Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2014, **5,** 961–970 doi: 10.1111/2041-210X.12232

Imputation of missing data in life-history trait datasets: which approach performs the best?

Caterina Penone[1]*, Ana D. Davidson[2,3], Kevin T. Shoemaker[2], Moreno Di Marco[4], Carlo Rondinini[4], Thomas M. Brooks[5], Bruce E. Young[3], Catherine H. Graham[2] and Gabriel C. Costa[1]

[1]*Departamento de Ecologia, Universidade Federal do Rio Grande do Norte, Natal, Brasil;* [2]*Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY, USA;* [3]*NatureServe, Arlington, VA, USA;* [4]*Global Mammal Assessment Program, Department of Biology and Biotechnologies, Sapienza Università di Roma, Rome, Italy; and* [5]*International Union for Conservation of Nature, 28 rue Mauverney, Gland 1196, Switzerland*

Summary

1. Despite efforts in data collection, missing values are commonplace in life-history trait databases. Because these values typically are not missing randomly, the common practice of removing missing data not only reduces sample size, but also introduces bias that can lead to incorrect conclusions. Imputing missing values is a potential solution to this problem. Here, we evaluate the performance of four approaches for estimating missing values in trait databases (K-nearest neighbour (kNN), multivariate imputation by chained equations (mice), missForest and Phylopars), and test whether imputed datasets retain underlying allometric relationships among traits.

2. Starting with a nearly complete trait dataset on the mammalian order Carnivora (using four traits), we artificially removed values so that the percent of missing values ranged from 10% to 80%. Using the original values as a reference, we assessed imputation performance using normalized root mean squared error. We also evaluated whether including phylogenetic information improved imputation performance in kNN, mice, and missForest (it is a required input in Phylopars). Finally, we evaluated the extent to which the allometric relationship between two traits (body mass and longevity) was conserved for imputed datasets by looking at the difference (bias) between the slope of the original and the imputed datasets or datasets with missing values removed.

3. Three of the tested approaches (mice, missForest and Phylopars), resulted in qualitatively equivalent imputation performance, and all had significantly lower errors than kNN. Adding phylogenetic information into the imputation algorithms improved estimation of missing values for all tested traits. The allometric relationship between body mass and longevity was conserved when up to 60% of data were missing, either with or without phylogenetic information, depending on the approach. This relationship was less biased in imputed datasets compared to datasets with missing values removed, especially when more than 30% of values were missing.

4. Imputations provide valuable alternatives to removing missing observations in trait databases as they produce low errors and retain relationships among traits. Although we must continue to prioritize data collection on species traits, imputations can provide a valuable solution for conducting macroecological and evolutionary studies using life-history trait databases.

Key-words: Phylopars, missForest, kNN, multivariate imputation by chained equations, phylogeny, carnivores, root mean squared error, body mass, longevity

Introduction

Trait-based analyses are widely used to address ecological and evolutionary processes from local to global scales (Lavorel & Garnier 2002; Graham *et al.* 2012; Huang, Stephens & Gittleman 2012). Life-history traits used in macroecological studies generally include physical characteristics (e.g., body mass or body length) and reproductive parameters (e.g., litter size or weaning age). The incorporation of trait dimensions of biodiversity in ecological studies is important for a range of research areas, including ecosystem functioning (Loreau 2010), community ecology (McGill *et al.* 2006), population ecology (Santini *et al.* 2013), extinction risk assessments (Pacifici *et al.* 2013), and conservation (Cardillo *et al.* 2008). Trait-based approaches often use large, ecoinformatic databases that require enormous efforts to compile data from primary literature or field observations. Often, trait data are scarce because many species are rare, cryptic, or occur in remote locations (Nakagawa & Freckleton 2008). Consequently, trait databases suffer from a chronic problem of missing data.

A common practice is to use only species, traits or locations for which complete data are available and ignore those for which some values are missing (e.g., Junker *et al.* 2013). When data are missing completely at random (MCAR) the

*Correspondence author. E-mail: caterina.penone@gmail.com

consequence of removing incomplete observations is a decrease in statistical power, due to decreased sample size (Nakagawa & Freckleton 2008). However, real data are usually missing at random (MAR). Data are considered to be MAR when the presence of missing data for a given variable (e.g., gestation length) is related to the values of another variable (e.g., body mass). In this case, values of missing data in both these variables can be inferred based on other variables in the dataset (Rubin 1976). When data are MAR, deleting missing values can lead to misleading results in comparative studies or biased estimates of evolutionary parameters (Hadfield 2008; González-Suárez, Lucas & Revilla 2012; Pakeman 2014). Taxonomic and phylogenetic bias is commonplace in trait databases; more information is available for charismatic or otherwise well-studied taxonomic groups (e.g., carnivores). Also, data on threatened or extinct species are generally less complete (Fisher, Blomberg & Owens 2003). But the prevalence of missing values can also depend on the traits themselves: in mammals, charismatic large-bodied species with large geographic ranges and long lifespans usually have more data entries (González-Suárez, Lucas & Revilla 2012). Finally, some traits are more complete because they are easier to measure (e.g., morphology). These issues in data availability can bias parameter estimates from models, potentially leading to erroneous conclusions (Nakagawa & Freckleton 2008). Some approaches deal with missing values in data analyses using phylogenetic comparative approaches (Cardillo *et al.* 2004), but do not fully address the issue of losing statistical power as rows with missing values are excluded from the analyses. Finding further solutions to address the problem of missing data is therefore important for improving our understanding of biological processes.

Many statistical alternatives to data deletion are available to impute values that are MAR using the other variables as predictors. The main strategies to substitute missing values include single, multiple or likelihood-based imputations. Many of these tools have been designed, tested and used in medical, biological or social sciences (Troyanskaya *et al.* 2001), but their use for trait datasets is relatively recent (Fisher, Blomberg & Owens 2003; Paine *et al.* 2011; Di Marco *et al.* 2012; Shan *et al.* 2012; Taugourdeau *et al.* 2014). Evaluating, comparing and examining the performance of several approaches for handling missing data in trait databases is a key consideration in selection of the appropriate software package for implementation of a given analysis (Joppa *et al.* 2013).

Our aim here is to determine whether imputation approaches can be used to impute traits, and specifically the relative performance of imputation approaches and how imputation alters relationships among traits. We tested four imputation approaches: K-nearest neighbour (kNN), multivariate imputation by chained equations (mice), a random forest technique (missForest), and an approach based on maximum likelihood that uses phylogenetic information (Phylopars). We performed our tests with a complete trait dataset where we artificially removed different percentages of missing values. We then evaluated the performance of the approaches by comparing the original values (from the complete dataset) and the imputed values. In order to assess if the imputed datasets recovered original biological patterns, we evaluated whether the allometric relationship between two traits changed in imputed datasets with increasing percentages of missing values.

