwise. We show such an analysis in supplementary Appendix S4. To get a closer insight into patterns of systematic variation in repeated-measures data, however, it is often necessary to analyze the data of the different measurement times separately, which is demonstrated in the next section.

Individual-level data and non-orthogonality

Focusing on a single time point, in the present case, individual-level plant height 14 months after planting, the previous model MM6 can be more easily expanded to ask additional biological questions, in particular whether contrasts for the plot-level presence of the dominant species *Eleocarpus decipiens* (`ed`), *Dalbergia hupeana* (`dh`) and *Sapindus mukorossi* (`sm`) could explain part of the large variation in individual-level height found among species compositions (`com`) and whether light-induced differences in linear species richness effects vary among species (three-way interaction `light × div × sp`). As recommended earlier, we here use LMs to obtain ANOVAs, because it allows more flexible model fitting than would be possible with the use of MMs, with preservation of fitting sequences of fixed-effects terms (option `keep.order=TRUE` of the function `terms`) and with comparison of these with error terms (random-effects terms in MMs) at the same 'currency', i.e. SSs. Using different fitting sequences shows the effects of non-orthogonality in data from BEF experiments and how this relates to tests of relevant biological hypotheses. The following models

```
aov(terms(y ~ block+light+ed+dh+sm+div+sp+
light:ed+light:dh+light:sm+light:div+
light:sp+div:sp+light:div:sp+
com+light:com+plot+com:sp+light:com:sp,
keep.order = TRUE))
```

(LM7)

and

```
aov(terms (y ~ block+light+div+ed+dh+sm+sp+
light:div+light:ed+light:dh+light:sm+
light:sp+div:sp+light:div:sp+
com+light:com+plot+com:sp+light:com:sp,
keep.order = TRUE))
```

(LM8)

have the same residuals but the Dfs and SSs of other terms vary with their position in the model and ANOVA table (supplementary Table S5a and b). The presence of any of the three dominant species in a plot has a very strong effect on individual-level tree height, which is not surprising, because these dominant species are included among those trees and other species may also grow taller in competition with them to reach the light. However, the three-way interaction `light × div × sp` is not significant, i.e. light-induced differences in linear species richness effects do not seem to vary among species.

The different fitting sequences in LM7 and LM8 lead to different results because the terms involved are not orthogonal, i.e. not independent. In the present case, the non-orthogonality is inherent to the design, because it is not possible, e.g. to have plots with four species of pool X but to exclude *E. decipiens*. Additionally, non-orthogonality occurs because of missing data, which are not distributed proportionally across the treatment combinations. In general, covariates also lead to non-orthogonality because these are almost never orthogonal to the other explanatory terms in a model.

Although the ANOVA tables derived from LM7 and LM8 use *F*- and *P* values based on error terms calculated with the LS approach and may therefore not be as perfect as those derived from MMs using the REML approaches, they give a good indication for the importance of the different explanatory terms, as sequentially fitted, because effect sizes of all terms can be inspected from the %SSs. These measure by how much the overall model fit, i.e. the multiple R^2, is improved when a term is added.

Effect sizes can also be shown by predicting data based on a model fit. Many R classes provide a `predict` method that can be used for this purpose. However, to interpret the effect size of a term according to its position in the model, it is safest to fit a simplified model that includes all terms up to the one of interest and then looking at the parameter estimate of this term (Payne *et al.* 1993). Thus, for the term `div`:

```
summary.lm(aov(y ~ block+light+ed+dh+sm+div))
```

yields an estimate of −11.8 cm height change per additional species in a plot. This is because, for plots with the dominant species, increasing species richness means fewer individuals of the dominant species per plot. If the term `div` is fitted before the dominant species, its effect is much smaller (supplementary Table S5b), but

```
summary.lm(aov(y ~ block+light+div))
```

yields an estimate of 1.64 cm height increase per additional species in a plot. This indicates that ignoring the presence of dominant species in a plot results in a slightly positive effect of diversity. Technically, these estimates for `div` are equivalent to the corresponding hypotheses tests in the so-called type-I (sequential) ANOVA, the type of ANOVA used throughout this article (Venables 2000).

