doi:10.1111/j.1420-9101.2009.01757.x

REVIEW

The seven deadly sins of comparative analysis

R. P. FRECKLETON

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

Keywords:

comparative method;
generalized least squares;
independent contrasts;
missing data;
R.

Abstract

Phylogenetic comparative methods are extremely commonly used in evolutionary biology. In this paper, I highlight some of the problems that are frequently encountered in comparative analyses and review how they can be fixed. In broad terms, the problems boil down to a lack of appreciation of the underlying assumptions of comparative methods, as well as problems with implementing methods in a manner akin to more familiar statistical approaches. I highlight that the advent of more flexible computing environments should improve matters and allow researchers greater scope to explore methods and data.

Introduction

Many important problems in ecology, evolution and behaviour cannot readily be addressed using experimental approaches. Thus, whilst the fully randomized experiment is, for many, the ideal approach for hypothesis testing, alternative methods frequently have to be adopted (Maynard Smith, 1978; Harvey *et al.*, 1983). To address questions about long-term processes, observational and comparative approaches have been developed and are frequently employed (Maynard Smith, 1978; Felsenstein, 1988; Harvey & Pagel, 1991).

The comparative approach is used to test evolutionary hypotheses on datasets collected across multiple species. Trait or ecological data are collected for a group of species, and then statistical analysis is used to seek patterns consistent with alternative hypotheses (Clutton-Brock & Harvey, 1984; Harvey & Pagel, 1991; Harvey & Purvis, 1991). This is potentially an extremely powerful approach as the data collected usually span groups that encompass long periods of evolutionary change in a wide range of environmental conditions. The patterns examined in comparative analysis thus encompass very broad evolutionary processes. Comparative analyses allow macroevolutionary patterns to be explored, looking at the broad outcome of evolutionary processes across species to be examined (e.g. Harvey *et al.*, 1996). This contrasts with experimental approaches that focus typically on within-species microevolutionary processes.

As has been well discussed in the literature, comparative analyses have to deal with issues of phylogenetic nonindependence (Clutton-Brock & Harvey, 1984; Felsenstein, 1985; Harvey & Pagel, 1991; Garland *et al.*, 1992). That is, within a multi-species dataset species are related to each other to differing degrees and the degree of relatedness between species is often reflected in the amount of trait similarity. This happens because closely related species share more evolutionary history and have had less time to diverge than more distantly related ones.

The basis for many, if not most, comparative analyses is the analysis of associations between traits using correlation or regression. In this type of analysis if phylogenetic nonindependence is not accounted for then statistical analyses may be compromised (e.g. Harvey & Pagel, 1991; Martins & Garland, 1991) and results could be misleading. The consequences of ignoring nonindependence are numerous. For example, in simple bivariate analyses the type I error rate of significance tests will be inflated (Martins & Garland, 1991) as the variances of the traits will be incorrectly estimated. Similarly in analyses of trait differences between groups that differ in discrete characters, the effective sample sizes will be incorrect (the 'radiation principle' of Grafen, 1989). Alternatively, without accounting for evolutionary history differences in evolutionary trajectories between groups cannot be accounted for and will confound analyses: for example, Garland *et al.* (1999) and McKechnie *et al.* (2006) show

Correspondence: R. P. Freckleton, Department of Animal & Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK.
Tel.: +44 114 2220017; e-mail: r.freckleton@sheffield.ac.uk

that the slope of the relationship between basal metabolic rate and body size in birds is incorrect unless the split between passerines and nonpasserines is controlled for. The bottom line is that it is dangerous to ignore phylogenetic structure in data and in the same way that it is risky to ignore autocorrelation in time series or spatial data. As noted below, diagnosing the extent to which phylogeny is important is actually relatively straightforward.

A suite of comparative tests have been developed to deal with issues of nonindependence, and of these the most commonly employed are the method of independent contrasts (Felsenstein, 1985) and the method of generalized least squares (GLS) (Martins & Hansesn, 1997; Pagel, 1997, 1999; Garland *et al.*, 1999). Although formulated in different ways, these two approaches are essentially the same (see below) and take an underlying Brownian model of trait evolution to model the expected variance and co-variance of traits amongst species (e.g. Felsenstein, 1985; Pagel, 1997, 1999).

Because these statistical methods make assumptions about the underlying model of trait evolution, these translate into assumptions and predictions about the way that data should be distributed across species. If these do not hold then the tests may be compromised in some way. Accordingly a series of diagnostics designed for interpreting the results of phylogenetic analysis have been developed (Garland *et al.*, 1992; Purvis & Rambaut, 1995; Freckleton, 2000). In addition to the assumptions about the evolutionary process, more familiar assumptions such as homscedasticity and the distribution of residuals also apply.

Recently, there has been an increasing realization that the way that statistics is practiced in ecology and evolutionary biology may need to be thought. Some key issues to have emerged include looking at effect sizes rather than relying on *P*-value (Hilborn & Mangel, 1997; Burnham & Anderson, 2002; Paradis, 2005); allowing for model uncertainty and not simply relying on parameter uncertainty for testing models (e.g. Burnham & Anderson, 2002); and including multiple forms of uncertainty into models (e.g. Clark, 2007).