Phylogenetic information can improve the estimation of missing trait values in the imputation process (Fisher, Blomberg & Owens 2003; Cardillo *et al.* 2008; Guénard, Legendre & Peres-Neto 2013; Swenson 2014) because closely related species tend to be more similar to each other (Pagel 1999) and many traits display high degrees of phylogenetic signal (Blomberg, Garland & Ives 2003). We tested whether including phylogenetic information improved the imputation process. Commonly, taxonomic ranks are used to impute trait data (e.g., Ter Steege *et al.* 2006), but this approach is subject to how rank is defined and does not use the full information available in a phylogeny (Swenson 2014). Here, we refine previous approaches based on taxonomy to include branch lengths from phylogenetic trees in the imputation process. We expected that adding phylogenetic information would improve our ability to accurately impute values, especially for traits with higher phylogenetic signal.

Methods

IMPUTATION APPROACHES

We evaluated four imputation methods that span a range of computation approaches and have been shown to perform better than other approaches in comparisons using non-trait data. Single imputations are the simplest approaches and replace missing values by a single value, without any estimate of the uncertainty of the imputation. Single imputations have been shown to be accurate for datasets with small percentages of missing values (Schafer 1999). Among single imputation approaches, the k-Nearest Neighbour (kNN) is one of the most precise (Troyanskaya *et al.* 2001). We used the R package 'VIM', kNN function (Templ *et al.* 2013) that allows control of some imputation parameters, such as using either the mean or median value for imputation. We also evaluated multiple imputations that take into account the imputation uncertainty by running single imputation multiple times and therefore may provide a more precise estimate of missing data. These approaches impute incomplete datasets n times and analyse the n imputed datasets using standard analytical methods. The n results of the analyses are then pooled in one final result that gives the uncertainty of the estimates (Nakagawa & Freckleton 2008). We chose Multivariate Imputation with Chained Equations, as implemented by the 'mice' package for R (van Buuren & Groothuis-Oudshoorn 2011), because it has smaller error and bias as compared to other multiple imputation approaches (Ambler, Omar & Royston 2007; test with medical data). Among the different possibilities of multiple imputation using mice, we chose predictive mean matching because it preserves non-linear relationships (van Buuren & Groothuis-Oudshoorn 2011), which occur in trait datasets (Santini *et al.* 2013). This method is the most frequently used in previous imputations of trait data (Fisher, Blomberg & Owens 2003; Baraloto *et al.* 2010; Paine *et al.* 2011; Di Marco *et al.* 2012). Imputations based on random forest algorithms can also be a valuable alternative, as they have been shown to be highly accurate and require little computational time (Pantanowitz & Marwala 2009). Random

forests are machine-learning techniques that grow many decision trees and output the clustering that appears most often in the individual trees (Breiman 2001). These approaches can deal with highly dimensional data, do not rely on distributional assumptions and are particularly appropriate for modelling complex interactions and non-linear relationships among variables. We used the 'missForest' package in R (Stekhoven & Bühlmann 2012) that has been shown to perform better than other approaches with various dataset types (Stekhoven & Bühlmann 2012), and has recently been used to impute mammal trait data by Verde Arregoitia, Blomberg & Fisher (2013). Finally, we tested a novel likelihood-based approach that estimates missing parameters using both phylogeny and allometric relationships among traits: Phylopars (Bruggeman, Heringa & Brandt 2009). This approach showed promising results for missing data estimation in traits (González-Suárez, Lucas & Revilla 2012; Riek & Bruggeman 2013), but has not yet been compared to other approaches. Phylopars uses a phylogenetic variance-covariance matrix, which is a component of phylogenetic generalized linear models (Bruggeman, Heringa & Brandt 2009). Phylopars, missForest and kNN can all be considered as single imputations because a single value is imputed for each missing datum. More details on the approaches are given in supporting information Data S1-1.

All these approaches use the relationships between traits to estimate the missing values but only Phylopars also uses phylogenetic trees directly for imputation. We thus evaluated how the addition of phylogenetic information influenced performance of the other three approaches (i.e., kNN, mice and missForest). For this purpose, we included phylogenetic information in the form of phylogenetic eigenvectors (Diniz-Filho, Ramos de Sant'Ana & Bini 1998) as additional predictor variables in the imputation process. We therefore tested kNN, mice and missForest in two ways: (i) with traits only and (ii) with both traits and phylogenetic information. Phylogenetic information for kNN, mice, and missForest was summarized by eigenvectors extracted from a principal coordinate analysis (PCoA), representing the variation in the phylogenetic distances among species [following Diniz-Filho *et al*. (2012a,b), PVR package].

To choose the number of eigenvectors to include in the analysis we ran preliminary tests where we introduced an increasing number of eigenvectors into the imputations and calculated the associated error (see 'imputation error calculation' section). The details of this test are given in supporting information Data S1-2. Error was minimized when including the first 10 eigenvectors as variables in the imputations (they represented 65% of the variation in the phylogenetic distances among species). This was consistent with the recommendations of the authors of mice who suggest including fewer than 15–25 variables in the imputation (van Buuren & Groothuis-Oudshoorn 2011). Note however that these eigenvectors are more representative of divergences closer to the root of the phylogeny so they do not include fine-scale differences among species (Diniz-Filho *et al*. 2012a). It is also possible to account for phylogenetic information after imputation through computation of phylogenetic independent contrasts as done by Fisher, Blomberg & Owens (2003). However, in the present approach phylogeny is directly used in the imputation process on the form of additional traits or variance-covariance matrix (Phylopars) potentially improving missing value estimation.

We ran all imputations (except Phylopars) and analyses in R 2.15.3 (R Core Team 2013). As we had to compute many imputations automatically, we ran Phylopars using Python (Enthought Canopy, Austin, TX, USA) with code provided by J. Bruggeman (Bruggeman, Heringa & Brandt 2009). Full details on the choice of tuning parameters used in kNN and missForest are given in supporting information Data S1-3. We ran all the analyses on log-transformed traits to reduce potential collinearity among predictors (van Buuren & Groothuis-Oudshoorn 2011) and because some of the approaches (e.g., kNN) could be sensitive to data with varying scales in the variables (Stekhoven & Bühlmann 2012). We back-transformed the data after imputations to calculate the error for each analysis, based on the original values (see below).

TEST DATASET

We used a newly compiled trait dataset on all mammals, based on Davidson *et al*. (2009) and Jones *et al*. (2009). We focused our analyses on mammals in the order Carnivora which has 273 species and has the most complete trait data. We chose the eight traits that were most complete in our dataset: body mass, litter size, maximum longevity, habitat breadth (these four traits had 6% of missing values) and adult body length, diet breadth, gestation length and weaning age (these four traits had 14% of missing values). From this initial dataset (hereafter, 'original dataset') we produced missing and imputed datasets (Fig. 1). All eight traits were imputed but in order to compare imputed data with original data (as complete as possible), all subsequent analyses were run on the four most complete traits (body mass, litter size, maximum longevity, and habitat breadth). Correlations between traits are given in supporting information Data S1-4. For the analyses including phylogeny, we used an interpolated smoothed tree of Mammals (S.B. Hedges, J. Marin, M. Suleski, M. Paymer, & S. Kumar, *submitted*) containing all mammal species. The eigenvectors were extracted from a pruned tree containing only the order Carnivora.