The above is not true if parameters are estimated after all terms have been added to a model, including ones that appear after a term of interest in the ANOVA. Then this term of interest is also adjusted for the terms following it in the fitting sequence. In other words, all terms are adjusted for all others in the model and all information about the sequence of fitting is lost. Such parameter estimates do no longer correspond to the biological hypotheses tested by *F*- and *P* values in the ANOVA table and can thus be very misleading. It may even be that the direction of an effect is changed, as would occur for `div` if, in the second model above, the contrasts for the three dominant species would be added after `div`. Parameter estimates that are obtained after all terms have been added to a model correspond to hypotheses tests in the so-called type-III (nonsequential) ANOVAs. Type-III ANOVAs are assembled from as many analyses as necessary to allow every term to appear last in sequence once and then extracting the corresponding SS (Driscoll and Borror 2000). This can lead to awkward hypotheses that have little biological meaning (Langsrud 2003; Venables 2000). This is especially the case if non-orthogonality is strong; for orthogonal data, type-I and type-III ANOVAs are the same.

We conclude our analyses here, emphasizing that with data from BEF experiments, such as the one presented here, there are a very large number of further modeling possibilities and ramifications, which should be broadly explored before drawing conclusions about relevant biological hypotheses. A key tool hereby is the decomposition of explanatory terms into contrasts. In the most extreme case, for an explanatory term with k levels, $k - 1$ single-Df contrasts can be formed in many different ways, like k twigs on a tree can be connected by $k - 1$ single branching events in many different ways. The goal of the analysis is to find contrasts that address interesting biological questions. These might be linear contrasts for diversity or time, contrasts between communities with or without particular species or contrasts between species belonging to particular functional groups or sharing specific phylogenetic relationships. Due to these many potential ways of analysis, the ANOVA for complex experimental designs is as much an exploratory tool as a confirmatory one. Nevertheless, compared with data from observational studies, data from BEF experiments have the advantage that they can be analyzed with *a priori* hypotheses about cause–effect relationships, because the explanatory terms have been manipulated by the experimenter and can be treated as fixed-effects terms. This also means that model comparisons can be based on biological judgment and do not require tools such as AIC-type information criteria (Burnham and Anderson 2002). We conclude this section with 'Question and Answers' presented in Table 6.

DISCUSSION

General recommendations

The aim of our presentation above is to demonstrate an ecologists' approach to analyze a BEF experiment with a typically complex hierarchical data structure. BEF experiments search for general effects of biodiversity on ecosystem functioning, but at the same time, they acknowledge that these general effects can only occur because different species (or functional groups, genotypes etc.) have more or less similar main and interactive effects on dependent variables. Thus, BEF experiments must include $k >> 1$ different species compositions (genotype compositions etc.) as treatments, which automatically leads to the possibility to test $k - 1$ orthogonal hypotheses in a single analysis and many more in different analyses. If in addition, modifying influences—in the example different levels of light availability or different time points—need to be tested, then the number of possible hypotheses increases even further.

Conceptually, the complexity of BEF experiments can be approached from two sides, either starting with tests for the most important fixed-effects terms and ignoring everything else (approach 1, exemplified by LM1) or starting with the most inclusive random-effects terms that give a predicted value for every observation (approach 2, exemplified by the equation `y ~ uind * ftime` used before fitting MM5). Also, different subsets of the full data set could be explored separately, but at a later stage these would have to be combined again. Otherwise, not all hypotheses of interest can be tested, and tests obtained with the different subsets may be interdependent, e.g. if separate analyses are done for different time points. We recommend that all these approaches should be followed to develop a good understanding of systematic and hierarchical random variation in the data. In addition, it is always helpful to tabulate and plot data according to the different groupings given by the levels of explanatory terms, together with the fitting of statistical models and production of ANOVA tables.

