**Research article**

A cautionary note of phylogenetic signal estimation on imputed databases

Lucas Jardim1\*, Luis Mauricio Bini1, José Alexandre Felizola Diniz-Filho1, Fabrício Villalobos1,2

1Departamento de Ecologia, Universidade Federal de Goiás, Goiânia, Goiás, Brasil;

2Red de Biología Evolutiva, Instituto de Ecología, A.C., Carretera antigua a Coatepec 351, El Haya, 91070 Xalapa, Veracruz, Mexico

\*Corresponding author: Lucas Jardim, Departamento de Ecologia, ICB, Universidade Federal de Goiás, 74690-900, Goiânia, Goiás, Brasil. E-mail: lucas.ljardim9@gmail.com.

Running title: *Phylogenetic signal and trait imputed databases*

**Acknowledgements**

We thank G. Guénard for advices on PEM and L. Revell for help on evolutionary simulations. We are also grateful to Tiago Quental, Matheus Ribeiro, Daniel Paiva, João Nabout, Cristian Dambros, Tiago Freitas, Bruno R. Ribeiro, Gracielle Higino, Pablo Silva, and Adriana Diaz for suggestion and commentaries that improved the study. L.J. was supported by a CAPES doctoral fellowship. FV was supported by a BJT “Science without borders” CNPq grant. L.M.B. and J.A.F.D.F. are continuously supported by CNPq productivity grants. This paper was developed in the context of the National Institutes for Science and Technology (INCT) project in “Ecology, Evolution and Biodiversity Conservation”, supported by MCTIC/CNPq (proc. 465610/2014-5) and FAPEG (proc. 201810267000023).

A**bstract**

Given the prevalence of missing data on species’ traits – Raunkiaeran shortfall-, several methods have been proposed to fill sparse databases. Analyses based on these imputed databases can introduce several biases. Here, we evaluated potential biases in descriptive statistics, regression parameters, and phylogenetic signal estimated from imputed databases under different missing and imputing scenarios. We found that percentage of missing data, missing mechanisms, and imputation methods were important in determining errors of estimates. We also found that imputation errors are not linearly related to estimate errors. Adding phylogenetic information provides better estimates of evaluated statistics, but must be combined with other variables. We advise researchers to share both their raw and imputed data and consider the pattern of missing data to evaluate methods that performs better for their goals. In addition, the perfomance of imputation methods must be based on statistical estimation instead of imputation error.

**Key-words**: bias, Multiple Imputation, trait databases, Phylogenetic Eigenvector Maps, phylogenetic signal, Phylogenetic Comparative Methods.

