Received: 14 August 2016 | Accepted: 16 February 2018

DOI: 10.1111/2041-210X.13164

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Comparative analysis of experimental data

Robert P. Freckleton | Mark Rees

Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK

Correspondence
Robert P. Freckleton
Email: r.freckleton@sheffield.ac.uk

Handling Editor: David Warton

Abstract

1. We consider the problem of how to analyse data from experiments conducted on multiple species. This seems to have been largely overlooked in the literature, and we highlight that the use of species as experimental units creates issues for both the design and analysis of experiments.
2. We distinguish fully randomized experiments in which all treatments are applied to all species from those experiments in which the factor of interest varies at the species level, that is, treatments are not randomly allocated to species. In this latter case, the distribution of the experimental factor across species may be random, phylogenetically structured or species may be chosen in order to phylogenetically balance the sample (e.g. through sister-species comparisons).
3. We show using simulations that the design of the experiment in terms of the phylogenetic distribution of the treatment can affect power and Type I error, and that commonly used approaches (Linear Mixed Models and ANOVAs) may have poor statistical properties (high Type I errors) when both the predictors and response data show strong phylogenetic signal.
4. We highlight that the true phylogenetic generalized least squares model yield has good statistical properties but show that in some cases the true variance structure may be difficult to identify empirically. Moreover, many current comparative tools do not allow such analyses to be easily applied, and we highlight some of those that do.

KEYWORDS

comparative analysis, evolutionary biology, experimental data, linear models, macroevolution, phylogenetics, statistics, type I errors

1 | INTRODUCTION

The comparative method is among the most widely used approaches for addressing questions in ecology and evolutionary biology (Freckleton & Pagel, 2010; Harvey & Pagel, 1991; Nunn 2011). The rationale of the comparative method is that a group of species will contain more variation than a single species or that it is possible to create using experimental manipulation (Maynard Smith, 1978). Consequently, comparative methods can be used to test extremely broad hypotheses. Moreover, comparative methods typically use data collated from the literature and are therefore extremely efficient in terms of time, expense of data collection and reuse.

The potential problems with comparative approaches are well known, and result from the statistical and evolutionary non-independence of species (Harvey, 1996; Harvey & Pagel, 1991). Evolutionary non-independence of species results in similarities within assemblages that are the result of common ancestry rather than independent evolution. As a consequence, comparative data cannot be safely regarded as being statistically independent, and a suite of approaches have been developed to analyse comparative data while incorporating the phylogenetic relationships between the species, and these are now routinely used (e.g. Grafen, 1989; Martins & Garland (1991); Hadfield & Nakagawa, 2010; Martins & Hansen, 1996; Pagel, 1997, 1999).

Conceptually, the opposite of the comparative method is the experimental approach (Maynard Smith, 1978). Comparative methods are typically observational, and rely on uncovering correlations, with possible confounding effects eliminated statistically. A correlative analysis can never eliminate all confounding effects, nor can causation be distinguished from correlation on purely statistical grounds. On the other hand, experiments are intended to manipulate only the factors of interest, with nuisance and confounding effects eliminated by design and randomization (Mead 1988). The randomized block experiment is, for example, described as the 'gold standard' in testing ecological hypotheses (e.g. Newman, Bergelson, & Grafen, 1997).

Experimental and comparative approaches need not be completely divorced, however, and the dichotomy between the two becomes blurred in experiments that use multiple species. Comparative experiments have always been common, and classic examples include mass screening experiments on plants (e.g. Grime, Hodgson, & Hunt, 1990). In principle, experiments should allow the 'species' effect to be accounted for through appropriate design and treatments randomized across species. Consequently, it might be expected that experimental comparative approaches should be less prone to confounding than purely correlative studies.

Unfortunately this is not likely to be a general or safe assumption to make. The reason for this is illustrated in Figure 1, which highlights four designs that are commonly used. The first approach (Figure 1a) is a factorial experiment, in which all species receive each of the treatments. Species are effectively treated as randomized blocking factors. This design would be used when it is possible to manipulate the factor of interest (e.g. when applying different nutrient regimes to plants in growth experiments).

Frequently, however, the factors analysed in an experiment are properties of the species themselves. For instance, in an example, we discuss below plant species may be characterized as possessing one of two photosynthetic pathways (C3 vs. C4). Each species is in only of one of two states, and consequently this factor cannot be randomly applied to species. In such situations, the design of the experiment will depend on how species are chosen and/or how traits are distributed across the phylogeny. Figure 1b–d illustrate three idealized

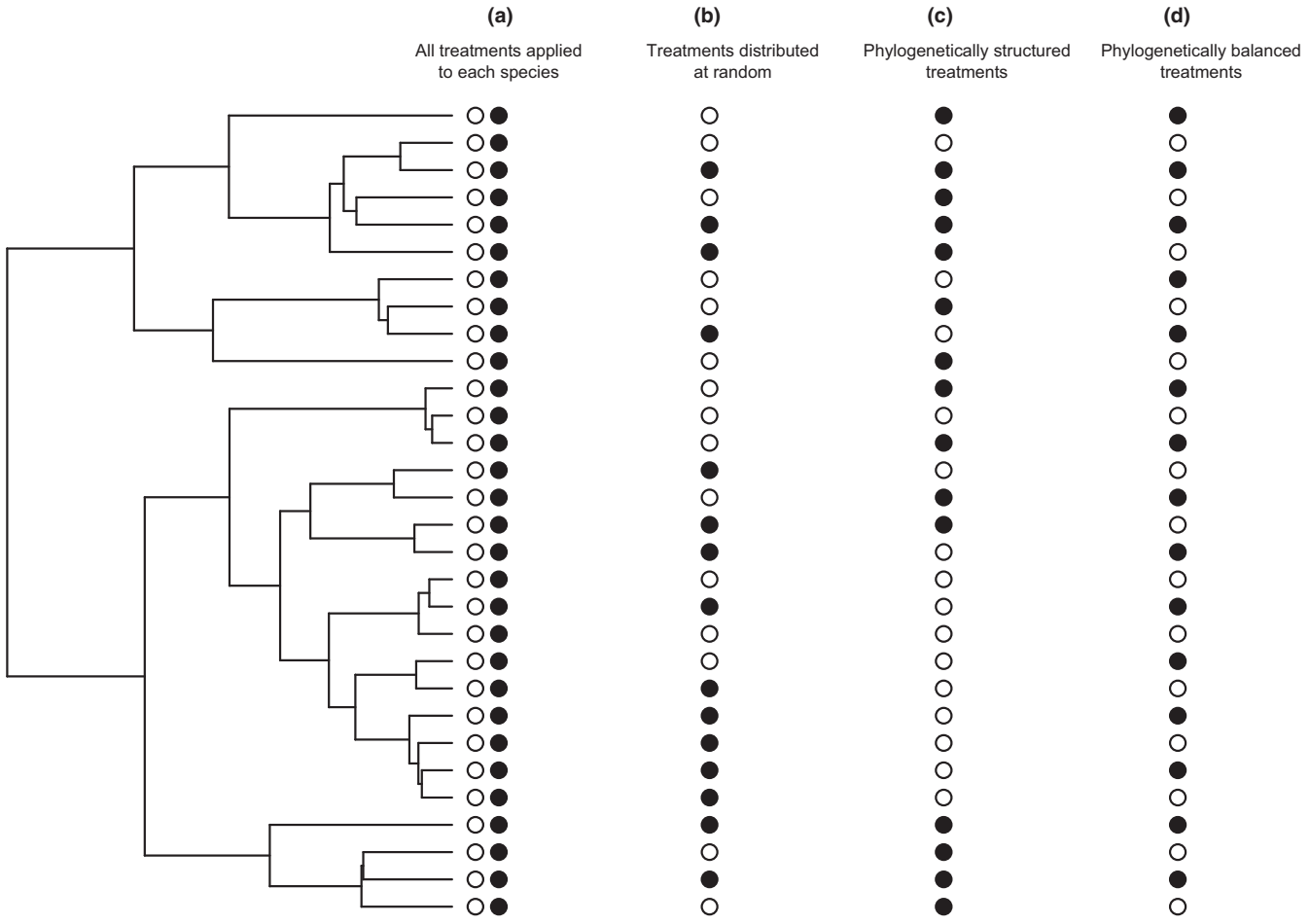

FIGURE 1 Possible experimental designs by which two treatments are assigned to multiple species. (a) A fully randomized design, with both treatments assigned to each species. In (b), (c) and (d), treatments are a property of the species and are not fully randomized. (b) Treatments are assigned at random to species. (c) Treatments are phylogenetically non-randomly distributed, such that closely related species are more similar to each other than to distantly related ones. (d) Phylogenetically balanced treatments, that is, species are chosen so that sister species differ in the factor studied

cases. Numerous approaches have been employed in the literature for the analysis of data from comparative experiments, although the question of how to analyse phylogenetically referenced experimental data does not appear to have been explicitly addressed.