For both, approaches 1 and 2, it is convenient to use graphs of the type shown in Figs 3 and 4 to move between models. With approach 1, new boxes are added to a graph, which are ignored (i.e. included in the residual random variation) in previous versions of the graph. With approach 2, terms in new boxes are contrasts of terms in old boxes of previous versions of the graph. Making contrasts is perhaps the most powerful aspect of ANOVA, because it allows the testing of focused biological hypotheses (Rosenthal and Rosnow 1985). Typical contrasts in BEF experiments are those between monocultures and mixtures, linear or log-linear species richness or the presence of particular species or species combinations.

Presenting results

Once a good model has been found, test results should always be shown together with effect sizes and the direction of effects in case of single-Df tests. One possible measure of effect size are %SSs, which can be easily added to ANOVA tables. One must, however, keep in mind that the amount of variance a term explains also contains a fraction that is explained by chance (VC(s) not related to the term, see EMS in Table 1). To

Table 6: Frequently asked questions by ecologists analyzing BEF experiments and suggested answers

What are the advantages of LM- and MM-based ANOVA, respectively, to analyze complex hierarchical data from BEF experiments? LM-based ANOVAs are flexible and compare the contributions of fixed-effects terms to systematic variation and the contributions of error terms to random variation in the dependent variable at the same scale, using SSs. Negative VCs are not constrained to zero. Unless error terms have very large numbers of levels, LMs can be fitted very efficiently. MM-based mixed models estimate variance components more efficiently and provide direct—and in the case of unbalanced data sets with non-orthogonal terms more accurate—tests for fixed-effects terms. However, MMs sometimes may not converge and if VCs are constrained to zero, test accuracy is reduced.
How can contrasts be used to test focused biological hypotheses? For a factor *F* with ordered levels, linear, log-linear or polynomial contrasts can be specified by using a continuous variable x, log(x), $x + x^2 + \ldots$ To compare particular treatment levels with others, e.g. monocultures with mixtures, contrasts can be specified giving each group of treatments a different label. Sub-contrasts must be fitted after contrasts, e.g. x (species richness) after monocultures vs. mixtures will test for a linear effect of species richness within mixtures only.
Which random-effects terms must be included in an MM to obtain correct tests for fixed-effects terms? All terms that define error strata and thus replication levels of treatments must be specified as random-effects terms in MM. This is particularly relevant when interactions of fixed-effects terms are being tested. For example, whenever a fixed-effects term F1 (e.g. `div`) requires a random-effects term R1 (e.g. `com`) then a fixed-effects interaction F1 × F2 (e.g. `div:light`) requires the interaction R1 × F2 (e.g. `com:light`) as random-effects term.
Does the sequence of terms matter in ANOVA? Yes, for fixed-effects terms; during construction of the ANOVA table, terms are added in sequence and tests are constructed by comparing nested models. In that sense, terms fitted earlier in the model eliminate variation that could be explained by later terms. For random-effects terms in MMs the sequence does not matter, but in LMs, error terms must follow fixed-effects terms that have been carved out from them and the sequence of error terms must respect the hierarchical data structure (from higher to lower error strata).
What can be done if *F* values are <1 and variance components <0? Check the statistical model for omitted terms that may affect the denominator but not the nominator term of the *F* value. Often this occurs with main-plot terms when split-plot terms are omitted. In other cases, it occurs when an error terms is affected by negative autocorrelation (e.g. due to competition). In such cases, negative variance components should be allowed to provide appropriate error terms. Alternatively, autocorrelations and other variance structures can be specified in a statistical model (not discussed in this article).
Should the statistical model follow the design or be selected according to model-selection criteria such as AIC or BIC? In experimental studies a statistical model should include terms for all *a priori* hypotheses for which an experimental manipulation was made (e.g. `light`, `div`, `com`). Terms with small *F* values can be pooled or included in error (LMs) or random-effects (MMs) term from which they have been carved out. It is better to work with *F* values of individual terms (see e.g. Murtaugh 2014) then with criteria assessing the whole model (R^2, AIC, BIC). The latter are more suitable for observational studies (Burnham and Anderson 2002).
Can parameter estimates and their standard errors be used to test the significance of explanatory terms in ANOVA? With balanced data and orthogonal terms this can be done for single-Df terms (then the *t*-value derived from the standard error of the estimate is the square root of the *F* value of the term). In other cases, it is not recommended. Parameter estimates are always calculated for the full model (i.e. eliminating all other terms) and thus may not reflect the sequential fitting sequence of terms used to address specific biological hypotheses in an ANOVA.
When and how should covariates be included in a model and what are their effects? In experimental studies, statistical analysis should begin without covariates. Usually, covariates destroy balance and orthogonality of experimental designs. They also change the interpretation of effects, which will be adjusted for the covariates. If covariates are introduced later in analysis, they may explain effects, e.g. by changing a significant effect into an insignificant one.
Can subsets of data be analyzed separately and what are the consequences? It is often convenient to analyze only particular sections of a data set, e.g. data from a single time point or for a particular species set. However, with multiple separate analyses, the problem of multiple testing arises. Therefore, a comprehensive analysis should always be carried out as well. Furthermore, if a term is significant in one but not in another analysis for two subsets of the same data, this does not mean that the effects of the term differ significantly between the two subsets. This can only be tested fitting a corresponding interaction between this term and the term defining the subsets in the comprehensive analysis.