MISSING DATASETS

From the original dataset we derived eight incomplete datasets (Fig. 1), removing 10%, 20%, 30%,.. to 80% of values in the four most complete traits. In order to simulate real situations where data are not missing completely at random, values were removed in three ways (hereafter, 'missing dataset type'): (i) missing completely at random (MCAR); (ii) missing at random with respect to body mass (MAR.BM); (iii) missing at random with respect to phylogeny (MAR.PH). In MAR.BM datasets we introduced 60% of the missing values in small carnivores (below the median weight of 3 kg) and 40% in large ones (above the median weight). In MAR.PH we introduced 60% of the missing values in closely related species (i.e. sharing the same node) and 40% in the other ones. The proportions of 40–60% were comparable to biases existing in real datasets [e.g., PanTHERIA (Jones *et al*. 2009), see supporting information Data S1-5 for details]. Missing values were introduced in order to have at least one trait value per species. To ensure representativeness of each dataset, we generated 10 different missing datasets per missing dataset type (MCAR, MAR.BM, MAR.PH) and missing values percentage (10–80%) (Fig. 1).

IMPUTED DATASETS

We imputed the missing datasets using the approaches described above. In missForest and mice, two imputations in the same dataset give slightly different results. Indeed, missForest averages the results from multiple randomly generated trees and mice repeats the imputation multiple times. Therefore, for these approaches, we repeated the imputations 10 times per missing dataset and missing values percentage

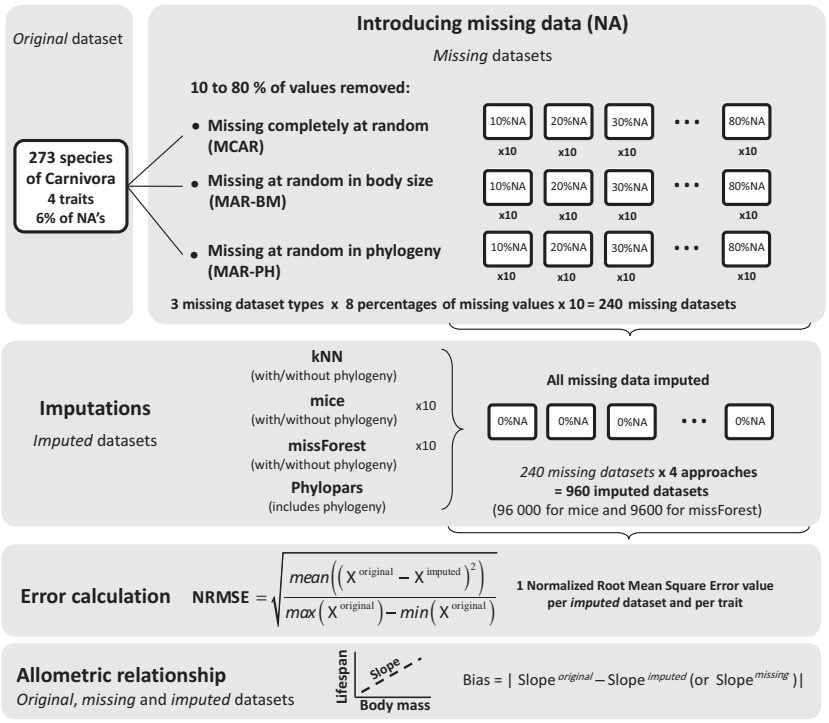

Fig. 1. Main steps used in our methods. From an initial *original* dataset we generated 3 types of *missing* datasets with 8 increasing percentages of missing values (NAs). We repeated this step 10 times. We imputed the previous *missing* datasets using the four different approaches (with and without phylogenetic information). We calculated an error by comparing the *original* dataset with the *imputed* datasets. We compared the slopes of the allometric relationship (between body mass and longevity) among *original*, *missing* and *imputed* datasets.

(Fig. 1). For mice, we extracted 10 imputed datasets per imputation (i.e., a total of 100 imputed datasets per missing dataset and per missing values percentage).

IMPUTATION ERROR CALCULATION

To assess the performance of each imputation approach, we calculated a Normalized Root Mean Squared Error [NRMSE (Oba *et al*. 2003)], which is the mean of squared imputation errors divided by the total range for each trait (see equation in Fig. 1). Lower values of NRMSE indicate better estimates of the variables. We calculated an error for each trait variable separately and a mean error across all traits for each imputed dataset. Because we made repeated imputations for missForest and mice, we calculated the error based on the mean trait value across all imputed datasets for each imputation approach (we also verified that using the median rather than the mean value would give very similar results). Note that the error was calculated only on the missing values that were artificially introduced in the datasets.

STATISTICAL ANALYSES

Error analyses

First, we used linear mixed-effects models (LME, nlme package, Pinheiro *et al*. 2014) to test the effects of the following predictor variables on the error: missing dataset types (MCAR, MAR.BM, MAR.PH), percentage of missing values, and imputation approach (fixed effects). We ran the analyses separately for imputations approaches with and without phylogeny. Second, to determine whether adding phylogenetic information into the imputations reduced the error, we ran the following model: error ~ missing dataset type + percentage of missing values + presence/absence of the phylogeny. We ran separate models using the error associated with each trait and the mean error as response variables. We examined the significance of each fixed effect with ANOVA. Because we generated 10 different datasets per missing type and per missing values percentage, we included the identity of the missing dataset as a random effect. The pairwise differences among approaches and missing dataset types were tested using Tukey tests (multcomp package, Hothorn, Bretz & Westfall 2008). Finally, we calculated the phylogenetic signal of traits using the K-statistic (Blomberg, Garland & Ives 2003) on the original dataset (phytools package, Revell 2012).

Allometric analyses

The value of the error itself does not indicate whether or not the imputed datasets reflect true biological relationships. We thus looked at the allometric relationships between two important mammalian life history traits, body mass and longevity, in our original dataset and compared it to the relationship in our imputed datasets. The comparisons were conducted separately for each imputed dataset, i.e. 10 comparisons per missing values percentage and per approach. We then counted the number of times we encountered significant differences between the original and the imputed slopes. We used this number as a response variable to test for differences between approaches, using ANOVA and glm with a Poisson distribution and percentage of missing values as a covariate.

Bias analyses

We compared imputation and data deletion using the same allometric relationship as above. In this case, we compared the bias in trait relationships when using datasets that include imputed data versus datasets that include only available raw data (i.e. removing those species that have missing data). We calculated the slopes of the relationship between body mass and longevity for the original dataset ($slope_{original}$), the missing datasets ($slope_{missing}$) and the imputed datasets ($slope_{imputed}$). We measured bias as the absolute value of the dif-

ference between $slope_{original}$ and $slope_{missing}$ or $slope_{imputed}$. We then used a GLM with normal distribution followed by Tukey tests to examine differences in bias between the missing datasets and mean bias of imputed datasets (because there were 10 imputed datasets per missing dataset). Percentage of missing values and missing dataset type were included as covariates. The two last analyses were done separately on approaches with and without phylogenetic information. To determine whether adding phylogenetic information into the imputations reduced the bias, we ran the following model: bias ~ missing dataset type + percentage of missing values + presence/absence of the phylogeny.