Where there is phylogenetic structure in the factor of interest, there are several possible statistical routes. Treating 'species' as a fixed or random factor is one approach, although it ignores the precise relationships between species. Another approach might be to regard 'species' as the level of replication in the analysis and therefore to stratify the analysis appropriately, for example, by effectively treating species as a 'split-plot factor' in classic ANOVA, or nesting. The efficacy of this approach would be expected to depend on the precise nature of the phylogenetic structure of the factor, however. Another approach is sister-species comparisons (e.g. see Weir & Lawson, 2015 for a recent method). This approach is based on differences between immediately related species and controls for phylogenetic structure at fine taxonomic scales. However, it potentially ignores higher level confounding, for instance if there are clade effects.

Currently the most widely employed approach to analysing comparative data is based on linear models and generalized least squares (e.g. Felsenstein 1973, Felsenstein 1985; Freckleton, Harvey, & Pagel, 2002; Grafen, 1989; Harvey et al., 1995; Martins & Hansen, 1996; Pagel, 1997, 1999). This approach is a generalization of classic regression and ANOVA methods. It accounts for phylogenetic non-independence by the explicit inclusion of a variance–covariance matrix that represents species' phylogenetic relationships. Conceptually there is little difference between this and the approaches that one might use in the analysis of comparative or experimental data. For example, fitting a block in an experimental design simultaneously accounts for uncontrolled variation and reduces non-independence between experimental units. This is akin to the function of a phylogenetic correction (Rees 1995).

Despite the conceptual links, there is currently a gulf between approaches used to analyse comparative, experimental data and the approaches used in correlative analyses. This is reflected in the software that is currently available (e.g. ape Paradis, Claude, & Strimmer, 2004; geiger Pennell et al., 2014; caper Orme, 2013) in which the response variable is assumed to be a single value for each species, and the ability to account for experimental design is not emphasized. For example, in such packages it is not clear how one might include blocking or more sophisticated designs such as error structures that vary between strata (but this may be possible with packages developed in other fields, e.g. see COXME package Therneau, 2015; this is also possible in the package MCMCGLMM (Hadfield & Nakagawa, 2010). This uses a Bayesian approach rather than maximum likelihood or least squares). Methods have been developed to account for intraspecific variability (Felsenstein, 2008; Ives, Midford, & Garland, 2007), which is similar to doing this. These analyses have shown that there are potentially impacts of accounting for intraspecific structure (e.g. Silvestro, Kosktikova, Litsios, Pearman, & Salamin, 2015), and the same would be expected to be true in experimental data.

In this paper, we consider the problem of how to analyse data within a comparative, experimental framework. To our knowledge there has not been a previous analysis of this problem. We use the four experimental schemes outlined in Figure 1 as the basis for exploring the performance of different analytical methods. Analysis of simulations and real data show that, depending on the structure of the data, application of the incorrect method can lead to either loss of power or Type I error. The results show that, with the correct variance structure, Generalized Least Squares approaches outperform other methods. However, our results also highlight that comparative experiments with small numbers (<20) of species have limited power to estimate the variance structure, and that in many such cases the correct analysis may remain in doubt.

2 | MATERIALS AND METHODS

2.1 | Simulations

The four schemes shown in Figure 1 were used as the basis for our simulations. To mimic the type of analysis commonly employed we used a randomized block experiment. For simplicity we assumed that the treatment of interest was a binary variable (as outlined in the discussion, it is straightforward to generalize our results to more complex treatments).

Phylogenies were generated according to a birth-death process using the package TREESIM (Stadler, 2015) in R (R Core Team 2015). Exploratory analysis indicated that the results below are relatively insensitive to the method used to generate the phylogeny, so we held the birth rate of the phylogeny constant at 0.9 and the death rate constant at 0.3. Phylogenies of between 3 and 100 species were generated, representing a typical range that might be encountered in real data.

The basis for the simulations is a general phylogenetic mixed model in which there is a treatment, a blocking structure and residual variance associated with both the individual measurement (e.g. plant-to-plant or pot-to-pot variation in a growth experiment) and with phylogeny.

The model for the vector of observations $\mathbf{y}$ is:

$$\mathbf{y} = \mathbf{B}\mathbf{b}_b + \mathbf{X}\mathbf{b}_x + \mathbf{e} \quad (1)$$

In the fully randomized block design with a binary treatment and b blocks and n species (Figure 1a), $\mathbf{B}$ is a b x $2n$ design matrix representing the blocking structure. We assume that the entries of $\mathbf{B}$ are ordered such that all members of the same block are ordered sequentially. $\mathbf{X}$ is a k x b x n design matrix representing the experimental factor with k levels, with the treatments ordered within blocks. $\mathbf{b}_b$ and $\mathbf{b}_x$ are coefficients that model the differences between blocks and the effects of the treatments respectively. Finally $\mathbf{e}$ is a vector of errors.

The errors $\mathbf{e}$ were assumed to follow a multivariate normal distribution with a variance–covariance matrix $\mathbf{W}$, given by:

$$W = \sigma\left[\lambda\left(1-\psi\right)\mathbf{D}\otimes\mathbf{V} + \psi\mathbf{D}\otimes\mathbf{I}_n + \left(1-\lambda\right)\left(1-\psi\right)\mathbf{I}_{kbn}\right] \quad (2)$$

In equation (2) σ is a scale parameter and $\otimes$ denotes the Kronecker product. **V** is an *n* x *n* variance–covariance matrix given by the phylogeny, and scaled so the leading diagonal entries are all one (all trees were ultrametric). **D** is *kb* x *kb* covariance matrix representing the covariance between species across *k* treatments and *b* blocks: this measures the degree to which the phylogenetic variance structure is the same across different treatments and blocks. We saw no reason to make any assumption about how this would vary, so the entries of **D** were all set to unity, that is, the phylogenetic effect is the same across the whole experiment. $\mathbf{D} \otimes \mathbf{I}_n$ is a *kbn* x *kbn* matrix that codes for species identity (entries corresponding to pairs of experimental units relating to the same species are 1, those corresponding to different species are 0).

There are three components to the variance in equation (2). The first component is the variance among species means that results from phylogenetic dependence (σ^2_p). The second describes variation in the species means that is independent of phylogeny (σ^2_s). The final variance is that between replicate experimental units independent of phylogeny or species identity, that is, the error variance (σ^2_e).

In equation (2), the parameters λ and ψ perform similar but slightly different roles. λ alters the degree of phylogenetic signal in the mean response for each species and thus alters the proportion of variance between species contributed by **V** relative to non-phylogenetic variation (σ^2_e). On the other hand, ψ generates variance in the species means independent of that generated by the phylogeny. This allows for differences between species, unrelated to phylogeny. If ψ is 1 then the phylogenetic component makes no contribution to overall variance. If ψ is 0 then there is no additional species-level variation, that is, all individuals from a species are identical. If $0 < \psi < 1$ and $0 < \lambda < 1$ then both phylogeny and independent species differences play a role.