Unfortunately in some respects, the application of comparative methods has failed to keep up with these developments. This is particularly true of analyses that use comparative analyses to measure simple correlations and associations between traits using phylogenetic counterparts to conventional nonphylogenetic statistical methods. In this paper, I review seven areas in which current practices often lag behind statistical developments in other areas of ecology and evolution (summarized in Table 1). I highlight that one problem is a barrier between the users of phylogenetic methods and the techniques themselves. I believe that one reason for this is that previously users of phylogenetic methods have had to rely on relatively inflexible proprietary software packages. However, it is increasingly possible to conduct phylogenetic analysis in flexible computing environments such as R (R-Development-Core-Team, 2008; e.g. reviewed in Paradis, 2006) which is beginning to break such barriers down, and has been the medium for implementing new tools for comparative analyses (see Table 1).

Putting undue faith in models with low R^2

It has been pointed out that in many comparative analyses the proportion of variance explained (R^2) by statistical models is often very low, frequently in the range 0.05–0.10 (Moller & Jennions, 2002; Roslin, 2002; Jennions & Moller, 2003). This low explanatory power is in spite of overall model significance, or terms within models being statistically significant. The explanation is simply that in large datasets it is possible to detect very small effects. As comparative datasets become larger it is possible to pick up statistically significant effects of ever weaker signals, yielding models with statistically significant terms yet low overall explanatory power.

According to one view of the world all variables influence each other to some (often trivial) degree; thus if we are examining the effects of a set of predictors on a response variable the likelihood is that the effects will show a tapering distribution of effect sizes (Burnham & Anderson, 2002). Showing that the effect of one variable is statistically significant is not terribly informative; what is of more interest is the effect of that variable and how that ranks relative to others (Hilborn & Mangel, 1997; Nakagawa & Cuthill, 2007), as well as to examine how well the sizes and relative sizes of observed effects relate to what is expected from prior theory (Stephens *et al.*, 2005). Comparative analysis has been slow to recognize this (but see Paradis, 2005 for a discussion of the issue), and frequently relies on binary interpretations of *P*-values without reference to the broader importance of variables.

The R^2 has rightly been criticized as a 'goodness-of-fit' measure. The goodness-of-fit measured by the R^2 is specifically the variance explained relative to the variance in the raw data. It does not directly compare the relative fit of models, nor does it say anything about whether the fitted model is the 'best' in some sense. This is because a model with higher R^2 may not be 'better' than a model with a lower value, amongst other things because a model with more parameters will always have a better fit than a nested model with fewer. More generally there is no absolute criterion for determining goodness-of-fit, only measures of relative fit or support: thus we can only measure the fit of one model relative to another. Although the R^2 is not a general measure for determining goodness-of-fit, in analysing the output of models the R^2 can be useful as an indicator of lack of fit; as noted above models with poor R^2 values should not be over-interpreted. The reasons why a model yields a low R^2 are various. One reason is simply that there is a lot of

Table 1 Seven deadly sins of comparative analysis.

	Problem		Consequence	Solution	Software
1	Putting undue faith in models with low R^2	Models with low explanatory power may be statistically significant. This is often a consequence of large sample sizes and in practice the effects of variables included in models are weak	The importance of weak predictors may be over-emphasized; R^2 is not a reliable measure of fit or relative fit	Use effect sizes as well as significance tests. AIC is a better measure for comparing model fit. Low R^2 is a diagnostic of lack of model fit	MasterBayes; gee function in ape; pglm function in caicr
2	Reporting both PI and PC analysis	PI and PC make very different assumptions about the distribution of data, and are best regarded as alternative models for the same data. As such they should not be treated equally	Models with alternative assumptions are treated equally; potentially conflicting results may be reported	Check residuals and data for phylogenetic dependence; use a correction if appropriate	BayesTraits; geiger; pdap; caicr
3	Not testing distributional assumptions	Phylogenetically corrected models make assumptions about the distribution of residuals that are the same as those made in nonphylogenetic analysis and are well known	Parameter estimates may be incorrect or biased. Reported *P*-values may be incorrect	Use conventional regression diagnostics – check for linearity, normality of residuals and homogeneity of variance (all adjusted for phylogeny)	gee function in ape; caic/caicr; MasterBayes
4	Data dredging	In analyses, comparing a large number of predictors, best fit models are selected by comparing a large number of alternative models, or by using significance tests on parameters to distinguish models	High probability of type I errors; degenerate sampling distributions for parameters. Selected model is often no better than many possible alternatives. Outcome is highly sensitive to collinearity	Clearly identify hypotheses to be tested and test those. Report all stages in the model selection process. Use the full model when appropriate; when selection is necessary use model averaging or a multi-model approach	gee function in ape; caicr; MasterBayes
5	Treating residuals as data	Residuals from regressions of the response on confounding variables are used to control for unwanted effects in multi-variable regressions	Results in biases, particularly when the predictors are collinear	Use multipredictor analyses rather than univariate methods; do not use residuals in model fitting	gee function in ape; caicr; MasterBayes
6	Ignoring alternative models	Methods such as contrasts and GLS assume that residuals are distributed according to the predictions of a Brownian model of trait evolution. This may not be the cases and other processes may be operating	The phylogenetic correction may not be fully effective. The effects of important processes such as stablizing selection, varying rates of evolution or other factors shaping trait variation may be missed	Consider alternative models, such as OU model, δ, λ or κ transformations of Pagel (1997, 1999), or models incorporating rate variations	BayesTraits, geiger, laser, ape, ouch
7	Ignoring quality control of data	Data from disparate sources vary in quality and may be erroneous. Data may be missing for significant numbers of species	Low quality data will compromise statistical power. Missing data can lead to biases in the outcome of analyses	Employ quality criteria for data inclusion. Analyse data to determine whether data are missing randomly with respect to other variables. Consider imputation methods	MasterBayes