Results

ERROR ANALYSES

The approaches that performed better without including phylogeny were mice and missForest (Fig. 2 and Table S1-6-t1); no significant differences were found between these two approaches. When phylogenetic information was added into the imputation process, mice, missForest and Phylopars performed equally well (Tukey pairwise comparisons were not significant) and kNN gave the highest errors (Table 1). However, there were some differences in the imputation error for different traits (see Table 1 and Tables S1-6-t1 and S1-6-t2): mice estimated the values of body mass better than Phylopars and missForest (either with or without phylogenetic information). The latter two approaches estimated the values of longevity and habitat breadth better than mice (with phylogenetic information). There was no difference in how the four approaches performed in their estimate of litter size. The estimated values of body mass and litter size were more precise than for longevity and habitat breadth by all approaches (Table S1-6-t2).

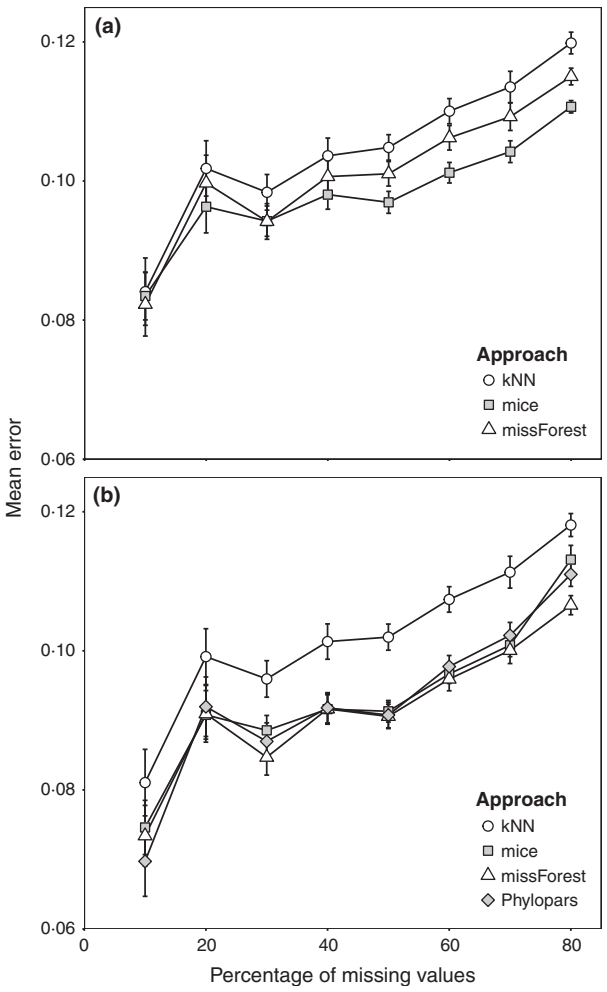

Fig. 2. Mean error (NRMSE) for kNN, missForest, mice and Phylopars on three types of missing datasets and eight levels of data removal (from 10% to 80%); (a) with and (b) without phylogenetic eigenvectors. The Phylopars approach can only be performed using phylogenetic information.

Overall, error differences among the three missing dataset types (MCAR, MAR.BM, MAR.PH) were marginally or not significant (Table 1 and S1-6-t1). However, there was a trend for poorer estimation of trait values when more data were missing in closely related species (MAR.PH) for body mass, longevity, and litter size, but not habitat breadth.

Adding phylogenetic information reduced the error for all the approaches except for kNN (Table 2). This result was stronger for missForest than for mice. Nevertheless, it did not improve the estimation of litter size in any approaches. Longevity was the trait whose estimation improved the most with the inclusion of phylogenetic information (even for kNN). Note that even if not significant, the phylogenetic imputation always improved the estimation of values for all traits, across all approaches (negative values in Table 2). The phylogenetic signal of traits was stronger for body mass (Blomberg's $K = 0{\cdot}84$, $P = 0{\cdot}001$) than for longevity ($K = 0{\cdot}43$, $P = 0{\cdot}001$) and habitat breadth ($K = 0{\cdot}25$, $P = 0{\cdot}001$). It was not significant for litter size ($K = 0{\cdot}05$, $P = 0{\cdot}95$).

ALLOMETRIC ANALYSES

For imputations without phylogeny, our analyses evaluating the allometric relationship between body mass and longevity showed that the relationships were preserved when up to 60% of the values were missing for mice, and up to 40% were missing for kNN and missForest (Figs 3, 4, S1-6-f1 and Table S1-6-t3); only differences between mice and missForest were significant ($\chi_{2,66} = 39$, $P = 0{\cdot}03$). In contrast, for imputations with phylogenetic information, analyses showed that the relationships were preserved when up to 60% of the values were missing for Phylopars and missForest, and up to 40% were missing for mice, and the relationships were not preserved at all for kNN (Figs 3, 4, S1-6-f1 and Table S1-6-t3). However, here we did not find any significant differences among approaches ($\chi_{3,88} = 59$, $P = 0{\cdot}6$).

BIAS ANALYSES

Overall, our analyses showed that bias was lower when missing data were imputed rather than deleted. For instance, when phylogenetic information was not included, bias was lower in datasets imputed with mice and kNN, compared to datasets

Table 1. Effects of missing dataset types, percentage of missing values and approach on the error for datasets imputed with phylogenetic information. Results of the ANOVA are given in italics, the intercepts + standard errors of Tukey tests are in non-italics. For results without phylogeny, see Table S1-6-t1

	Analyses with phylogeny - Error				
	Mean	Body mass	Litter size	Longevity	Habitat Breadth
Percent of missing values	*$F_{1,944}$ = 369****	*$F_{1,944}$ = 125****	*$F_{1,944}$ = 74****	*$F_{1,944}$ = 206****	*$F_{1,944}$ = 76****
Missing dataset type	*$F_{2,944}$ = 5**	*$F_{2,944}$ = 8**	*$F_{2,944}$ = 4**	*$F_{2,944}$ = 1*	*$F_{2,944}$ = 13****
MCAR-MAR.BM	−0·001 ± 0·001	0·002 ± 0·002	0·004 ± 0·003	0·002 ± 0·001	−0·011 ± 0·002***
MAR.PH-MAR.BM	0·002 ± 0·001	0·009 ± 0·002***	0·009 ± 0·003*	0·001 ± 0·001	−0·009 ± 0·002***
MAR.PH-MCAR	0·003 ± 0·001*	0·007 ± 0·002**	0·005 ± 0·003	−0·0001 ± 0·001	0·002 ± 0·002
Method	*$F_{3,944}$ = 23****	*$F_{3,944}$ = 77****	*$F_{3,944}$ = 0·4*	*$F_{3,944}$ = 41****	*$F_{3,944}$ = 6****
mice-kNN	−0·008 ± 0·001***	−0·039 ± 0·002***	0·003 ± 0·003	−0·001 ± 0·001	0·002 ± 0·002
missForest-kNN	−0·010 ± 0·001***	−0·011 ± 0·002***	−0·0002 ± 0·004	−0·012 ± 0·001***	−0·007 ± 0·002*
Phylopars-kNN	−0·009 ± 0·001***	−0·022 ± 0·002***	0·0003 ± 0·004	−0·009 ± 0·001***	−0·005 ± 0·002
missForest-mice	−0·001 ± 0·001	0·018 ± 0·002***	−0·004 ± 0·004	−0·012 ± 0·001***	−0·009 ± 0·002**
Phylopars-mice	−0·001 ± 0·001	0·018 ± 0·002***	−0·003 ± 0·004	−0·009 ± 0·001***	−0·007 ± 0·002*
Phylopars-missForest	0·001 ± 0·001	−0·001 ± 0·002	0·001 ± 0·004	0·003 ± 0·001	0·001 ± 0·002