In order to understand the differences between these different components, it is useful to highlight that the net covariance **W** contains three different types of (co)variances. The first is the expected variance of a species *i* in a given block and treatment:

$$w_{i,i} = \sigma^2 \left[\lambda \left(1-\psi\right) v_{i,i} + \psi + \left(1-\lambda\right)\left(1-\psi\right)\right] \quad (3)$$

The second is the expected covariance of species *i* in a given block and treatment with the value measured from species *i* in another block (*b*) and treatment (*k*):

$$w_{i,kbi} = \sigma^2 \left[\lambda \left(1-\psi\right) v_{i,i} + \psi\right] \quad (4)$$

Although both (3) and (4) refer to intraspecific (co)variance, the covariance (4) is less than the expected variance for an observation (3) because of the additional observation error in (3). This does not arise in (4) because (4) refers only to covariance, that is, expected similarity. Finally there is the expected covariance of species *i* with species *j*:

$$w_{i,j} = \sigma^2 \left[\lambda \left(1-\psi\right) v_{i,j}\right] \quad (5)$$

Although alternative parameterizations to that used in equation (2) are possible they are not necessarily simpler. One possibility would be to redefine λ as the proportion of the total variance explained by phylogeny (i.e. $\sigma^2_p/\sigma^2_{\text{Total}}$). However, this adds complexity to the error component of equation (2), which simple algebra shows to be proportional to $\left[\left(1-\lambda\right)\left(1-\psi\right)-\lambda\psi\right]$. The subtracted variance $\lambda\psi$ in this formulation relates to any variance attributable to both phylogeny and species differences, which is a difficult quantity to understand intuitively. The formulation in equation (2) ensures that λ and ψ are independent, with the components due to phylogeny and species differences clearly separated.

For the cases in Figure 1b–d, it was assumed that each species was only assigned one treatment, that is, the dimensions of **X** were *k* x *n*, the dimension of **B** were *b* x *n* and the dimensions of **y** were 1 x *n*. To generate random samples according to the scheme in Figure 1b, half of the species were chosen at random and assigned to one treatment, the other half were assigned to another. In order to generate the phylogenetically structured distribution of treatments in Figure 1c, we used the phylogeny to create a variance–covariance matrix and then generate a multivariate normal distribution using the mvtnorm package in R (Genz & Bretz, 2009; Genz et al., 2014). Species in the bottom 50% quartile of this random variable were assigned to one treatment, the rest were assigned to the other treatment. Finally, in order to generate the scheme shown in Figure 1d, species were ordered according to phylogenetic relatedness and assigned alternating states.

2.2 | Methods of analysis

In the comparison of methods we outline six different approaches for analysing the data generated according to these models. (1), (2) and (4)–(6) may be used to fit to the experimental scheme shown in Figure 1a, while, (1) and (3)–(6) can be applied to data generated according to the schemes shown in Figure 1b–d. To make the models fully explicit, we include R formulae for the models where appropriate. The full simulation code is available as a supplement.

In the models below, $y_{i,j,k}$ is the observation. $\beta_{0,i}$ is the intercept term with separate intercepts for the $i = 1 \ldots b$ blocks. $\beta_{X,j}$ is the effect of treatment *j*, where $j = 1 \ldots t$ representing the *t* treatments. We assume that both block and treatment are fitted as fixed effects. Typically there are either too few blocks in an experiment or no within-block replication, so that it is not possible to efficiently or reliably treat blocks as random factors. Subscript *k* refers to species $k = 1 \ldots n$. The approaches to modelling the data differ with respect to how the species-specific effects are included. All models incorporate an error term $e_{i,j,k}$ which is a random error for each experimental observation, which is assumed to be normally distributed $e_{i,j,k} \sim N\left(0, \sigma^2_e\right)$. Appendix S1 summarizes R commands for fitting models (1)–(4) which can be fit using straightforward commands.

2.2.1 | Model (1): OLS ANOVA

This is undoubtedly an inappropriate model whenever data contain a phylogenetic signal. However, this approach is still commonly employed in the literature. We show the results obtained using this approach to demonstrate unequivocally the importance of accounting for phylogeny in such analyses. We simply fit 'block' and 'treatment' as fixed factors and species identity was ignored. The model fitted is:

$$y_{i,j,k} = \beta_{0,i} + \beta_{X,j} + e_{i,j,k} \quad (6)$$

2.2.2 | Model (2): OLS ANOVA with species as a 'fixed' factor

This is the simplest approach to including phylogenetic effects, by adding 'species' as an additional fixed factor, that is,

$$y_{i,j,k} = \beta_{0,i} + \beta_k + \beta_{X,j} + e_{i,j,k} \quad (7)$$

In this model, the additional term β_k models species-specific variation.

2.2.3 | Model (3): OLS ANOVA stratified by species

Model 2 will not fit traits that species are assigned only one treatment for (i.e. schemes b–d in Figure 1). This is because species and treatment are confounded. In this case, the appropriate error for the model is the species level. This is the same as treating species as a split-plot factor. In this case the fitted model is:

$$y_{i,j,k} = \beta_{0,i} + \beta_{X,k} + e_k + e_{i,j,k} \quad (8)$$

The term e_k is a species-specific error term, $e_k \sim N\left(0, \sigma_S^2\right)$. The treatment is also species-specific, that is, $\beta_{X,k}$ does not include the j subscript and consequently, the effective level of replication is the species. Hence, the appropriate error term for testing the treatment effect is the variance at the species level.

2.2.4 | Model (4): Linear Mixed Model with species as a 'random' factor

In this model, we treat 'species' as a random factor. For an orthogonal balanced design (e.g. scheme (a) in Figure 1), this offers no advantage over Model 2, and would yield identical results in terms of parameter errors and variances for the fixed effects. This is because the fixed effects in Model 4 are calculated marginally with respect to the random effects. The complication with Model 4 is the calculation of statistical significance for the main effects as the calculation of residual degrees of freedom is not straightforward (e.g. Pinheiro & Bates, 2000). We therefore fitted model 4 by maximum likelihood (rather than REML) and used likelihood ratio tests to test the fixed effect 'treatment'. The model fitted is:

$$y_{i,j,k} = \beta_{0,i} + b_k + \beta_{X,j} + e_{i,j,k} \quad (9)$$

$$b_k \sim N\left(0, \sigma_S^2\right)$$

In this model, the random effects term b_k is assumed to be normally distributed with mean 0 and variance σ_S^2.

2.2.5 | Model (5): PGLS model with known variance components

This model directly fits the 'true' variance structure with λ and ψ set to their simulation values. The reason for fitting this model is to demonstrate how the 'ideal' model for the system behaves relative to other approaches. We would expect this approach to behave well, but a key issue is whether other approaches can approximate this model.

We fitted model (1) directly to the data, using a PGLS approach (e.g. Freckleton et al., 2002; Hansen & Martins, 1996; Pagel, 1997, 1999). The additional element here is that we account for the blocking and treatment structure inherent in an experiment of the design. In contrast, the majority of previous analyses of comparative and cross-species experimental analyses focus on the analysis of data in which each species is represented by only a single measurement. Models for the analysis of data in which each species is measured more than once are described in several papers (Felsenstein, 2008; Hadfield & Nakagawa, 2010 and Ives et al., 2007). These take different approaches. Felsenstein (2008) allows for repeated measurements within species, including species-specific covariance matrices. Ives et al. (2007) use an approach based on PGLS, allowing for intraspecific variance. Hadfield and Nakagawa (2010) describe a Bayesian framework in which it is possible to flexibly specify different intra- and interspecific variance functions. In the simulations, assuming that λ is known, we fitted a pgls model with block and treatment as fixed effects using maximum likelihood.

2.2.6 | Model (6): Phylogenetic generalized least squares with estimated variance components

To pre-empt our main result, our simulations show that Model 5 is the *only* method that behaves well in *all* circumstances. Model 5 assumes that the parameters λ and ψ are known. However, in real applications these parameters are unknown. Using the technique of 'Estimated' Generalized Least Squares (EGLS) (e.g. mentioned in a phylogenetic context by Freckleton et al., 2002 and Ives & Zhu, 2006), the variance structure may be estimated from the data. We used code written by RF, checked against the function lmekin() in the R package COXME (Therneau, 2015) to fit equation (1) to the data and estimate the variance components in equation (2).

Models (1) and (4) can be expressed as special cases of (5), specifically:

$\lambda = 0$ and $\psi = 1$: this is model (1)

$\lambda = 0$ and $0 < \psi < 1$: this is model (4)

$0 < \lambda < 1$ and $0 < \psi < 1$: this is model (5)

The middle model requires some explanation: in this model the phylogenetic variance structure is reduced to a diagonal matrix representing the model for the mean species responses. This implies no covariance among species resulting from phylogeny, but does represent a variance structure for within species variance. In the models considered here, this means that replicates from the same species are more similar to each other than to other species because $\sigma_S^2 > 0$. The difference between models (4) and (5) is that model (4) allows for species differences unrelated to phylogeny, whereas (5) models these differences as a function of phylogenetic distance.