Current URLs for the software mentioned are given below.
ape: http://cran.r-project.org/web/packages/ape/
MasterBayes: http://cran.r-project.org/web/packages/MasterBayes/index.html
caic: http://www.bio.ic.ac.uk/Evolve/software/caic/index.html
caicr: http://r-forge.r-project.org/R/?group_id=140
laser: http://cran.r-project.org/web/packages/laser/index.html
geiger: http://cran.r-project.org/web/packages/geiger/index.html
pdap: http://mesquiteproject.org/pdap_mesquite/index.html
ouch: http://tsuga.biology.lsa.umich.edu/ouch/

random noise in the data, for example resulting from measurement error. In this case, the low R^2 should not deter us from interpreting the model. However, in the absence of other knowledge, when a low R^2 is returned it cannot be concluded with great certainty that other key variables have not been omitted, and the low explanatory power may be a consequence of a failure to include important predictors.

It is also worth noting that there are also circumstances under which high R^2 values can also be potentially misleading. For example, when data are examined on a logarithmic scale, if the scale encompasses several orders of magnitude, a high R^2 may be estimated even though there is considerable variance in the underlying untransformed data (Smith, 1980; Nee *et al.*, 2005). Nee *et al.* (2005) argue that this can lead to misleading interpretations of quantities such as life-history invariants.

Although we like to be reassured that our models provide as good a fit as possible to the data, and methods for testing and rejecting models do exist (e.g. Freckleton & Harvey, 2006; Rabosky, 2006), and we should be careful to not over-interpret models that have a high lack of fit, in many ways the focus on explanatory power is misleading: the question should be how the effect of variables of interest compares with other factors, and how this tallies with the predictions of theory. Thus, if hypotheses and alternative hypotheses are framed carefully, by comparing the fit of different models it should be possible to distinguish these and, using methods such as information theoretic or Bayesian methods, to quantify the relative weight of evidence in favour of each model (Hilborn & Mangel, 1997; Burnham & Anderson, 2002; Stephens *et al.*, 2005; Link & Barker, 2006). By using metrics such as effect sizes or relative weight of evidence of different models, it is possible to test hypotheses without using the R^2 for a purpose for which it is unsuited.

Reporting both PI and PC analyses

Frequently, both across-species and phylogenetically corrected analyses of the same data are reported simultaneously. This is despite the fact that the two forms of analysis make very different assumptions about the distribution of the data. Because of this, if one of the analyses is valid in terms of the data and meets the assumptions of the analysis, the other will not be. A number of methods exist for diagnosing and controlling for phylogenetic nonindependence (Lynch, 1991; Pagel, 1999; Freckleton *et al.*, 2002; Blomberg *et al.*, 2003; Houswarth *et al.*, 2004). In the specific case of distinguishing between phylogenetic analysis using contrasts or GLS and simple across-species analysis, this is readily done by comparing likelihoods of the alternative models (Pagel, 1999; Freckleton *et al.*, 2002).

Three arguments may be used to justify the use of both types of analysis on the same data. First, it is argued that across-species analysis reveals different factors, such as ecological ones, compared with phylogenetic analysis which reveals historical ones. However, this argument is generally not accepted, following heated debate (Harvey *et al.*, 1995a,b; Westoby *et al.*, 1995). In purely statistical terms, it is clear that across-species analyses should at least be regarded with suspicion if data show evidence of nonindependence in any case (Martins & Garland, 1991).

The second argument is that in some models of trait evolution, across-species analysis performs better statistically, even if data show strong phylogenetic dependence (Price, 1997; Harvey & Rambaut, 2000; Freckleton & Harvey, 2006). However, these models are rather specialized models referring to ecological traits in closely related species occupying a confined niche space in an adaptive radiation. They do not apply if species are not connected ecologically, and for instance would not apply to many of the very broad datasets frequently analysed that encompass large numbers of species across biogeographically widely separated areas. Moreover, the diagnostics now exist with which to detect such processes, if they are suspected (Harmon *et al.*, 2003; Freckleton & Harvey, 2006).