Significance codes: ***: $P < < 0{\cdot}001$, **: $0{\cdot}001 < P < 0{\cdot}01$,*: $0{\cdot}01 < P < 0{\cdot}05$.
MCAR, missing completely at random; MAR.BM, more missing values in small species; MAR.PH, more missing data in closely related species.

Table 2. Comparison between imputations with and without phylogeny. Results of the ANOVA testing the effect of presence/absence of phylogenetic eigenvectors on the error are given in italics, the intercepts + standard errors of Tukey tests are in regular type

Error					
	Mean	Body mass	Litter size	Longevity	Habitat Breadth
All methods	*$F_{1,1424}$ = 50****	*$F_{1,1424}$ = 10***	*$F_{1,1424}$ = 0·5*	*$F_{1,1424}$ = 86****	*$F_{1,1424}$ = 39****
mice	−0·005 ± 0·001***	−0·005 ± 0·002**	−0·0009 ± 0·003	−0·004 ± 0·001**	−0·008 ± 0·002***
missForest	−0·009 ± 0·001***	−0·006 ± 0·002*	−0·0003 ± 0·003	−0·015 ± 0·001***	−0·014 ± 0·003***
kNN	−0·002 ± 0·002	−0·002 ± 0·003	−0·002 ± 0·001	−0·003 ± 0·002*	−0·003 ± 0·002

Significance codes: ***: $P < < 0{\cdot}001$, **: $0{\cdot}001 < P < 0{\cdot}01$,*: $0{\cdot}01 < P < 0{\cdot}05$.

with missing data or datasets imputed with missForest (Fig. 5 and Table 3). When phylogenetic information was included, all approaches had a comparable bias that was significantly lower than bias in datasets with missing data (Fig. 5 and Table 3). These differences seemed to increase when more than 30% of values were missing. Figure 5 presents the results from the allometric analyses (Fig. 3, 4, S1-6-f1 and Table S1-6-t3) and shows that for all approaches, bias increased when up to 60% of the values were missing. Phylogenetic imputation reduced the bias for missForest ($F_{1,43} = 71$, $P < 0{\cdot}001$) but not for mice ($F_{1,43} = 0{\cdot}4$, $P = 0{\cdot}5$) and kNN ($F_{1,43} = 0{\cdot}01$, $P = 0{\cdot}93$). Note that for high percentages of missing values, bias seemed to be higher for mice when phylogenetic information was added (Fig. 5), but this was not significant. Finally, we did not find any significant differences in bias among missing dataset types (MCAR, MAR.BM and MAR.PH) in this analysis ($F_{2,110} = 1$, $P = 0{\cdot}3$).

Discussion

All imputation approaches we tested provided valuable alternatives to data deletion in our trait database, and were reliable with up to 60% of missing values (Table S1-6-t3). We found that allometric relationships, particularly between body mass and longevity, were preserved in the imputed datasets (Figs 3, 4 and 5). Additionally, removing species with missing values from the analysis created more bias than imputing data, especially when more than 30% of the data were missing. However, as expected, the accuracy of estimated trait values decreased with increasing percentage of missing values. This has to be taken into account as datasets on poorly studied species are likely characterised by higher proportions of missing data.

Adding phylogenetic information to the imputation process improved the estimation of the traits evaluated in our analysis for all the approaches, and decreased the bias in trait relationships for missForest. Imputation approaches with phylogenetic variance-covariance matrix (Phylopars) and phylogenetic eigenvectors (missForest and mice) gave similar results. However, phylogenetic information did not improve estimation equally among traits. Estimation of litter size, for example, did not improve (or worsen) with the inclusion of phylogenetic information, but this trait did not exhibit phylogenetic signal in our dataset. Phylogenetic imputation, however, did improve estimation of body mass, but only weakly. While this trait exhibited a relatively strong phylogenetic signal compared to the other traits evaluated here, signal was still lower than expected under Brownian motion (under Brownian motion Blomberg's K $\approx$ 1; in our dataset body mass $K = 0{\cdot}84$). Addi-

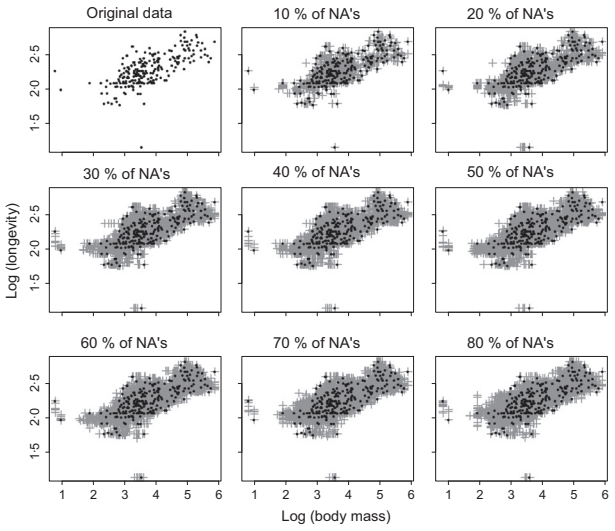

Fig. 3. Plots of the relationship between logged values of body mass and longevity for the original data (black points) and all the imputed datasets (grey crosses) for missForest (phylogenetic imputations). mice and Phylopars outputs were similar to missForest (see Fig. S1-6-f1).