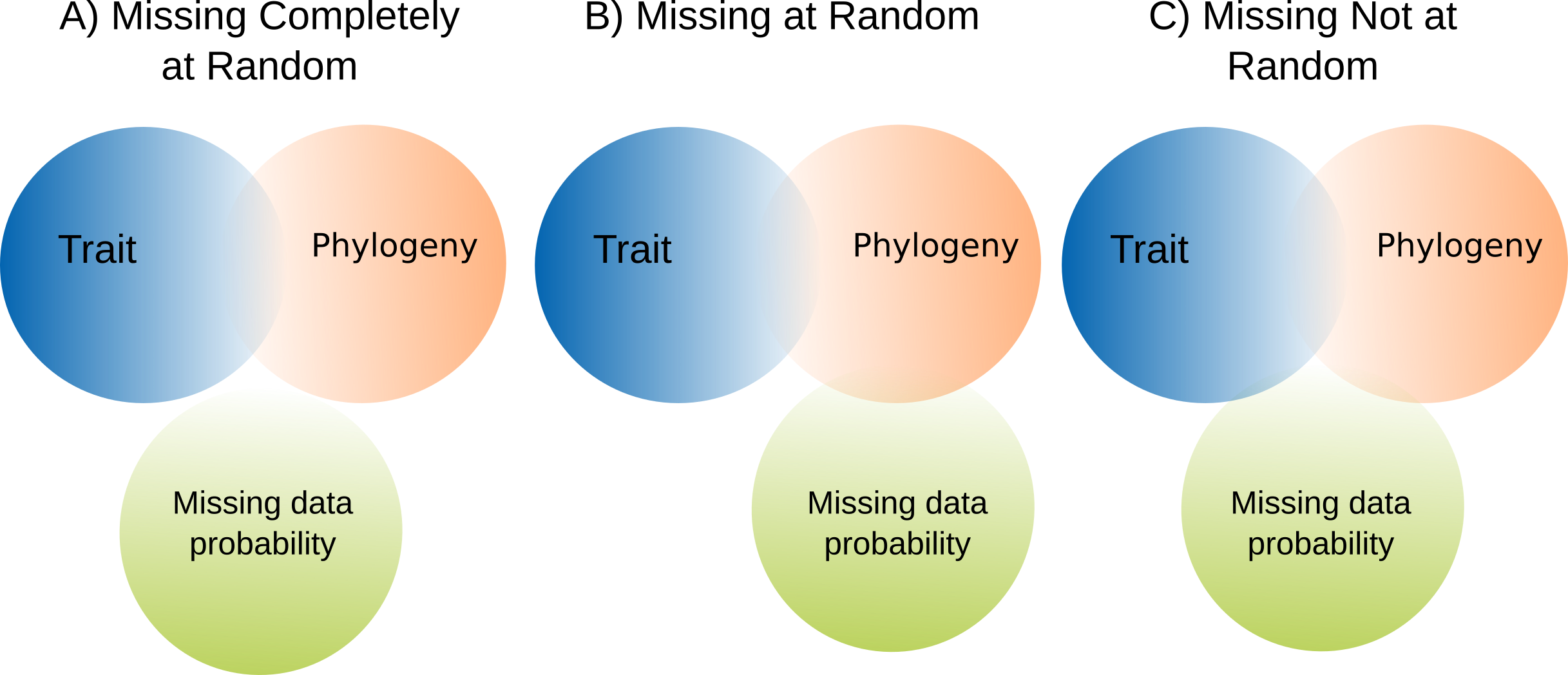
**Introduction**

Missing data are an ubiquitous feature of real-world datasets (Nakagawa & Freckleton 2008). Lack of information may limit the application of statistical analysis and can lead to biased estimates and conclusions on the phenomena of interest (Rubin 1976). Research in ecology and evolutionary biology usually requires data about species and their traits to answer different questions from community assembly and ecogeographical rules to correlated evolution, diversification rates and extinction probability, among others (Gaston et al. 2008; Goldberg et al. 2010; Jetz and Freckleton 2015; Lukas and Clutton-Brock 2013; Purvis et al. 2000; Webb et al. 2002). Thus, to facilitate research and make it reproducible and data more accessible (Reichman et al. 2011), ecologists and evolutionary biologists usually create databases that include information on huge amounts of species and their traits (e.g., Jones et al. 2009; Kattge et al. 2011; Wilman et al. 2014). However, as databases become larger, the probability of having all the necessary data for all species rapidly decreases. This lack of knowledge about species’ traits and their ecological functions was recently defined as the Raunkiaeran shortfall (Hortal et al. 2015) or Eltonian shortfall (Rosado et al. 2015).

Owing to the ubiquity of the Raunkiaeran shortfall, some researchers are interested in filling such gaps in their databases for their own analyses but also to make them available for other researchers (Schrodt et al. 2015; Swenson 2014). To do so, recent studies suggest the use of phylogenetic information in the imputation process (Guénard et al. 2013; Schrodt et al. 2015; Swenson 2014). Phylogenetic information is important in imputation because closely related species resemble, on average, each other more than distantly related species. This phenomenon is commonly known as phylogenetic signal (Simon P Blomberg et al. 2003). Consequently, knowing the phylogenetic position of species could, in principle, be used to perform a good estimation of missing trait values. However, the relationship between trait divergence and phylogenetic distance may be more complex (due to distinct evolutionary models) than usually assumed (Hansen and Martins 1996; Münkemüller et al. 2012). For instance, under an Ornstein-Uhlenbeck evolutionary model traits may evolve under selection constrains where species track a trait single or multiple optima, causing phenotypic resemblance even among phylogenetically distant species (Beaulieu et al. 2012a; Hansen and Martins 1996). Alternatively, traits may evolve accelerating trait’s evolutionary rates resulting in phylogenetically closely related species having different trait values (Simon P Blomberg et al. 2003; Harmon et al. 2010). Finally, but not a complete list of evolutionary model (Simone P. Blomberg et al. 2020; Landis and Schraiber 2017), trait evolution may behave like a drift process (e.g., Brownian motion) where species trait differences are directly correlated with time since divergence (Felsenstein 1985; R. P. Freckleton et al. 2002; Hansen and Martins 1996) Therefore, imputation methods should explicitly consider or assume a trait evolutionary model determining the relationship between species resemblance and phylogenetic proximity (Guénard et al. 2013).

In 1976, Donald B. Rubin proposed a missing-data theory to allow analysis of incomplete datasets (Rubin 1976), explaining how unbiased parameters could be estimated with missing data by considering the mechanisms causing missing data. These mechanisms were classified into three categories: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). They mean, respectively, that missing values are equally probable across a dataset, probability of missing data is correlated with other variables (eg. phylogeny) rather than to the variable with missing data (target variable), and probability of missing data is itself correlated to the target variable and dependent on the missing data (Enders 2010; Nakagawa and Freckleton 2008; Rubin 1976; van Buuren 2012) (Figure 1).