A third argument is that presenting both an across-species and a phylogenetically corrected analysis is a 'belt-and-braces' approach to analysis. This would be particularly the case if both analyses yield comparable results, for example in indicating whether the effect of a particular predictor in a linear model were significant or not. It may not seem to matter a great deal in such a case if both analyses are reported, and a reader may be comforted that the result obtained is robust to making contrasting assumptions about the structure of the data. The problem arises, of course, when the two analyses do not agree and there is a difference between them. In this case, the only course of action is to use an index of phylogenetic dependence to try to distinguish the models. For example the λ statistic of Pagel (1997, 1999) is a straightforward way to do this. This index varies between 0 (phylogenetic independence) and 1 (traits covary as assumed by the Brownian model). In a linear modelling context this parameter allows varying levels of phylogenetic dependence in the model residuals to be modelled, and is readily estimated via maximum likelihood (Freckleton *et al.*, 2002). The maximum likelihood value might be zero (nonphylogenetic analysis should be preferred), 1 (Brownian model preferred) or an intermediate value reflecting a more complex pattern of trait distribution. Using several freely available R-packages, this parameter can be estimated 'in the background' with minimum additional effort.

The problem of presenting both across-species and phylogenetic analyses can be made clearer by analogy with the problem of whether to transform data or not in a conventional parametric analysis: one is often faced with the problem of whether to do this or not, but it would not usually be the case that one would present in a publication the results of analyses of the same response variable both with and without transformation. Instead one would use a combination of diagnostics and maximum likelihood (e.g. Box–Cox transformation) to decide which analysis is better justified. It would seem sensible to apply the sample type of quantitative criteria to decide between phylogenetic and nonphylogenetic models.

Not testing distributional assumptions

It is well appreciated that in conventional statistical analysis specific assumptions about the distribution of data are made and should be tested (e.g. Grafen & Hails, 2004). Apart from the issue of independence, these include the shape of residual distributions, linearity (in linear models) and homogeneity of variance (or more generally, that the residual variation is distributed as it should be as the predictors change). Such assumptions should also hold in phylogenetic comparative analyses, otherwise results of statistical tests may be invalid.

In analyses using PICs or GLS, this is very readily done and those developing methods have often been clear on how to do this (e.g. see http://www.bio.ic.ac.uk/Evolve/software/caic/assumptions.html), although this does not seem to be widely appreciated. In a PIC analysis (assuming that the data have not been regressed through the origin) residuals from fitted models can be interpreted and tested under the same assumptions as any other model (Garland *et al.*, 1992). In a GLS analysis, Garland & Ives (2000) present a simple and elegant calculation for generating normalized residuals from a GLS model. Specifically, these residuals have the expected property of being normally, independently distributed with constant variance and zero mean. These residuals can be analysed using the conventional diagnostics used in nonphylogenetic linear models.

Data dredging

One of the biggest criticisms of the comparative approach is that because data are not generated experimentally, the relationships found are only correlative and may not stand up to further examination. The possibility of hidden variables or type I error, for example, mean that significant correlations may be mistakenly taken as implying a causal relationship when in fact no such relationship exists. Moreover, if a large number of potential explanatory variables are used these problems may be exacerbated, and subjectivity may enter into the decision as to which results to present and which subsequent tests to perform. The net result of this is the phenomenon of 'data dredging', whereby statistical tests are presented as if they are conducted independently and objectively whereas in reality they are not. The danger here is particularly that the importance of some predictors is overstated: so unimportant variables may be reported as 'significant', when they are not, or the effects of weak predictors are over estimated. This problem is a well-known phenomenon (Burnham & Anderson, 2002).

There are three solutions to these problems. First, authors need to be explicit about which variables have been analysed, which hypotheses were anticipated to be tested *a priori*, and how the tests reported are chosen (and which were not) (Burnham & Anderson, 2002; Stephens *et al.*, 2005). Second, techniques such as model averaging should be considered and formulated appropriately (Burnham & Anderson, 2002; Link & Barker, 2006). Model averaging is a particularly attractive approach as it involves fitting all models that are biologically plausible, assigning to each model a weight that depends on its relative fit, then basing inference on this set of weights. There is no need for selection and inclusion or exclusion of models using this approach. Third, when large amounts of data have been collected with few *a priori* expectations of how different variables will relate to each other, statistical testing may not be appropriate and alternative techniques such as data mining should be considered (see Kantarszic, 2002 for an overview of techniques).

In essence, the problem is that too frequently the exploratory analyses that precede the analysis presented are not reported or considered as being part of the overall analysis (Stephens *et al.* 2005). When considering large complex datasets, it is essential that all steps in the analysis process are considered and that the consequences of decisions made in the exploratory phase are understood as there is evidence that this can lead to biased reporting of results in the literature (Ridley *et al.*, 2007).