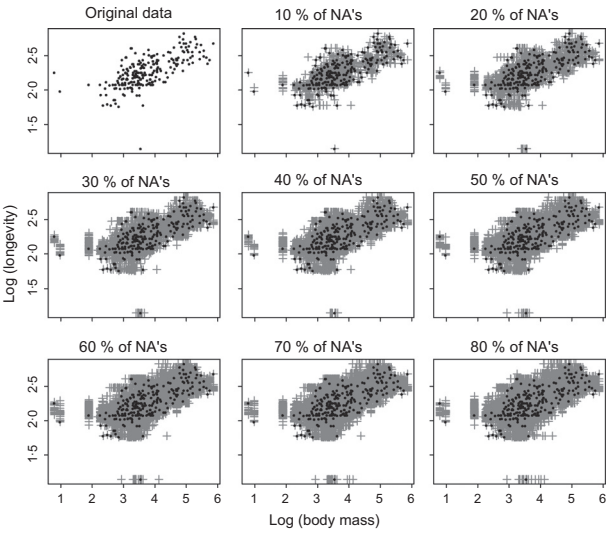

Fig. 4. Plots of the relationship between logged values of body mass and longevity for the original data (black points) and all the imputed datasets (grey crosses) for kNN (phylogenetic imputations).

tionally, both body mass and litter size were well estimated even without the inclusion of phylogenetic information. Estimates of longevity and habitat breadth substantially improved with phylogenetic imputation. The imputation approaches that we tested use both allometric relationships and phylogenetic relatedness as predictors. Including phylogeny can account for similarities among taxa that cannot be explained by simple allometries. The importance of adding phylogeny will likely vary in different traits being more important where phylogenetic signal is stronger and when there are no other traits with strong signal. Our results suggest that although trait databases often contain traits with different levels of phylogenetic signal, phylogenetic imputations should be used when possible because in no cases did they decrease the quality of trait imputation. In our study we used 10 eigenvectors. As discussed in Diniz-Filho *et al.* (2012a,b), the first eigenvectors contain mainly information on basal divergences, so by using only 10 eigenvectors we miss the most recent splits in the phylogeny. However we found that using 10 eigenvectors optimized the imputation error for missForest and mice (see supporting information Data S1-2). This result was surprising and may be related to the fact that adding many eigenvectors might dilute the information contained in the other traits. In order to avoid potential circularity in the imputation, the phylogeny should not be built using shared traits but based primarily on molecular data, as was the case in this study. Finally, as highlighted by Swenson (2014), the quality of the phylogenetic tree may affect the quality of the imputation, so analyses using imputed datasets should consider the resolution of the trees used for imputation.

We did not find large differences in the estimation of traits when data were missing completely at random (MCAR) or missing at random (MAR) with respect to body mass (MAR.BM) or phylogeny (MAR.PH). This was found both in analyses considering the error and the bias of imputations. Our result was different from a previous study based on all mammalian orders which found that datasets biased in body mass led to poor estimates of other traits (González-Suárez, Lucas & Revilla 2012). The difference in these results could be a consequence of the differences in values of body mass between the two datasets; body mass ranges from 6 g to 743 kg in Carnivores and 2 g to 3824 kg in all mammals. Evaluations of imputation approaches on other datasets with additional traits and taxonomic groups or on simulated data would be informative to confirm the robustness of these approaches.

Our study suggested that mice, missForest, and Phylopars performed similarly, while kNN performed less well. This is likely because kNN implements a single evaluation of values, thus ignoring the variation in estimation due to imputation and is hence less precise than methods that account for uncertainty caused by estimating missing data. However, when no phylogenetic information was added, mice gave better results than missForest in the allometric analyses. This difference was not significant when the phylogeny was used in the imputations.

Given that the three best approaches performed similarly, other insights might help the user to choose one approach over another (Table 4). Phylopars and missForest do not require specific user skills or knowledge about the relationship among traits. Indeed, they make very few assumptions about structural aspects of the data. Nevertheless, a minimal knowledge of datasets is always recommendable in order to detect potential anomalous estimates of trait values. Phylopars is a web application, which is very easy to use, but does not permit batch analyses. Moreover, due to limited computational resources, it does not handle datasets with >20 000 values, and even imputations with smaller datasets can take a long time. Both issues can be resolved by obtaining (and modifying) the python source code from the author of Phylopars (Bruggeman, Heringa & Brandt 2009). In mice, linear dependencies between

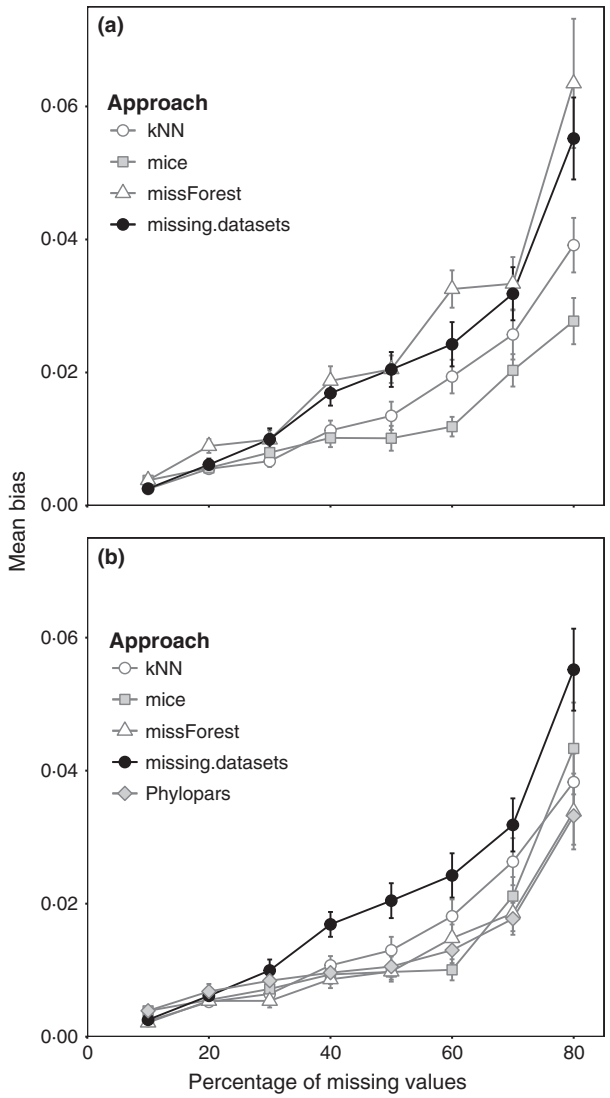

Fig. 5. Mean bias for four imputation approaches and the missing datasets (i.e. raw data, with missing data and no imputation) on eight levels of data removal (from 10% to 80%); (a) with and (b) without phylogenetic information. Bias = $|Slope_{original} - Slope_{imputed}$ (or $Slope_{missing})|$, where Slope = slope of the relationship between logged values of body mass and longevity.

Table 3. Results of Tukey tests for the bias. Significant relationships are in bold

Methods	Estimate	SE	*z* value	*P*
Without phylogeny				
mice-kNN	−0·05	0·04	−1·153	0·65
missForest-kNN	**0·19**	**0·04**	**4·783**	**< 0·001**
missing.datasets-kNN	**0·12**	**0·04**	**2·919**	**0·02**
missForest-mice	**0·24**	**0·04**	**5·936**	**< 0·001**
missing.datasets-mice	**0·16**	**0·04**	**4·072**	**< 0·001**
missing.datasets-missForest	−0·07	0·04	−1·864	0·24
With phylogeny				
mice-kNN	−0·02	0·04	−0·55	0·98
missForest-kNN	−0·08	0·04	−1·77	0·39
missing.datasets-kNN	**0·13**	**0·04**	**3·04**	**0·02**
Phylopars-kNN	−0·01	0·04	−0·18	0·99
missForest-mice	−0·05	0·04	−1·22	0·74
missing.datasets-mice	**0·16**	**0·04**	**3·59**	**0·003**
Phylopars-mice	0·02	0·04	0·37	0·99
missing.datasets-missForest	**0·21**	**0·04**	**4·81**	**< 0·001**
Phylopars-missForest	0·07	0·04	1·59	0·5
Phylopars-missing.datasets	**−0·14**	**0·04**	**−3·22**	**0·01**