2.3 | Simulation details

Phylogenies of 3–100 species were generated. The value of λ was set at 1 (high phylogenetic signal) generating a situation where species mean responses show a strong phylogenetic signal or set to zero, that is, no phylogenetic signal in the mean response.

The effect of between-species variation unrelated to phylogeny was assumed to be low ($\psi = 0.1$) or high ($\psi = 0.8$). Block effects were included throughout, although initial analysis showed that the assumptions about these were unimportant so long as all species x treatment combinations are present in each block. The treatment was assumed to be binary, and we either assumed no treatment effect (i.e. testing Type I error), or that there was a low to moderate treatment effect (i.e. testing power; $b_0 = 0$, $b_1 = 0.05, 0.1$).

All of the methods described above would be expected to yield unbiased estimates of the model coefficients ($\mathbf{b}_b$ and $\mathbf{b}_X$) in equation (1) (McCullagh & Nelder, 1989), so we did not study the behaviour of the parameter estimates further.

For each set of parameters we calculated the *p*-value for the inclusion of the treatment variable in the model as the main output of the simulation. This focus on *p*-values might seem unusual given the modern emphasis on effect sizes and model selection. However, in this context, in which we are studying the effect of including a binary factor in the model, the test effectively reduces to a *t* test on the inclusion of an additional parameter, that is, $t = (b_{\text{obs}} - b_{\text{test}})/se(b)$ with $n - k$ degrees of freedom, where k is the number of estimated parameters. Given that we know from theory that $(b_{\text{obs}} - b_{\text{test}})$ is unbiased (McCullagh & Nelder, 1989), the *p*-value effectively summarizes the information on both the variance and degrees of freedom (Murtaugh, 2014). Moreover, in an experimental context, the primary aim of the analysis is usually to test the significance of the treatment variable, so this is a practically relevant measure.

2.4 | Case study

In order to demonstrate the consequences of phylogenetic structure and method of analysis for the conclusions drawn from experimental data, we present an analysis of data taken from Taylor, Ripley, Woodward, & Osborne, 2011). We chose this dataset because the fully blocked experiment included treatments that were both fully randomized and phylogenetically structured. This dataset thus encompasses a broad range of scenarios considered in the simulations.

The data are taken from 13 species of grasses, 7 of which have C_4 photosynthesis and 6 of which have C_3 photosynthesis. Five experimental blocks were set up in which plants were either exposed to drought or were watered. One plant was assigned to each treatment within each block. There were 5 (blocks) x 2 (treatments) x 13 (species) = 130 experimental units at the start of the experiment. Full details are given in (Taylor et al., 2011). We report the analysis of measurements of photosynthetic rate (log transformed).

The two factors of interest in this experiment are the watering treatment and the photosynthetic pathway. The watering treatment was assigned in a fully randomized manner, while the photosynthetic pathway was not. The photosynthetic pathway is a property of the species, and cannot be allocated at random to species. The experimental design involved some degree of phylogenetic balancing, so this treatment is somewhere between case (C) and (D) (see Figure 2).

We used this approach to analyse the effects of watering and photosynthetic type. The effect of watering treatment (watering vs. drought) was first analysed using models (1)–(3). The effect of photosynthetic type (C_3 vs. C_4) was analysed using models (1), (2) and (4). The results are summarized in Table 1. To analyse the data from this experiment, we used Model (6).

3 | RESULTS

3.1 | Simulation results

The simulations show that the results of analyses of experimental data on multiple species are dependent on both the structure of the data and the method of analysis employed (Tables 1 & 2). The only situation where results are robust to varying the analysis method employed is a fully randomized experiment in which all treatments are applied to each species (scenario (a) in Figure 1). Apart from Model 1, which ignores species identity completely, all methods have acceptable Type I error (Table 1a) and the same power (Table 2a,b). Model 1 ignores species identify and so is anti-conservative. This is because the variance resulting from species differences is not accounted for.

When each species does not receive every treatment, the results of the analyses are sensitive to the experimental design, method of analysis and degree of phylogenetic dependence. When phylogenetic dependence is strong, the difference in Type I error between phylogenetically balanced versus randomly distributed treatments in only evident for the non-phylogenetic analysis (Table 1b,d). Otherwise, in these cases, the performance of the phylogenetically corrected approaches is unaffected by the distribution of the traits. These methods are thus reasonably robust in terms of Type I errors.

On the other hand, phylogenetic structure in the treatment variable leads to high Type I errors for all methods apart from PGLS when phylogenetic signal is strong (Table 1c). Notably, the Type I error rate increases with the number of species in the analysis, reflecting the misattribution of variance. Even for methods such as mixed models

TABLE 1 Type I error rates for different methods of analysis of experimental data generated according to a range of designs and assuming a 5% threshold for statistical significance

	$\lambda = 1/\psi = 0.8$						$\lambda = 1/\psi = 0.1$						$\lambda = 0/\psi = 0.8$					
	Number of species						Number of species						Number of species					
	3	5	10	20	50	100	3	5	10	20	50	100	3	5	10	20	50	100
(a) Fully randomized																		
(1) OLS	0.03	0.04	0.03	0.04	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.03	0.04	0.03	0.03	0.02
(2/3) ANOVA	0.04	0.06	0.04	0.05	0.04	0.05	0.05	0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.04
(4) LMM	0.07	0.07	0.04	0.06	0.04	0.05	0.07	0.07	0.05	0.05	0.05	0.05	0.09	0.06	0.06	0.06	0.05	0.04
(5) PGLS	0.07	0.07	0.04	0.05	0.04	0.05	0.07	0.07	0.03	0.02	0.01	0.01	0.09	0.06	0.06	0.06	0.05	0.04
(6) EGLS	0.07	0.07	0.04	0.05	0.04	0.05	0.07	0.07	0.05	0.05	0.05	0.05	0.09	0.06	0.06	0.06	0.05	0.04
(b) Phylo random																		
(1) OLS	0.13	0.13	0.12	0.14	0.13	0.16	0.56	0.51	0.43	0.41	0.40	0.43	0.15	0.17	0.17	0.17	0.17	0.19
(2/3) ANOVA	0.05	0.04	0.05	0.04	0.04	0.05	0.07	0.06	0.04	0.04	0.05	0.05	0.04	0.04	0.04	0.06	0.05	0.07
(4) LMM	0.17	0.11	0.08	0.05	0.05	0.05	0.34	0.16	0.09	0.06	0.06	0.05	0.18	0.13	0.08	0.07	0.06	0.07
(5) PGLS	0.06	0.05	0.05	0.05	0.05	0.06	0.05	0.05	0.04	0.05	0.05	0.06	0.05	0.05	0.05	0.05	0.05	0.06
(6) EGLS	0.11	0.09	0.07	0.06	0.05	0.05	0.37	0.19	0.11	0.06	0.04	0.05	0.12	0.12	0.08	0.08	0.06	0.07
(c) Phylo structured																		
(1) OLS	0.12	0.13	0.18	0.26	0.33	0.44	0.53	0.46	0.59	0.62	0.67	0.75	0.17	0.15	0.17	0.19	0.18	0.16
(2/3) ANOVA	0.05	0.04	0.09	0.14	0.22	0.30	0.04	0.04	0.16	0.22	0.31	0.44	0.05	0.04	0.05	0.06	0.06	0.05
(4) LMM	0.16	0.10	0.13	0.16	0.23	0.31	0.31	0.16	0.22	0.25	0.32	0.45	0.21	0.12	0.08	0.07	0.07	0.05
(5) PGLS	0.05	0.05	0.04	0.05	0.05	0.05	0.06	0.05	0.04	0.06	0.05	0.04	0.04	0.04	0.06	0.05	0.05	0.05
(6) EGLS	0.10	0.08	0.12	0.11	0.09	0.06	0.35	0.19	0.18	0.11	0.05	0.04	0.13	0.10	0.08	0.07	0.07	0.05
(d) Balanced																		
(1) OLS	–	–	0.13	0.14	0.13	0.14	–	–	0.41	0.44	0.41	0.44	–	–	0.16	0.16	0.16	0.13
(2/3) ANOVA	–	–	0.06	0.06	0.05	0.05	–	–	0.04	0.05	0.05	0.04	–	–	0.05	0.06	0.06	0.03
(4) LMM	–	–	0.08	0.07	0.05	0.05	–	–	0.07	0.06	0.06	0.04	–	–	0.07	0.07	0.06	0.04
(5) PGLS	–	–	0.06	0.06	0.04	0.05	–	–	0.04	0.06	0.05	0.08	–	–	0.04	0.06	0.05	0.04
(6) EGLS	–	–	0.08	0.07	0.04	0.05	–	–	0.08	0.06	0.04	0.05	–	–	0.07	0.07	0.06	0.04