Treating residuals as data

One of the most common problems in comparative analyses is the use of residuals as data. This arises particularly when researchers wish to control for one variable, particularly body mass, whilst analysing the relationship between other variables. The approach commonly used is to regress the trait of interest on body mass, take the residuals from this regression and use these as data in the subsequent analysis. Despite warnings from the literature (e.g. Garcia-Berthou, 2001; Freckleton, 2002), this approach nevertheless continues to be used in numerous studies.

The problem with doing this is that if the variable controlled for covaries with other variables in the analysis, then subsequent analyses will be biased (Freckleton, 2002). The obvious, and straightforward, way to deal with this is to use a multiple regression/linear model in which the confounding variable is used to control for unwanted effects. Subsequently, if required, alternative decomposition of variance can be used to eliminate variables in different orders. For example, if we have two predictors, then the model sequential sums of squares can be examined to see whether the variance explained by predictor 1 is dependent or not of predictor 2 in the model, and to decide whether the ordering of the variables is important. Detailed examples of this process, and consequences for interpretation, are given by Grafen & Hails (2004).

One argument frequently used in favour of using residuals to control effects is that this decomposition may better reflect the order in which causal variables act. This is not correct, however. To illustrate this, consider a response variable y which is generated as a function of x_1 and x_2. The variables act sequentially, such that y is generated in the following way. First x_1 acts, generating a first prediction:

$$y_1 = a_1 + b_1 x_1 + e \tag{1}$$

In eqn 1 e is the variance unexplained by x_1, i.e. the residual term. This is then acted upon by x_2, i.e. $e = b_2x_2 + e'$, such that the effects of $b_1$and b_2 are sequential and independent. The net model then becomes:

$$y_1 = a_1 + b_1 x_1 + b_2 x_2 + e' \tag{2}$$

It is straightforward to see that the order of x_1 and x_2 in this model is irrelevant, and that eqn 2 would be the same irrespective of whether x_1 acts before x_2 or vice versa. Consequently, there is no argument for using residual regression to estimate the effects of x_2 and x_1 separately.

Ignoring alternative models

In analysing comparative data, the bulk of studies have used the method of contrasts to look for simple correlations between pairs of variables. For example, the most commonly employed software for conducting comparative analysis by contrasts (CAIC; Purvis & Rambaut, 1995) has been used in over 700 studies at the time of writing (source: Thomson ISI). There are two issues with this reliance on one technique, however. First, the method assumes a specific evolutionary model for the data; and second, this is often the simplest of a suite of more complex statistical models.

Increasingly more sophisticated evolutionary models are being used to analyse comparative data (Hansen, 1997; Pagel, 1997, 1999; Garland *et al.*, 1999; Thomas *et al.*, 2006; Ives *et al.*, 2007; Felsenstein, 2008; Hansen *et al.*, 2008; reviewed by Freckleton & Pagel, 2009; Revell & Collar, 2009). These models can incorporate a range of processes, such as punctuational and speciational evolution, accelerating and decelerating rates of evolution and adaptive variation. These models allow data to be explored and more sophisticated interpretations of data to be developed than to simply ask whether traits are correlated or not, as is frequently done.

A second way in which it is possible to develop more sophisticated analyses is via the use of more flexible statistical models. The method of contrasts is identical to the method of GLS (Pagel, 1997, 1999; Garland *et al.*, 1999; Freckleton & Jetz, 2009). However, by expressing statistical problems in the form of linear models it is possible to generate less restrictive models that more realistically reflect the structure of the data. For instance software that performs GLS analysis will typically allow a model to be written as a compact formula, combining both continuous and categorical predictors. The same analysis using independent contrasts would require dummy coding of some variables (e.g. for multi-level factors and interaction terms) making the process of model fitting more laborious and less transparent.

Ignoring quality control of data

Comparative analysis frequently relies on data that have been collated from various sources across a number of different trait variables and using the phylogenetic information at hand. Although many authors take great trouble in describing the process of data collation, this is not always done. There are also issues in data quality that are not always dealt with and here I wish to discuss three, the quality of the trait data, missing data and robustness of the phylogenetic data.

In a review of databases on primate body size data, Smith & Jungers (1997) pointed out that many of the data used in comparative analyses are collected in a somewhat haphazard manner. They revisited several existing published databases and found that the reporting of the process of how the data were collected was patchy to the point where it was not clear where data had come from. For example during the evolution of one dataset in successive publications, the meaning of data had become distorted with initially species means substituting for genus means, then back again in subsequent papers with the consequence that species were in the final version assigned ‘mean’ values even though they had never been measured; or in some cases there is insufficient checking of data, for example they cite the example of human body mass, which in one database was represented by a value appropriate for pygmies.

In terms of data quality one important issue is that often in comparative analyses, it is necessary to use constructed variables that are indices or scores. This is inevitable as for many important variables it is impossible to obtain data from the literature that have been measured in the same way in all studies. In such cases, it is important to ensure that the index or score constructed is meaningful and robust to alternative formulations (e.g. see Olson *et al.*, 2008 for an example).