Bias = $|Slope_{original} - Slope_{imputed}$ (or $Slope_{missing})|$, where Slope = slope of the relationship between logged values of body mass and longevity.

variables cause fatal errors and should be eliminated before imputation. This is potentially a recurrent and considerable problem because trait data tend to be correlated with each other. Mice is a rich tool that includes many multiple imputation methods and options. However, this diversity of choices may also be bewildering for basic users and jeopardize repeatability. Predictive mean matching is considered an effective overall imputation method by its authors (van Buuren & Groothuis-Oudshoorn 2011). However, we recommend specifying function details when using mice for trait imputations. Finally, our study showed that kNN did not perform as well as the other approaches because it produced larger errors and induced more bias in the allometric relationship. In addition, for kNN the user must specify a value of the tuning parameter *k* (number of neighbours used in the analysis – see supporting information Data S1-3), which is difficult to determine *a priori* and can have a substantial impact on the performance of imputation.

In this study, we did not test datasets with categorical traits (all categorical or mixed continuous/categorical traits). MissForest, kNN, and mice can analyse categorical variables (either nominal or ordinal) as they are; whereas Phylopars requires the use of dummy variable coding (transforming trait categories into many dichotomous variables). A test of imputation performance with non-trait categorical data found that missForest performed better than mice and another function based on nearest-neighbour (using dummy coding) for both categorical and mixed datasets (Stekhoven & Bühlmann 2012). Because trait databases have both categorical and continuous variables, approaches that can easily evaluate both are of greater interest. A study similar to ours, evaluating a combination of continuous and categorical variables (both nominal and ordinal) would be valuable for further guiding the future use of imputation approaches on traits datasets.

Data imputation is dependent on data quality and quantity. Some traits are more likely to be accurate than others. For instance, it is easier to measure body length than longevity because the former can be obtained quickly from multiple sources (e.g., field, museum and captivity data). Conversely, longevity may be difficult to measure, especially for long-lived species, and may vary in the wild and in captivity. Trait values can vary across a species range, therefore multiple measurements are preferable. Although our results suggest that imputation can be helpful when dealing with missing data, the addition of new data from literature, museum specimens and fieldwork is important and remains greatly needed.

In summary, we show that imputation is a promising and viable solution to help fill gaps in large trait databases where missing data can cause statistical biases and mask biological

Table 4. Key information and conclusions about the tested approaches

	kNN	mice	missForest	Phylopars
Description	Imputation using k-nearest neighbours	Multivariate imputation by chained equations	Nonparametric missing value imputation using random forest	Estimation of missing parameter values using phylogeny
R package	VIM	mice	missForest	Online at http://www.ibi.vu.nl/programs/phylopars/
Author	Templ *et al.* (2013)	Van Buuren & Groothuis-Oudshoorn (2011)	Stekhoven & Bühlmann (2012)	Bruggeman, Heringa & Brandt (2009)
Detailed information	For each missing value, find its k nearest variables. Impute the missing value using the weighted mean of the k variables. Weights depend on distance to neighbour variables.	Use predictive mean matches (many other methods could be used) to perform multiple imputation.	Train a random forest on observed values, predict the missing values using other variables and trained random forest, then proceed iteratively.	Estimate phylogenetic covariances using maximum likelihood estimation. Missing values are estimated using observed values and associated weights based on covariances.
Type of variables	Continuous, categorical	Continuous, categorical	Continuous, categorical	Only continuous
Advantages	Very fast [0·1/0·2 s]*/Few assumptions about data	Fast [2·5/3·4 s]* Choice among many imputation tools and parameters	Fast [2·3/2·8 s]*/Few assumptions about data	Few assumptions about data/Accounts for intraspecific variation among traits
Disadvantages	Needs prior knowledge of tuning parameters	Fails when variables are too correlated/Difficult set of parameters/Less efficient with many variables		Very slow [/168·9 s]*
Conclusions of our study		Best method without phylogeny	Best method with phylogeny	Best method with phylogeny

*The computation time is a mean of 30 imputations of the same dataset with 10% of data missing completely at random. Values are given for imputations [without/with phylogeny].

patterns. We hope that our study stimulates additional research on imputation approaches, perhaps by considering trait characteristics or multiple taxonomic groups. As trait databases continue to be improved, both by imputation approaches and further data collection, we will increase our ability to advance macroecological theory and address global conservation issues.

Acknowledgements

We thank Jorn Bruggeman for his help with Phylopars, Patrick R Stephens for his advice on the use of phylogenetic eigenvectors and Stef van Buuren for answering our questions on mice. We are also grateful to J.A.F. Diniz-Filho and two anonymous reviewers for their helpful comments on the manuscript. CP was supported by Capes grant PVE 018/2012, GCC by CNPq Grant #302776/2012-5. KTS was supported by NSF grant DEB-1146198. CHG and ADD were supported by NSF Dimensions grants #DEB-1136586 and 1136588 and also thank CAPES/Science without Borders grant PVE 018/2012.

Data accessibility

Carnivora trait dataset: uploaded as online supporting information Data S2.

References

Ambler, G., Omar, R.Z. & Royston, P. (2007) A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Statistical Methods in Medical Research*, **16**, 277–298.

Baraloto, C., Paine, C.E.T., Poorter, L., Beauchene, J., Bonal, D., Domenach, A.-M. *et al.* (2010) Decoupled leaf and stem economics in rain forest trees. *Ecology Letters*, **13**, 1338–1347.

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

Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5–32.

Bruggeman, J., Heringa, J. & Brandt, B.W. (2009) PhyloPars: estimation of missing parameter values using phylogeny. *Nucleic Acids Research*, **37**, 179–184.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011) mice: Multivariate Imputation by Chained. *Journal of Statistical Software*, **45**, 1–67.

Cardillo, M., Purvis, A., Sechrest, W., Gittleman, J.L., Bielby, J. & Mace, G.M. (2004) Human population density and extinction risk in the world's carnivores. *PLoS Biology*, **2**, e197.

Cardillo, M., Mace, G.M., Gittleman, J.L., Jones, K.E., Bielby, J. & Purvis, A. (2008) The predictability of extinction: biological and external correlates of decline in mammals. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **275**, 1441–1448.

Davidson, A.D., Hamilton, M.J., Boyer, A.G., Brown, J.H. & Ceballos, G. (2009) Multiple ecological pathways to extinction in mammals. *Proceedings of the National Academy of Sciences*, **106**, 10702–10705.

Di Marco, M., Cardillo, M., Possingham, H.P., Wilson, K.A., Blomberg, S.P., Boitani, L. & Rondinini, C. (2012) A novel approach for global mammal extinction risk reduction. *Conservation Letters*, **5**, 134–141.