Type I error proportions of 0.1 and higher (i.e. twice the nominal rate and higher) are highlighted in red. The simulated experiment involved applying a binary treatment to 10, 20, 50 or 100 species in a randomised block design. The four experimental designs were (see Figure 1 for a graphical illustration): (a) Fully randomized design in which each treatment was applied to every species; (b) each species was assigned to one treatment or another at random; (c) each species was assigned to one treatment or another in a phylogenetically structured manner (note this is not feasible for small sample sizes of three or five species); (d) the treatments were phylogenetically balanced. The amount of phylogenetic signal (λ) and species-specific variation (ψ) were varied as shown.

TABLE 2 Power of different methods of analysis of experimental data generated according to a range of designs

	$\lambda = 1/\psi = 0.8$						$\lambda = 1/\psi = 0.1$						$\lambda = 0/\psi = 0.8$					
	Number of species						Number of species						Number of species					
	3	5	10	20	50	100	3	5	10	20	50	100	3	5	10	20	50	100
(a)																		
(A) Fully Randomized																		
(1) OLS	0.00	0.00	0.03	0.02	0.05	0.08	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.02	0.03	0.09
(2/3) ANOVA	0.01	0.00	0.04	0.04	0.07	0.11	0.05	0.07	0.17	0.34	0.77	0.91	0.00	0.00	0.02	0.03	0.06	0.14
(4) LMM	0.02	0.00	0.04	0.04	0.07	0.11	0.05	0.08	0.19	0.35	0.77	0.91	0.00	0.00	0.01	0.02	0.06	0.14
(5) PGLS	0.01	0.00	0.04	0.04	0.06	0.08	0.06	0.08	0.14	0.24	0.58	0.83	0.00	0.00	0.02	0.02	0.06	0.14
(6) EGLS	0.02	0.00	0.04	0.04	0.07	0.11	0.05	0.08	0.18	0.35	0.77	0.91	0.00	0.00	0.01	0.02	0.06	0.14
(B) Phylo random																		
(1) OLS	0.02	0.01	0.00	0.01	0.03	0.03	0.00	0.00	0.02	0.00	0.02	0.02	0.01	0.00	0.00	0.00	0.00	0.04
(2/3) ANOVA	0.00	0.03	0.00	0.01	0.02	0.04	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.00	0.00	0.03
(4) LMM	0.03	0.00	0.00	0.02	0.01	0.04	0.02	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.02	0.00	0.00	0.03
(5) PGLS	0.00	0.00	0.01	0.00	0.01	0.03	0.00	0.01	0.01	0.01	0.05	0.09	0.01	0.00	0.02	0.00	0.00	0.03
(6) EGLS	0.01	0.00	0.00	0.00	0.01	0.03	0.01	0.00	0.00	0.01	0.03	0.07	0.00	0.00	0.01	0.00	0.00	0.03
(C) Phylo structured																		
(1) OLS	–	–	0.01	0.01	0.04	0.06	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.04	0.02	0.03	0.00	0.03
(2/3) ANOVA	–	–	0.00	0.01	0.01	0.04	0.00	0.01	0.00	0.01	0.03	0.00	0.01	0.02	0.00	0.01	0.00	0.01
(4) LMM	–	–	0.01	0.02	0.01	0.04	0.00	0.00	0.01	0.01	0.03	0.00	0.01	0.03	0.01	0.02	0.00	0.01
(5) PGLS	–	–	0.00	0.01	0.03	0.05	0.00	0.01	0.03	0.01	0.01	0.02	0.02	0.02	0.00	0.02	0.01	0.01
(6) EGLS	–	–	0.00	0.01	0.02	0.04	0.00	0.00	0.01	0.01	0.01	0.02	0.02	0.03	0.00	0.02	0.00	0.01
(D) Balanced																		
(1) OLS	–	–	0.17	0.20	0.20	0.20	–	–	0.04	0.00	0.04	0.00	–	–	0.02	0.04	0.04	0.07
(2/3) ANOVA	–	–	0.04	0.07	0.08	0.08	–	–	0.02	0.00	0.02	0.03	–	–	0.00	0.01	0.02	0.04
(4) LMM	–	–	0.08	0.09	0.09	0.08	–	–	0.02	0.01	0.02	0.03	–	–	0.02	0.02	0.03	0.04
(5) PGLS	–	–	0.05	0.07	0.08	0.08	–	–	0.01	0.00	0.04	0.09	–	–	0.01	0.01	0.03	0.04
(6) EGLS	–	–	0.08	0.09	0.09	0.08	–	–	0.03	0.00	0.04	0.07	–	–	0.01	0.02	0.03	0.04
(b)																		
(A) Fully Randomized																		
(1) OLS	0.01	0.02	0.05	0.06	0.21	0.37	0.02	0.01	0.02	0.02	0.15	0.58	0.01	0.02	0.04	0.06	0.15	0.41
(2/3) ANOVA	0.02	0.01	0.06	0.09	0.26	0.43	0.21	0.32	0.61	0.88	0.95	0.95	0.02	0.03	0.05	0.07	0.19	0.49
(4) LMM	0.03	0.01	0.07	0.09	0.26	0.43	0.26	0.34	0.62	0.88	0.95	0.95	0.01	0.02	0.05	0.08	0.20	0.49

(Continues)

TABLE 2 (Continued)

	$\lambda = 1/\psi = 0.8$						$\lambda = 1/\psi = 0.1$						$\lambda = 0/\psi = 0.8$					
	Number of species						Number of species						Number of species					
	3	5	10	20	50	100	3	5	10	20	50	100	3	5	10	20	50	100
(5) PGLS	0.03	0.01	0.06	0.08	0.24	0.41	0.26	0.35	0.55	0.80	0.98	0.99	0.01	0.02	0.05	0.08	0.20	0.49
(6) EGLS	0.03	0.01	0.07	0.09	0.26	0.43	0.26	0.34	0.62	0.88	0.95	0.95	0.01	0.02	0.05	0.08	0.20	0.49
(B) Phylo random																		
(1) OLS	0.03	0.01	0.02	0.03	0.10	0.16	0.00	0.00	0.03	0.04	0.06	0.08	0.02	0.01	0.03	0.03	0.06	0.13
(2/3) ANOVA	0.00	0.01	0.02	0.02	0.06	0.11	0.00	0.00	0.02	0.01	0.03	0.03	0.00	0.01	0.02	0.00	0.04	0.07
(4) LMM	0.03	0.01	0.01	0.03	0.06	0.12	0.00	0.01	0.02	0.02	0.03	0.03	0.01	0.00	0.01	0.00	0.04	0.08
(5) PGLS	0.00	0.00	0.01	0.03	0.07	0.16	0.00	0.01	0.04	0.06	0.19	0.39	0.00	0.00	0.02	0.02	0.04	0.08
(6) EGLS	0.01	0.02	0.01	0.03	0.06	0.15	0.00	0.02	0.02	0.06	0.14	0.33	0.01	0.00	0.01	0.00	0.04	0.08
(C) Phylo structured																		
(1) OLS	0.01	0.00	0.05	0.01	0.08	0.06	0.00	0.03	0.00	0.01	0.03	0.00	0.02	0.03	0.03	0.01	0.07	0.17
(2/3) ANOVA	0.00	0.00	0.01	0.02	0.05	0.05	0.00	0.02	0.00	0.03	0.06	0.01	0.00	0.02	0.02	0.01	0.03	0.10
(4) LMM	0.00	0.01	0.02	0.02	0.06	0.05	0.00	0.00	0.00	0.04	0.06	0.01	0.02	0.02	0.02	0.01	0.03	0.10
(5) PGLS	0.01	0.00	0.01	0.03	0.06	0.06	0.00	0.02	0.03	0.02	0.05	0.10	0.02	0.02	0.01	0.01	0.03	0.09
(6) EGLS	0.00	0.00	0.01	0.03	0.05	0.08	0.00	0.00	0.00	0.03	0.04	0.10	0.03	0.02	0.02	0.01	0.03	0.11
(D) Balanced																		
(1) OLS	—	—	0.19	0.20	0.26	0.30	—	—	0.03	0.03	0.04	0.04	—	—	0.03	0.03	0.10	0.17
(2/3) ANOVA	—	—	0.07	0.07	0.10	0.14	—	—	0.03	0.02	0.02	0.05	—	—	0.02	0.01	0.04	0.10
(4) LMM	—	—	0.09	0.08	0.10	0.14	—	—	0.04	0.03	0.02	0.05	—	—	0.03	0.01	0.05	0.11
(5) PGLS	—	—	0.08	0.06	0.10	0.14	—	—	0.03	0.05	0.19	0.36	—	—	0.04	0.00	0.05	0.10
(6) EGLS	—	—	0.09	0.09	0.10	0.14	—	—	0.05	0.05	0.16	0.31	—	—	0.03	0.01	0.05	0.11