LM-based ANOVA refers to approaches used in the R functions `aov` and `lm`, i.e. linear models. MM-based ANOVA refers to approaches used in the R functions `lme`, `lmer` and `asreml`, i.e. mixed models.

find out the direction of an effect, it is necessary that the corresponding parameter estimate is obtained, which as explained in the section on non-orthogonality can best be done by fitting a model with all terms up to the one of interest as the last one and then look at its estimate. With this approach, the biological explanation is that an effect has been adjusted for all the terms preceding it in the analysis but not the ones that follow.

Because effect sizes and their significances in complex analyses depend on the particular model, it is also important to use appropriate graphical representations of data. It is often useful to show individual raw data in scatter plots because these do not depend on the statistical model used for analysis. Furthermore, different subsets of the raw data can be shown in separate panels or symbols can be used for individual raw data belonging to different levels of explanatory factors (e.g. see Figs 2 and 5), still avoiding any dependence from statistical models. However, beyond this, dependency cannot be avoided. Therefore, every figure showing means, regression lines, standard errors etc. should specify clearly from which model these parameter estimates are taken. Particular care is required with parameter estimates after a hierarchical data analysis has been carried out. In this case, the estimates of fixed effects must be adjusted for the random effects and thus can be different from the estimates obtained with nonhierarchical data analysis or using original data.

SUPPLEMENTARY DATA

Supplementary material is available at *Journal of Plant Ecology* online.

FUNDING

This study was supported by the Swiss National Science Foundation (grant number 310030B_147092 to B.S.) and the University Research Priority Program Global Change and Biodiversity of the University of Zürich.

ACKNOWLEDGEMENTS

We acknowledge the very helpful comments of two anonymous reviewers on an earlier version of this article.
Conflict of interest statement. None declared.

REFERENCES

Balvanera P, Pfisterer AB, Buchmann N, *et al.* (2006) Quantifying the evidence for biodiversity effects on ecosystem functioning and services. *Ecol Lett* **9**:1146–56.

Bruelheide H, Böhnke M, Both S, *et al.* (2011) Community assembly during secondary forest succession in a Chinese subtropical forest. *Ecol Monogr* **81**:25–41.

Bruelheide H, Nadrowski K, Assmann T, *et al.* (2014) Designing forest biodiversity experiments: general considerations illustrated by a new large experiment in subtropical China. *Methods Ecol Evol* **5**:74–89.