A major problem with data quality is that it is rarely the case that data are available for all species within a clade. Usually data are missing for a proportion of species, often from just one or two variables per species. This generates several problems. The usual way to deal with missing data is via case-wise deletion. Thus, all cases containing missing entries are simply removed from the dataset prior to analysis. For model selection or comparison the set of included cases should be the same for all models considered, so included cases has to be based on the set of complete variables for the maximal model. This can either lead to the loss of quite considerable proportions of cases, reducing sample sizes, or restrict the set of predictors that can be included in the analysis. In either case the resultant analysis can be compromised.

Case-wise deletion is only a valid course of action under certain conditions. This is the case only when data are missing completely at random, which is when the probability of a value being absent is not related to any other variable in the analysis, or indeed to phylogenetic position. The problems of missing data in evolutionary biology have been reviewed recently by Nakagawa & Freckleton (2008). In essence the problem is that if data are not missing completely at random, but are missing nonrandomly with respect to one or more of the other variables being considered, then there is a strong possibility that results will be biased. It is easy to see why this may be the case in a comparative analysis. For example, if species are more likely to have missing data on life-history variables because they are small or because they are rare, the sample of included species will be biased with respect to these variables.

If the phylogeny is available for the whole clade then it should be relatively straightforward to test whether missingness is a function of any other variable. Missingness can be coded as a binary variable, then tested for phylogenetic position or correlated with other variables. Any significant correlate of missingness should be taken seriously and subsequent results interpreted with caution. Techniques exist for imputing missing values and dealing with some forms of missing data (see Nakagawa & Freckleton, 2008). Fisher *et al.* (2003), for example, used this approach in a phylogenetic comparative analysis to deal with missing data.

Extinction of species during the course of evolution can generate problems akin to those of missing data. If extinction is higher for species with given traits then the results of analyses can be affected (Maddison, 2006; Freckleton *et al.*, 2008). The difference here is that extinct species are usually not included within the working phylogeny and hence this effect is difficult to test for. However, approaches are beginning to be developed (e.g. Bokma, 2008; Paradis, 2008), although it is too early to judge how successful these are, or whether biased extinction generates widespread problems for comparative analyses.

The final issue with data quality concerns the robustness of the phylogeny. Commonly the limiting factor in conducting comparative analyses has been the availability of phylogenetic information (Harvey *et al.*, 1995b). In many cases, a phylogeny is not available and hence a taxonomy has been substituted, which is used to generate a branching tree for the group examined. Sometimes more than one tree is available, or alternative resolutions exist for trees in which there is great uncertainty at some of the nodes. The first point to make is that, whatever the quality of the phylogeny, a test for phylogenetic dependence can be used to determine whether the phylogeny improves the quality of the statistical model. Thus, even a rough tree based on a taxonomy should be preferred over a nonphylogenetic model if diagnostics indicate that the resultant model is preferable. Second, as suggested by Pagel (1993), the uncertainty in the phylogenetic model can potentially be reduced by using an estimated GLS approach (EGLS), whereby areas of phylogenetic uncertainty are removed by resolving the phylogeny matrix on a variable known to contain high levels of phylogenetic information. For example, if we have a polytomy connecting a set of species, and we have body size information for those, body size is known to contain strong phylogenetic signal and could be used to generate a resolution (Pagel, 1993). Finally, increasingly it is possible to quantify the uncertainty in phylogenetic reconstructions and to directly incorporate this into comparative tests using Bayesian methods (e.g. Huelsenbeck, 2000).

Concluding remarks

The aim of this paper was to highlight some areas where comparative analyses are currently lagging behind statistical practice in other areas of ecology and evolutionary biology. This may be partly because methods for analysing comparative data have been formulated in a different way and authors are unclear on the assumptions. A critical limitation in the past is that researchers have had to rely on relatively inflexible bespoke packages for applying individual tests. With the increasing availability of software for R via contributed packages (e.g. Paradis *et al.*, 2004), together with detailed instructional texts (Paradis, 2006), data can be analysed in a more interactive environment, and using techniques in ways that are more akin to those used in conventional analyses, This offers a great deal of hope for those conducting comparative analyses and should see these techniques evolve and become more sophisticated in the near future.

Acknowledgments

RPF is funded by a Royal Society University Research Fellowship. Thanks to Tom Webb and Niclas Kolm for comments on a draft of this MS and to Niclas for encouraging me to write this.

References

Blomberg, S.P., Garland, T. & Ives, A.R. 2003. Testing for phylogenetic signal in comparative data: behavioural traits are more labile. *Evolution* **57**: 717–745.

Bokma, F. 2008. Detection of ''punctuated equilibrium'' by Bayesian estimation of speciation and extinction rates, ancestral character states and rates of anagenetic and cladogenetic evolution on a molecular phylogeny. *Evolution* **62**: 2718–2726.

Burnham, K.P. & Anderson, D.R. 2002. *Model Selection and Multimodel Inference*. Springer-Verlag, New York.

Clark, J.S. 2007. *Models for Ecological Data*. Princeton University Press, Princeton, NJ.