In this Table, power is measured as the proportion of times that a statistically significant result is recorded, minus the proportion of times a Type I error is expected (Table 1). The details of the simulations are as described in Table 1/Figure 1. The effect size simulated was low (0.05) and higher (0.1).

and ANOVA that 'control' for species differences, Type I error can greatly exceed the nominal 5% for larger phylogenies.

In terms of power, PGLS yields the highest or equal-highest power for all parameter combinations (Table 2). This emphasizes the importance of accurately identifying the correct variance structure even in experimental analyses. Table 2 measures effective power, that is, the proportion of times a statistically significant result is recorded, minus the proportion of times a Type I error is expected. Thus, the power of tests that have high Type I errors in Table 1 is typically low in Table 2.

In terms of design, fully randomized experiments have consistently high power (Table 2). Phylogenetically random and balanced designs combined with PGLS methods yield moderate to high power, with phylogenetically structured treatments yielding experiments with considerably lower effective power. If there is a choice of strategies for the allocation of treatments in an experiment, therefore, the design is important.

There is a qualitative difference between the effect of varying λ and Ψ. When λ is 1 and ψ is 0.1, PGLS (Model 5) has highest power (Table 2), appropriate Type I error (Table 1) and other methods perform poorly. However, when ψ is high (0.8), then irrespective of the value of λ, the power of all techniques is lower. The reason is that when ψ is high, there are large unknown differences between species, and these differences have to be estimated from the data. The effect of setting ψ greater than zero is akin to including phylogenetic branch lengths that are unknown resulting in unaccounted differences between species. In contrast, when λ is high and the phylogeny is known, the expected covariance among species is known and can be potentially corrected for. The need to estimate species effects from the data when ψ is high therefore reduces power. Note, that the same is true for models (2)–(4) when applied to data in which ψ is 0, but λ is high. In this case, these methods are attempting to deal with phylogenetic independence without an estimate of the covariance structure among the species and yield similarly low power (Table 2).

In summary, each of models (1)–(5) can yield statistically acceptable or equivalent results under some circumstances. However, PGLS is the only method to perform well under all tested scenarios. The problem, of course is that our simulations with PGLS assume that the true variance structure is known. In reality the true variance structure can never be known, with one course of action being that we estimate this from the data. The EGLS simulations in which the variance structure is estimated from the data indicate that the outcome of this can be mixed.

Of the techniques, PGLS and EGLS performs as well or better in terms of Type I errors (Table 1) or power (Table 2) when phylogenetic signal is low ($\lambda = 0$). In this case, the method is accurate at identifying a lack of phylogenetic signal (e.g. see simulations in Freckleton et al., 2002) and the results are almost identical to those from the PGLS in which ψ is fixed to its true value.

When phylogenetic signal is high ($\lambda = 1$), then there are two circumstances under which EGLS performs below PGLS in terms of power: (a) when the number of species is small and experiments are not fully randomized; (b) when the treatments are phylogenetically structured. Particularly in the latter case the rates of Type I error can be considerable, although not as high as those of methods that do not appropriately account for phylogeny. Unfortunately when phylogenetic signal exists in the treatment variable, our results indicate that it may be difficult to reliably fit models to experimental data if the number of species is small.

3.2 | Analysis of real data

As shown in Figure 2, the structure of the data mirrors the situations envisaged in Figure 1. The watering treatment was fully randomized as in Figure 1a. The photosynthetic type varies with a combination of phylogenetic balancing by design (Figure 1d) and phylogenetic structure (Figure 1c). The analysis of the effect of the watering treatment indicated that the conclusions drawn were relatively invariant to the choice of method of analysis. In this case, the results obtained from ANOVA, random effects model and the PGLS were very similar indeed (Table 3a). The probability value for the OLS model ignoring phylogeny altogether was larger reflecting the lower power of this method observed in the simulations. Note that in this dataset the maximum likelihood value of λ is zero, which means that the PGLS model and the REML mixed model produce equivalent parameter estimates (to within a trivial numerical difference).

However, the analysis of the effect of photosynthetic type indicated that the choice of method was important. Depending on the approach chosen, the result was statistically highly significant (OLS), clearly non-significant (ANOVA), marginally non-significant (random effects model) or marginally significant (EGLS) (Table 1b). This reinforces the conclusion from the modelling, namely that when phylogenetic signal is strong in both the data and the test variables, the results obtained may be very sensitive to the choice of model.

4 | DISCUSSION

The strongest message from the results presented above is that the power of comparative experiments is maximized when the correct variance structure is incorporated into the analysis. This, of course, applies to all statistical models and in that respect there is nothing different about comparative analyses. However, we are aware of no previous review of the analysis of comparative data in an experimental context. Depending on the design of the experiment, there may be increased Type I error rates when the correct variance structure is not included. This occurs particularly if the treatments are naturally varying ones that show phylogenetic structure, and in this case the more commonly used methods for experimental data analysis can perform poorly. The acceptable rates for PGLS rely on knowing the correct variance structure, however, this can only ever be estimated empirically.

The only experimental design that is completely robust is the fully randomized design, with all treatments applied to each species. For many factors of interest, however, this is not possible and consequently the method of analysis is important. To illustrate this, we

presented a simple case study in which it is possible to get a range of answers depending on the method employed.

The analysis of experimental data seems to have been overlooked in the comparative literature. In the development of new methods it is typically assumed that phylogenetic comparative methods are applied in a correlative manner, frequently to literature-derived observational data. On the other hand, many studies report experiments performed on multiple species and our results highlight that these should carefully consider the choice of analysis.

One approach to analysing the data from multiple species in an experiment is to generate a single mean for the response for each species, and use this in conventional phylogenetic analysis (e.g. PGLS). In the example considered in the simulations above, this would involve averaging values across the blocks. In the case of the fully randomized design (Figure 1a), this would generate two values per species; in the case of the others (Figure 1b–d) this would yield one value per species. This analysis is effectively the same as model (3) if no further phylogenetic correction is applied. This is justifiable, but with three limitations: (a) the design of the experiment must permit simple averaging of species traits (e.g. ideally completely balanced). (b) Consequently, there can be no missing data as the means are taken across experimental replicates. (c) The estimates of phylogenetic signal and experimental error cannot be disentangled, that is, λ and ψ cannot be separately estimated. If the experimental design is not simple (e.g. split plot or repeated measures are included) then the model could not be fit using this approach without carefully and appropriately averaging across clusters, and we are aware of no examples of studies that have analysed such designs in a phylogenetic comparative context.