Bu WS, Schmid B, Liu XJ *et al.* (2017). Interspecific and intraspecific variation in specific root length drives aboveground biodiversity effects in young experimental forest stands. *J Plant Ecol* **10**:158–69.

Burnham KP, Anderson DR (2002) *Model Selection and Multimodel Inference*. New York, NY: Springer-Verlag.

Butler DG, Cullis BR, Gilmour AR, *et al.* (2009) *ASReml-R Reference Manual*. Queensland, Australia: Queensland Government, Department of Primary Industries and Fisheries.

Clark JS, Gelfand AE (2006) A future for models and data in environmental science. *Trends Ecol Evol* **21**:375–80.

Cohen J (2002) *Statistical Power Analysis for the Behavioral Sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Dimitrakopoulos PG, Schmid B (2004) Positive biodiversity effects increase linearly with biotope space. *Ecol Lett* **7**:574–83.

Driscoll MF, Boror CM (2000) Sums of squares and expected means squares in SAS. *Qual Reliab Engng Int* **16**:423–33.

Ebeling A, Pompe S, Baade J, *et al.* (2014) A trait-based experimental approach to understand the mechanisms underlying biodiversity–ecosystem functioning relationships. *Basic Appl Ecol* **15**: 229–40.

Gelman A (2005) Analysis of variance—why it is more important than ever. *Ann Stat* **33**:1–53.

Gilmour AR, Gogel BJ, Cullis BR, *et al.* (2009) *ASReml User Guide Release 3.0.* www.vsni.co.uk (10 October 2016, date last accessed).

Green BF, Tukey JW (1960) Complex analysis of variance: general problems. *Psychometrika* **25**:127–52.

Hahn CZ, Niklaus PA, Bruelheide H, *et al.* (2017). Opposing intraspecific vs. interspecific diversity effects on herbivory and growth in subtropical experimental tree assemblages. *J Plant Ecol* **10**:242–51.

Hector A, Bell T, Hautier Y, *et al.* (2011) BUGS in the analysis of biodiversity experiments: species richness and composition are of similar importance for grassland productivity. *PLOS ONE* **6**:e17434.

Hector A, Loreau M, Schmid B; BIODEPTH Project (2002) Biodiversity and the functioning of grassland ecosystems: multi-site comparisons. In Kinzig AP, Pacala SW, Tilman D (eds). *Functional Consequences of Biodiversity: Empirical Progress and Theoretical Extensions*. Princeton, NJ: Princeton University Press, 71–95.

Hector A, Schmid B, Beierkuhnlein C, *et al.* (1999) Plant diversity and productivity experiments in European grasslands. *Science* **286**:1123–27.

Hector A, von Felten S, Schmid B (2010) Analysis of variance with unbalanced data: an update for ecology & evolution. *J Anim Ecol* **79**:308–16.

Kirwan L, Connolly J, Finn JA, *et al.* (2009) Diversity-interaction modeling: estimating contributions of species identities and interactions to ecosystem function. *Ecology* **90**:2032–8.

Langsrud O (2003) ANOVA for unbalanced data: use type II instead of type III sum of squares. *Stat & Comp* **13**:163–7.

Le Roux X, Schmid B, Poly F, *et al.* (2013) Soil environmental conditions and microbial build-up mediate the effect of plant diversity on soil nitrifying and denitrifying enzyme activfities in temperate grasslands. *PLOS ONE* **8**:e61069.

Li SS, Tong YW, Wang ZW (2017). Species and genetic diversity affect leaf litter decomposition in subtropical broadleaved forest in southern China. *J Plant Ecol* **10**:232–41.

McCullagh P, Nelder JA (1989) *Generalized Linear Models*. London, UK: Chapman and Hall.

Murtaugh PA (2014) In defence of *P* values. *Ecology* **95**:611–7.

Naeem S, Thompson LJ, Lawler SP, *et al.* (1994) Declining biodiversity can alter the performance of ecosystems. *Nature* **368**:734–7.