Clutton-Brock, T.H. & Harvey, P.H. 1984. Comparative approaches to investigating adaptation. In: *Behavioural Ecology: An Evolutionary Approach* (J.R. Krebs & N.B. Davies, eds), pp. 7–29. Blackwell Scientific Publications, Oxford.

Felsenstein, J. 1985. Phylogenies and the comparative method. *Am. Nat.* **126**: 1–25.

Felsenstein, J. 1988. Phylogenies and quantitative characters. *Annu. Rev. Ecol. Syst.* **19**: 445–471.

Felsenstein, J. 2008. Comparative methods with sampling error and within-species variation: contrasts revisited and revised. *Am. Nat.* **171**: 713–725.

Fisher, D.O., Blomberg, S.P. & Owens, I.P.F. 2003. Extrinsic versus intrinsic factors in the decline and extinction of Australian marsupials. *Proc. R. Soc. Lond. B Biol. Sci.* **270**: 1801–1808.

Freckleton, R.P. 2000. Phylogenetic tests of ecological and evolutionary hypotheses: checking for phylogenetic independence. *Funct. Ecol.* **14**: 129–134.

Freckleton, R.P. 2002. On the misuse of residuals in ecology: regression of residuals vs. multiple regression. *J. Anim. Ecol.* **71**: 542–545.

Freckleton, R.P. & Harvey, P.H. 2006. Non-Brownian trait evolution in adaptive radiations. *PLoS Biol.* **4**: 2104–2111.

Freckleton, R.P. & Jetz, W. 2009. Space versus phylogeny: disentangling phylogenetic and spatial signals in comparative data. *Proc. R. Soc. Lond. B Biol. Sci.* **276**: 21–30.

Freckleton, R.P. & Pagel, M. 2009. Recent advances in comparative methods. *Social Behaviour: Genes, Ecology and Behaviour* (eds, T. Székely, A.J. Moore & J. Komdeur), in press.

Freckleton, R.P., Harvey, P.H. & Pagel, M. 2002. Phylogenetic analysis and comparative data: a test and review of evidence. *Am. Nat.* **160**: 712–726.

Freckleton, R.P., Phillimore, A.B. & Pagel, M. 2008. Relating traits to diversification: a simple test. *Am. Nat.* **172**: 102–115.

Garcia-Berthou, E. 2001. On the misuse of residuals in ecology: testing regression residuals vs. the analysis of covariance. *J. Anim. Ecol.* **70**: 708–711.

Garland, T.J. & Ives, A.R. 2000. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. *Am. Nat.* **155**: 346–364.

Garland, T.J., Harvey, P.H. & Ives, A.R. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Syst. Biol.* **41**: 18–32.

Garland, T.J., Midford, P.E. & Ives, A.R. 1999. An introduction to phylogenetically-based statistical methods, with a new method for confidence intervals based on ancestral values. *Am. Zool.* **39**: 374–388.

Grafen, A. 1989. The phylogenetic regression. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **326**: 119–157.

Grafen, A. & Hails, R. 2004. *Modern Statistics for the Life Sciences*. Oxford University Press, Oxford.

Hansen, T.F. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution* **51**: 1341–1351.

Hansen, T.F., Pienaar, J. & Orzack, S.H. 2008. A comparative method for studying adaptation to a randomly evolving environment. *Evolution* **62**: 1965–1977.

Harmon, L.J., Schulte, J.A., Larson, A. & Losos, J.B. 2003. Tempo and model of evolutionary radiation in Iguanan lizards. *Science* **301**: 961–964.

Harvey, P.H. & Pagel, M.D. 1991. *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford.

Harvey, P.H. & Purvis, A. 1991. Comparative methods for explaining adaptations. *Nature* **351**: 619–624.

Harvey, P.H. & Rambaut, A. 2000. Comparative analyses for adaptive radiations. *Philosophical Transactions of the Royal Society Series B* **355**: 1599–1606.

Harvey, P.H., Colwell, R.K., Silvertown, J. & May, R.M. 1983. Null models in ecology. *Annu. Rev. Ecol. Syst.* **14**: 189–211.

Harvey, P.H., Read, A.F. & Nee, S. 1995a. Further remarks on the role of phylogeny in comparative ecology. *J. Ecol.* **83**: 733–734.

Harvey, P.H., Read, A.F. & Nee, S. 1995b. Why ecologists need to be phylogenetically challenged. *J. Ecol.* **83**: 535–536.

Harvey, P.H., Leigh Brown, A.J., Maynard Smith, J. & Nee, S., eds. 1996. *New Uses for New Phylogenies*. Oxford University Press, Oxford.

Hilborn, R. & Mangel, M. 1997. *The Ecological Detective: Confronting Models with Data*. Princeton University Press, Princeton, NJ.

Housworth, E.A., Martins, E.P. & Lynch, M. 2004. The phylogenetic mixed model. *Am. Nat.* **163**: 84–96.

Huelsenbeck, J.P. 2000. Accommodating phylogenetic uncertainty in evolutionary studies. *Science* **288**: 2349–2350.