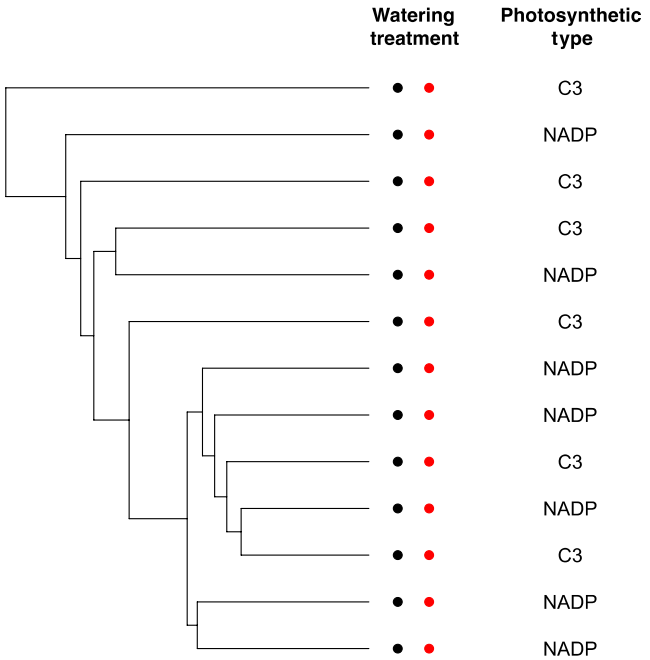

FIGURE 2 Experimental design in an example dataset. There are 13 species related by the phylogeny shown. There were two factors considered. Watering treatment was a binary variable, with each treatment applied to each species. Photosynthetic type is a binary variable, but each species was either C_3 or NADP

One approach is to model measurement errors in comparative analyses through adding a vector of errors (e.g. Silvestro et al., 2015). However, this would be inadvisable in the analysis of experimental data, as the individual species-level values will individually be poor estimates of the error variance. This is particularly so if the number of blocks is small (typical values being in the range 3–10 for most experiments). It is advised that measurement error cannot be accurately calculated from fewer than ~12 observations (Hansen & Bartoszek, 2012).

Models for intraspecific variation in comparative data have been described, and these are closely related to the approaches for analysing experimental data described here (e.g. Felsenstein, 2008)). Such models allow for intraspecific variation and are specifically designed to consider the variation within measured units, which is typically assumed to have a deterministic basis. These methods are essentially the same as the pgls approach: for example, through including additional variables experimental designs such as blocking could be included.

Linear mixed models are frequently used to analyse data from experiments, with species specified as a random factor (e.g. Brown et al., 2014; Jamil, Ozinga, Kleyer, & ter Braak, 2013). The problems with using LMMs are twofold. First, although they account for differences between species, they do not allow for higher level phylogenetic dependence. They assume that $\lambda = 0$, although the inclusion of higher level taxonomic grouping factors could improve the performance of this approach providing the taxonomic relationships reflect the underlying phylogeny. The second problem with LMMs is that it is not straightforward to calculate degrees of freedom relating to the random components of the model. If the random components of the models are not varied then this is not an issue for examining the fixed effects. It is worth noting that in the situations simulated above, the parameter estimates and error variances from the LMM and the fixed effect models were identical. This strongly suggests that in the LMMs the number of grouping factors (i.e. # of species) is a number of estimated parameters.

Our analyses of real data showed that in one case the results obtained were relatively insensitive to the choice of method. But in another, the results in terms of the fixed factor of interest varied widely. Depending on the choice of method, the treatment effect was highly significant, non-significant, marginally non-significant or marginally significant. We also found that the results were relatively insensitive to the model by which the phylogeny was generated. We also simulated with trees that were pectinate and in which trees had very short tip distances (coalescent trees). In terms of the results in Table 1, we found that the pectinate tress yielded slightly lower Type I errors, while coalescent trees yielded slightly higher (results not shown). But overall the effect was not as important as varying the phylogenetic distribution of the experimental factor.

The comparative method is typically thought of as an observational approach (Harvey & Pagel 1991). In practice, however, there are many studies that integrate comparative and experimental methods (Weber & Agrawal 2012). Arnold & Nunn (2010) considered one aspect of this integration, namely how one chooses species in order

TABLE 3 Analysis of real experiment data using different approaches

Model	(a) Watering Treatment			(b) Photosynthetic type	
OLS	*b*	−0.078		0.520	
	se	0.131		0.123	
	t	−0.598		4.237	
	P	0.551		0.000	
ANOVA	*b*	−0.083		—	
	se	0.094		—	
	t	−0.884		$F = 1.081$	
	P	0.379		0.370	
Random Effects		ML	REML	ML	REML
	b	−0.083	−0.083	0.499	0.498
	se	0.092	0.094	0.258	0.281
	t	−0.900	−0.880	1.934	1.776
	P[a]	0.369		0.070	
PGLS	*b*	−0.083	$\lambda = 0$	0.493	$\lambda = 0.483$
	se	0.092	$\psi = 0.487$	0.234	$\psi = 0$
	t	−0.900		2.100	
	P	0.370		0.036	

The data are taken from Taylor et al. (2011). The response variable is log(photosynthetic rate). The predictors in the model are watering treatment (assigned in a randomized, fully factorial manner) and photosynthetic type (C3 or C4, with each species being one or the other). The data were analysed in four ways: (i) OLS: the predictors were fitted singly, not accounting for phylogeny. (ii) ANOVA: in the case of watering treatment, species was fitted as an additional fixed factor; photosynthetic type varies at the species level, so a split-plot ANOVA was used; (iii) Random effects: species identity was fitted as a random effect. This was fitted by Maximum Likelihood (ML) and the effect of the predictor tested using a likelihood ratio test, relative to a simpler model. (iv) PGLS: as described in the text, a phylogenetic variance matrix was included in the model and the parameters λ and ψ were estimated to measure the effects of phylogenetic structure and species-specific variation respectively

[a]tested by likelihood ratio test.

to maximize the power of analyses. They highlighted that appropriate phylogenetic targeting could considerably enhance the power of tests and that even for the same number of species, power could vary considerably depending on how they are distributed phylogenetically. Our results similarly show that the phylogenetic distribution of traits play an important role in determining both the Type I error and power of experimental studies.

4.1 | Recommendations

If the effects of naturally varying traits are examined, then the phylogenetic distribution of these is extremely important. More extreme cases are possible than we considered in our simulations. For example, 'grade shifts' occur when the values of entire clades are identical. Comparative analysis on such traits would have to be approached with care.

We recommend that the phylogenetic distribution of experimental factors is checked prior to undertaking an experiment. For example, the phylogenetic signal of a binary predictor can be assessed using the D statistic (Fritz & Purvis, 2010). For a multi-level predictor, the phylogenetic distribution of each level could be assessed using the same approach. Our results clearly indicate that all empirically applicable methods, including EGLS, may be severely compromised if experimental factors show strong phylogenetic dependence.

The technique of phylogenetic balancing is closely related to the sister-species approach that has been used a great deal over several decades (e.g. see Weir & Lawson, 2015 for a recent application). The sister-species approach is based on choosing closely related pairs of species that differ in some key characteristic. Differences between other traits of the sister species are then tested. Sister-species comparisons would ignore higher level confounding (being based on comparisons at the species-pair level only) and also reduce sample sizes relative to fully phylogenetically explicit methods (being based on $n/2$ comparisons for n species).

We would also recommend the use of simulations to test the statistical performance of experimental designs prior to analyses. Such simulations are relatively straightforward to implement and would give insights into the likely pitfalls of any proposed experimental schemes, or the relative performance of alternative designs.

The results with varying ψ indicated that high values of ψ led to low power for all tests apart from fully randomized experiments, irrespective of sample size. As noted above, high values of ψ yield large species-specific mean differences that detract from the power to compare species in terms of other traits. Fully randomized experiments perform better because these differences are controlled through design and, indeed, the results are thus insensitive to the method of analysis. It cannot be predicted in advance whether species differ in this way, and hence a risk of any experiment that relies

on natural variation is that species-specific non-phylogenetic variation might obscure treatment effects.

ACKNOWLEDGEMENTS

We thank Sam Taylor and Colin Osborne for access to data.

AUTHORS' CONTRIBUTIONS

R.P.F. and M.R. conceived the project, undertook the analysis and co-wrote the paper.

DATA ACCESSIBILITY

The data and code for running the simulations and analysis are available here: https://figshare.shef.ac.uk/s/cee0d9a301673bd7578b (https://doi.org/10.15131/shef.data.7642628) (Freckleton & Rees, 2019).