Nelder JA (1977) A reformulation of linear models. *J R Statist Soc A* **140**:48–63.

Nelder JA (1994) The statistics of linear-models: back to basics. *Stat & Comp* **4**:221–34.

Nelder JA, Lane PW (1995) The computer analysis of factorial experiments: in memoriam—Frank Yates. *Amer Statist* **49**:382–5.

Niklaus PA, Leadley PW, Schmid B, *et al.* (2001) A long-term field study on biodiversity x elevated CO_2 interactions in grassland. *Ecol Monogr* **71**:341–56.

Ogle K, Barber JJ (2008) Bayesian data–model integration in plant physiology and ecosystem ecology. *Prog Bot* **69**:281–311.

Payne RW, Lane PW, Digby PGN, *et al.* (1993) *Genstat 5, Release 3 Reference Manual*. Oxford: Clarendon Press.

Peng SY, Schmid B, Niklaus PA (2017). Leaf area increases with species richness in young experimental stands of subtropical trees. *J Plant Ecol* **10**:128–35.

Qian AA, Shen Z (2007) Ecological applications of mutlilevel analysis of variance. *Ecology* **88**:2489–95.

Rockström J, Steffen W, Noone K, *et al.* (2009) A safe operating space for humanity. *Nature* **461**:472–5.

Roscher C, Schumacher J, Baade J, *et al.* (2004) The role of biodiversity for element cycling and trophic interactions: an experimental approach in a grassland community. *Basic Appl Ecol* **5**:107–21.

Rosenthal R, Rosnow RL (1985) *Contrast Analysis: Focused Comparisons in the Analysis of Variance*. Cambridge, UK: Cambridge University Press,.

Schläpfer F, Pfisterer BA, Schmid B (2005) Non-random species extinction and plant production: implications for ecosystem functioning. *J Appl Ecol* **42**: 13–24.

Schmid B, Hector A (2004) The value of biodiversity experiments. *Basic Appl Ecol* **5**:535–42.

Schmid B, Hector A, Huston M, *et al*. (2002) The design and analysis of biodiversity experiments. In Loreau M, Naeem S, Inchausti P (eds). *Biodiversity and Ecosystem Functioning: Synthesis and Perspectives*. Oxford: Oxford University Press, 61–75.

Schmitz M, Flynn DBF, Mwangi PN, *et al*. (2013) Consistent effects of biodiversity on ecosystem functioning under varying density and evenness. *Folia Geobotanica* **48**:335–53.

Searle SR (1971) *Linear Models*. New York, NY: John Wiley & Sons.

Snedecor GW, Cochran WG (1989) *Statistical Methods*. Ames, IA: Iowa State University Press.

Sun ZK, Liu XJ, Schmid B, *et al.* (2017). Positive effects of tree species richness on fine-root production in a subtropical forest in SE-China. *J Plant Ecol* **10**:146–57.

Tilman D, Wedin D, Knops J (1996) Productivity and sustainability influenced by biodiversity in grassland ecosystems. *Nature* **379**:718–20.

Van Kleunen M, Stephan MA, Schmid B (2006) [CO_2]- and density-dependent competition between grassland species. *Global Change Biol* **12**:2175–86.

Venables, WN (2000) Exegeses on linear models. In: *Paper presented at the SPlus User's Conference, Washington DC, 8–9 October 1998*.

Wilkinson GN, Rogers CE (1973) Symbolic description of factorial models for analysis of variance. *Appl Stat* **22**:392–9.

Zeng XQ, Durka W, Fischer M (2017). Species-specific effects of genetic diversity and species diversity of experimental communities on early tree performance. *J Plant Ecol* **10**: 252–8.

Zuur AF and Ieno EN (2016) A protocol for for data exploration to avoid common statistical problems. *Methods Ecol Evol* **1**:3–14.

Zuur AF, Ieno EN, Elphick CS (2010) A protocol for conducting and presenting results of regression-type analyses. *Methods Ecol Evol* **7**:636–45.