Ives, A.R., Midford, P.E. & Garland, T. 2007. Within-species variation and measurement error in phylogenetic comparative methods. *Syst. Biol.* **56**: 252–270.

Jennions, M.D. & Moller, A.P. 2003. A survey of the statistical power of research in behavioural ecology and animal behaviour. *Behav. Ecol.* **14**: 438–445.

Kantarszic, M. 2002. *Data Mining: Concepts, Models, Methods and Algorithms*. Wiley-Blackwell, Oxford.

Link, W.A. & Barker, R.J. 2006. Model weights and the foundations of multimodel inference. *Ecology* **87**: 2626–2635.

Lynch, M. 1991. Methods for the analysis of comparative data in evolutionary biology. *Evolution* **45**: 1065–1080.

Maddison, W.P. 2006. Confounding asymmetries in evolutionary diversification and character change. *Evolution* **60**: 1743–1746.

Martins, E.P. & Garland, T. 1991. Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. *Evolution* **45**: 534–557.

Martins, E.P. & Hansesn, T.F. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of inter-specific data. *Am. Nat.* **149**: 646–667.

Maynard Smith, J. 1978. Optimization theory in evolution. *Annu. Rev. Ecol. Syst.* **9**: 31–56.

McKechnie, A.E., Freckleton, R.P. & Jetz, W. 2006. Phenotypic plasticity in the scaling of avian basal metabolic rate. *Proc. R. Soc. Lond. B Biol. Sci.* **273**: 931–937.

Moller, A.P. & Jennions, M.D. 2002. How much variance can be explained by ecologists and evolutionary biologists? *Oecologia* **132**: 492–500.

Nakagawa, S. & Cuthill, I.C. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol. Rev.* **82**: 591–605.

Nakagawa, S. & Freckleton, R.P. 2008. Missing inaction: the dangers of ignoring missing data. *Trends Ecol. Evol.* **23**: 592–596.

Nee, S., Colgreave, N., West, S.A. & Grafen, A. 2005. The illusion of invariant quantities in life histories. *Science* **309**: 1236–1239.

Olson, V.A., Liker, A., Freckleton, R.P. & Székely, T. 2008. Parental conflict in birds: comparative analyses of offspring development, ecology and mating opportunities. *Proc. R. Soc. Lond. B Biol. Sci.* **275**: 301–307.

Pagel, M. 1993. Seeking the evolutionary regression coefficient: an analysis of what comparative methods measure. *J. Theor. Biol.* **164**: 191–205.

Pagel, M. 1997. Inferring evolutionary processes from phylogenies. *Zool. Scr.* **26**: 331–348.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. *Nature* **401**: 877–884.

Paradis, E. 2005. Statistical analysis of diversification with species traits. *Evolution* **59**: 1–12.

Paradis, E. 2006. *Analysis of Phylogenetics and Evolution with R*. Springer, New York.

Paradis, E. 2008. Asymmetries in phylogenetic diversification and character change can be untangled. *Evolution* **62**: 241–247.

Paradis, E., Claude, J. & Strimmer, K. 2004. APE: analyses of phylogenetic and evolution in R language. *Bioinformatics* **20**: 289–290.

Price, T. 1997. Correlated evolution and independent contrasts. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **352**: 519–529.

Purvis, A. & Rambaut, A. 1995. Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. *Comput. Appl. Biosci.* **11**: 247–251.

Rabosky, D.L. 2006. Likelihood methods for inferring temporal shifts in diversification rates. *Evolution* **60**: 1152–1164.

R-Development-Core-Team 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Revell, L.J. & Collar, D.C. 2009. Phylogenetic analysis of the evolutionary correlation using likelihood. *Evolution* **63**: 1090–1100.

Ridley, J., Kolm, N., Freckleton, R.P. & Gage, M.J.G. 2007. An unexpected influence of widely used significance thresholds on the distribution of reported P-values. *J. Evol. Biol.* **20**: 1082–1089.

Roslin, T. 2002. Explaining a little is often a lot. *Trends Ecol. Evol.* **17**: 498.

Smith, R.J. 1980. Rethinking allometry. *J. Theor. Biol.* **87**: 97–111.

Smith, R.J. & Jungers, W.L. 1997. Body mass in comparative primatology. *J. Human Evol.* **32**: 523–559.

Stephens, P.A., Buskirk, S.W., Hayward, G.D. & Martinez Del Rio, C. 2005. Information theory and hypothesis testing: a call for pluralism. *J. Appl. Ecol.* **42**: 4–12.

Thomas, G., Freckleton, R.P. & Szekely, T. 2006. Comparative analyses of the influence of developmental mode on phenotypic diversification rates in shorebirds. *Proc. R. Soc. Lond. B Biol. Sci.* **273**: 1619–1624.

Westoby, M., Leishman, M.R. & Lord, J.M. 1995. On misinterpreting the ‘phylogenetic correction’. *J. Ecol.* **83**: 531–534.

Received 26 January 2009; revised 25 March 2009; accepted 30 March 2009