ORCID


Robert P. Freckleton https://orcid.org/0000-0002-8338-864X

REFERENCES

Arnold, C., & Nunn, C. L. (2010). Phylogenetic targeting of research effort in evolutionary biology. *American Naturalist*, *176*, 601–612.

Brown, A. M., Warton, D. I., Andrew, N. R., Binns, M., Cassis, G., & Gibb, H. (2014). The fourth-corner solution - using predictive models to understand how species traits interact with the environment. *Methods in Ecology and Evolution*, *5*, 344–352. https://doi.org/10.1111/2041-210X.12163

Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, *25*, 471–492.

Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist*, *126*, 1–25.

Felsenstein, J. (2008). Comparative methods with sampling error and within-species variation: Contrasts revisited and revised. *Amercian Naturalist*, *171*, 713–725. https://doi.org/10.1086/587525

Freckleton, R. P., Harvey, P. H., & Pagel, M. (2002). Phylogenetic analysis and comparative data: A test and review of evidence. *American Naturalist*, *160*, 712–726. https://doi.org/10.1086/343873

Freckleton, R. P., & Pagel, M. (2010). Recent advances in comparative methods. In T. Székely, A. J. Moore, & J. Komdeur (Eds.), *Social behaviour: Genes, ecology and evolution*. Cambridge, UK: Cambridge University Press.

Freckleton, R. P., & Rees, M. (2019). Data from: Comparative analysis of experimental data. *figshare*, https://figshare.shef.ac.uk/s/cee0d9a301673bd7578b

Fritz, S. A., & Purvis, A. (2010). Selectivity in mammalian extinction risk and threat types: A new measure of phylogenetic signal strength in binary traits. *Conservation Biology*, *24*, 1042–1051. https://doi.org/10.1111/j.1523-1739.2010.01455.x

Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and t probabilities*. Heidelberg, Germany: Springer-Verlag. https://doi.org/10.1007/978-3-642-01689-9

Genz, A., Bretz, F., Miwa, T., Mi, X., Liesch, F., Scheipl, F., & Hothorn, T. (2014) mvtnorm: Multivariate normal and t distributions. R package 1.0-2.

Grafen, A. (1989). The phylogenetic regression. *Philosophical Transactions of the Royal Society of London Series B*, *326*, 119–157. https://doi.org/10.1098/rstb.1989.0106

Grime, J. P., Hodgson, J. G., & Hunt, R. (1990). *Comparative Plant Ecology*. London, UK: Chapman & Hall.

Hadfield, J. D., & Nakagawa, S. (2010). General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology*, *23*, 494–508. https://doi.org/10.1111/j.1420-9101.2009.01915.x

Hansen, T. F., & Bartoszek, K. (2012). Interprerting the evolutionary regression: The interplay between observational and biological errors in phylogenetic comparative studies. *Systematic Biology*, *61*, 413–425. https://doi.org/10.1093/sysbio/syr122

Hansen, T. F., & Martins, E. P. (1996). Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution*, *50*, 1404–1417. https://doi.org/10.1111/j.1558-5646.1996.tb03914.x

Harvey, P. H. (1996). Phylogenies for ecologists. *Journal of Animal Ecology*, *65*, 255–263. https://doi.org/10.2307/5872

Harvey, P. H., & Pagel, M. D. (1991). *The comparative method in evolutionary biology*. Oxford, UK: Oxford University Press.

Harvey, P. H., Read, A. F., & Nee, S. (1995). Why ecologists need to be phylogenetically challenged. *Journal of Ecology*, *83*, 535–536. https://doi.org/10.2307/2261606

Ives, A. R., Midford, P. E., & Garland, T. (2007). Within-species variation and measurement error in phylogenetic comparative methods. *Systematic Biology*, *56*, 252–270. https://doi.org/10.1080/10635150701313830

Ives, A. R., & Zhu, J. (2006). Statistics for correlated data: Phylogenies, space and time. *Ecological Applications*, *16*, 20–32. https://doi.org/10.1890/04-0702

Jamil, T., Ozinga, W. A., Kleyer, M., & ter Braak, C. J. F. (2013). Selecting traits that explain species-environment relationships: A generalized linear mixed model approach. *Journal of Vegetation Science*, *24*, 988–1000. https://doi.org/10.1111/j.1654-1103.2012.12036.x

Martins, E. P., & Garland, T. (1991). Phylogenetic analyses of the correlated evolution of continuous characters: A simulation study. *Evolution*, *45*, 534–557. https://doi.org/10.1111/j.1558-5646.1991.tb04328.x

Martins, E. P., & Hansen, T. F. (1996). The statistical analysis of interspecific data: A review and evaluation. In E. P. Martins (Ed.), *Phylogenies and the comparative method in animal behaviour* (pp. 22–75). Oxford, UK: Oxford University Press.

Maynard Smith, J. (1978). Optimization theory in evolution. *Annual Review of Ecology and Systematics*, *9*, 31–56. https://doi.org/10.1146/annurev.es.09.110178.000335

McCullagh, P., & Nelder, J. A. (1989). *Generalisd linear models*. London, UK: Chapman & Hall. https://doi.org/10.1007/978-1-4899-3242-6

Mead, R. (1988). *The Design of Experiments*. Cambridge, NY: Cambridge University Press.

Murtaugh, P. A. (2014) In defense of P values. *Ecology*, *95*, 611–617. https://doi.org/10.1890/13-0590.1

Newman, J. A., Bergelson, J., & Grafen, A. (1997). Blocking factors and hypothesis tests in ecology: is your statistics text wrong? *Ecology*, *78*(5), 1312–1320.

Nunn, C. L. (2011). *The Comparative Method in Evolutionary Anthropology and Biology*. University of Chicago Press.

Orme, D. (2013) The caper package: comparative analysis of phylogenetics and evolution in R.

Pagel, M. (1997). Inferring evolutionary processes from phylogenies. *Zoologica Scripta*, *26*, 331–348. https://doi.org/10.1111/j.1463-6409.1997.tb00423.x

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, *401*, 877–884. https://doi.org/10.1038/44766

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analysis of phylogenetics and evolution in R language. *Bioinformatics*, *20*, 289–290. https://doi.org/10.1093/bioinformatics/btg412

Pennell, M. W., Eastman, J. M., Slater, G. J., Brown, J. W., Uyeda, J. C., FitzJohn, R. G., … Harmon, L. J. (2014). geiger 2.0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*, *30*, 2216–2218. https://doi.org/10.1093/bioinformatics/btu181

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed effect models in S and Splus*. New York, NY: Springer. https://doi.org/10.1007/978-1-4419-0318-1

Rees, M. (1995). EC-PC comparative analyses? *Journal of Ecology*, *83*, 891–892.

R Core Team. (2015) *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Silvestro, D., Kosktikova, A., Litsios, G., Pearman, P. B., & Salamin, N. (2015). Measurment errors should always be incorporated in phylogenetic comparative analysis. *Methods in Ecology and Evolution*, *6*, 340–346. https://doi.org/10.1111/2041-210X.12337

Stadler, T. (2015). TreeSim: Simulating phylogenetic Trees. *R package version*, *2*, 2.

Taylor, S. H., Ripley, B. S., Woodward, F. I., & Osborne, C. P. (2011). Drought limitation of photosynthesis differs betwee C3 and C4 grass species in a comparative experiment. *PLant*, *Cell and Environment*, *34*, 65–75. https://doi.org/10.1111/j.1365-3040.2010.02226.x

Therneau, T. M. (2015) coxme: mixed effects cox models. R package version 2.2-5.

Weber, M. G., & Agrawal, A. A. (2012). Phylogeny, ecology, and the coupling of comparative and experimental approaches. *Trends in Ecology and Evolution*, *27*, 394–403.

Weir, J. T., & Lawson, A. (2015). Evoluitonary rates across gradients. *Methods in Ecology and Evolution*, *6*, 1278–1286. https://doi.org/10.1111/2041-210X.12419

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Freckleton RP, Rees M. Comparative analysis of experimental data. *Methods Ecol Evol*. 2019;10: 1308–1321. https://doi.org/10.1111/2041-210X.13164