Diniz-Filho, J.A.F., Ramos de Sant'Ana, C.E. & Bini, L.M. (1998) An eigenvector method for estimating phylogenetic inertia. *Evolution*, **52**, 1247–1262.

Diniz-Filho, J.A.F., Bini, L.M., Rangel, T.F., Morales-Castilla, I., Olalla-Tárraga, M.Á., Rodríguez, M.Á. & Hawkins, B.A. (2012a) On the selection of phylogenetic eigenvectors for ecological analyses. *Ecography*, **35**, 239–249.

Diniz-Filho, J.A.F., Rangel, T.F., Santos, T. & Bini, L.M. (2012b) Exploring patterns of interspecific variation in quantitative traits using sequential phylogenetic eigenvector regressions. *Evolution*, **66**, 1079–1090.

Fisher, D.O., Blomberg, S.P. & Owens, I.P.F. (2003) Extrinsic versus intrinsic factors in the decline and extinction of Australian marsupials. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **270**, 1801–1808.

González-Suárez, M., Lucas, P.M. & Revilla, E. (2012) Biases in comparative analyses of extinction risk: mind the gap. *Journal of Animal Ecology*, **81**, 1211–1222.

Graham, C.H., Parra, J.L., Tinoco, B.A., Stiles, F.G. & McGuire, J.A. (2012) Untangling the influence of ecological and evolutionary factors on trait variation across hummingbird assemblages. *Ecology*, **93**, S99–S111.

Guénard, G., Legendre, P. & Peres-Neto, P. (2013) Phylogenetic eigenvector maps: a framework to model and predict species traits. *Methods in Ecology and Evolution*, **4**, 1120–1131.
Hadfield, J.D. (2008) Estimating evolutionary parameters when viability selection is operating. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **275**, 723–734.
Hothorn, T., Bretz, F. & Westfall, P. (2008) Simultaneous inference in general parametric models. *Biometrical Journal*, **50**, 346–363.
Huang, S., Stephens, P.R. & Gittleman, J.L. (2012) Traits, trees and taxa: global dimensions of biodiversity in mammals. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **279**, 4997–5003.
Jones, K.E., Bielby, J., Cardillo, M., Fritz, S.A., O'Dell, J., Orme, C.D.L. *et al.* (2009) PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, **90**, 2648.
Joppa, L.N., McInerny, G., Harper, R., Salido, L., Takeda, K., O'Hara, K., Gavaghan, D. & Emmott, S. (2013) Troubling trends in scientific software use. *Science*, **340**, 814–815.
Junker, R.R., Blüthgen, N., Brehm, T., Binkenstein, J., Paulus, J., Martin Schaefer, H. & Stang, M. (2013) Specialization on traits as basis for the niche-breadth of flower visitors and as structuring mechanism of ecological networks. *Functional Ecology*, **27**, 329–341.
Lavorel, S. & Garnier, E. (2002) Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the Holy Grail. *Functional Ecology*, **16**, 545–556.
Loreau, M. (2010) Linking biodiversity and ecosystems: towards a unifying ecological theory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**, 49–60.
McGill, B.J., Enquist, B.J., Weiher, E. & Westoby, M. (2006) Rebuilding community ecology from functional traits. *Trends in Ecology & Evolution*, **21**, 178–185.
Nakagawa, S. & Freckleton, R.P. (2008) Missing inaction: the dangers of ignoring missing data. *Trends in Ecology & Evolution*, **23**, 592–596.
Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K. & Ishii, S. (2003) A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, **19**, 2088–2096.
Pacifici, M., Santini, L., Di Marco, M., Baisero, D., Francucci, L., Grottolo Marasini, G., Visconti, P. & Rondinini, C. (2013) Generation length for mammals. *Nature Conservation*, **5**, 89–94.
Pagel, M. (1999) Inferring the historical patterns of biological evolution. *Nature*, **401**, 877–884.
Paine, C.E.T., Baraloto, C., Chave, J. & Hérault, B. (2011) Functional traits of individual trees reveal ecological constraints on community assembly in tropical rain forests. *Oikos*, **120**, 720–727.
Pakeman, R.J. (2014) Functional trait metrics are sensitive to the completeness of the species' trait data? *Methods in Ecology and Evolution*, **5**, 9–15.
Pantanowitz, A. & Marwala, T. (2009) Missing data imputation through the use of the Random Forest Algorithm. *Advances in Computational Intelligence*, **116**, 53–62.
Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Development Core Team (2014). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-108. http://CRAN.R–project.org/package = nlme.
R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Revell, L.J. (2012) phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, **3**, 217–223.
Riek, A. & Bruggeman, J. (2013) Estimating field metabolic rates for Australian marsupials using phylogeny. *Comparative Biochemistry and Physiology. Part A*, **164**, 598–604.
Rubin, D.B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.
Santini, L., Di Marco, M., Visconti, P., Baisero, D., Boitani, L. & Rondinini, C. (2013) Ecological correlates of dispersal distance in terrestrial mammals. *Hystrix, the Italian Journal of Mammalogy*, **24**, 181–186.
Schafer, J.L. (1999) Multiple imputation: a primer. *Statistical methods in medical research*, **8**, 3–15.
Shan, H., Kattge, J., Reich, P.B., Arindam, B., Schrodt, F. & Reichstein, M. (2012) Gap filling in the plant kingdom — trait prediction using hierarchical probabilistic matrix factorization. *Proceedings of the 29th International Conference on Machine Learning*.
Stekhoven, D.J. & Bühlmann, P. (2012) MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, **28**, 112–118.
Swenson, N.G. (2014) Phylogenetic imputation of plant functional trait databases. *Ecography*, **37**, 105–110.
Taugourdeau, S., Villerd, J., Plantureux, S., Huguenin-Elie, O. & Amiaud, B. (2014) Filling the gap in functional trait databases: use of ecological hypotheses to replace missing data. *Ecology and Evolution*, **4**, 944–958.
Templ, M., Alfons, A., Kowarik, A. & Prantner, B. (2013) VIM: Visualization and Imputation of Missing Values. R package version 4.0.0. http://CRAN.R–project.org/package = VIM.
Ter Steege, H., Pitman, N.C.A., Phillips, O.L., Chave, J., Sabatier, D., Duque, A. *et al.* (2006) Continental-scale patterns of canopy tree composition and function across Amazonia. *Nature*, **443**, 444–447.
Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T.J., Tibshirani, R., Botstein, D. & Altman, R. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.
Verde Arregoitia, L.D., Blomberg, S.P. & Fisher, D.O. (2013) Phylogenetic correlates of extinction risk in mammals: species in older lineages are not at greater risk. *Proceedings. Biological sciences/The Royal Society*, **280**, 20131092.

Received 20 February 2014; accepted 8 July 2014
Handling Editor: Robert Freckleton

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Data S1-1. Details on the tested methods.

Data S1-2. Number of eigenvectors used in phylogenetic imputations.

Data S1-3. Details on tuning parameters for kNN and missForest.

Data S1-4. Correlations between traits.

Data S1-5. Proportions of missing values in trait datasets.

Data S1-6. Additional tables and figures.

Data S2. Carnivora trait dataset.