When dealing with missing data, the above mechanisms need to be taken into account before analysis (Rubin 1976). This is because different methods that handle missing data assume different mechanisms, so using them indiscriminately may bias parameter estimates (Enders 2010; Rubin 1976; van Buuren 2012).

**Figure 1**. Correlation structure among variables in each missing data mechanism. Circles represent model components and their intersection represents correlations among them. Blue circles are traits missing data, orange circles are phylogenetic information and green circles are missing data probability. Then, in **A**, traits have phylogenetic signal and their values are missing randomly. In **B**, phylogenetic information explains missing data probability of the traits. In **C**, phylogenetic information is not able to explain all missing data probability.

Multiple Imputation and Full Information Maximum Likelihood methods are currently regarded as the most appropriate methods to handle missing data, because they work under MAR and MCAR scenarios and provide unbiased estimates (Enders 2010). In contrast, it is very difficult to model missing data under a MNAR scenario (van Buuren 2012). This is so due to the need of considering a model that represents the probability of missing values to occur and because the shape of the probability density function is not known (Enders 2010; van Buuren 2012).

Nowadays, large imputed databases already exist using taxonomic, ecological or allometric relationships to fill in missing values (Jones et al. 2009; Wilman et al. 2014). This highlights the need to critically evaluate the use of imputed databases given that the reliability of statistical analysis under missing data is dependent on how many values were missing in the original data, what mechanism caused data to be missing and which methods were used in the imputation process (Enders 2010; Schafer and Graham 2002; van Buuren 2012). Moreover, other problems can also arise when testing for phylogenetic signal (Cavender-Bares et al. 2009; Münkemüller et al. 2012). In such cases, if analysis were to be conducted on phylogenetically imputed data, results could be misleading given that missing values would have been already filled based on their phylogenetic structure, thus potentially inflating the level of phylogenetic signal. This potential issue can have important consequences for studies evaluating, for example, niche conservatism, trait lability, community assembly and diversification (Simon P Blomberg et al. 2003; Cavender-Bares et al. 2009; Goldberg et al. 2010; Wiens and Graham 2005).

Considering the current need for complete databases and the use of imputation methods to accomplish this, we evaluate how the estimation of common descriptive statistics, regression coefficients and phylogenetic signal can be misled by the percentage of missing data, the mechanism of missing data, the model of trait evolution and the choice of methods used to handle missing values. To accommodate these scenarios, we use simulated phylogenies and traits under different combinations of such conditions. In addition, to address the impact of imputation error on phylogenetic signal estimates, we evaluated the relationship between their errors.

**Methods**

*Phylogeny simulation*

To evaluate the effect of imputing missing values into sparse databases (i.e. with missing data), we first simulated 100 yule phylogenies. Each phylogeny had 200 species and were simulated using the function *pbtree* from the R package *phytools* (Revell 2012).

*Trait simulation*

For each phylogeny, we simulated two traits: a target trait and an auxiliary trait. The first trait represented the one that would be imputed (i.e. missing-value trait), whereas the second trait represented an auxiliary trait that would be used to impute values for the target trait.

The target trait was simulated using the *rTraitCont* function from the *ape* package (Paradis et al. 2004). We modeled this trait under an Ornstein-Uhlenbeck evolutionary process (OU) (Gillespie 1996), because it allowed us to simulate trait evolution within a continuum from Brownian motion (BM) to weak and strong levels of selection strength (OU’s α) of Ornstein-Uhlenbeck model of evolution (Hansen 1997a; Hansen and Martins 1996). Thus, we could evaluate the performance of imputation methods under different levels of phylogenetic signal. We fixed the target trait’s optimum (ϴ) to zero and the trait interspecific variation (σ) equal to one. Also, we simulated different selective strengths by varying α (0.05, 0.1, 0.2, 0.5, 1, 2). Such values covered evolutionary scenarios from Brownian motion (OU α = 0.05) to strong selective strength (OU α = 2).

The auxiliary trait represented a variable used to impute values into the target trait. We simulated auxiliary traits in two ways: (i) correlated with the phylogeny and (ii) correlated with the target trait but uncorrelated with phylogeny. For (i), we simulated the trait as follow:

eqn 1

where *y* is the target trait, *x* the auxiliary trait, and *r* the correlation coefficient between both traits, which was set to *r* = 0.6. Species covariance matrix (**∑**) (Felsenstein 1985; Revell et al. 2008) and σ² of the target trait variation rate calculated as the mean of squared phylogenetic independent contrasts (Robert P Freckleton and Jetz 2009), which was estimated using the *pic* function from *ape* (Paradis et al. 2004). Multivariate Normal Distribution (MVN) was simulated using the *fastBM* function from the *phytools* R package (Revell 2012). This auxiliary trait was later used when simulating the MCAR (Missing Completely at Random) and MAR.PHYLO (Missing at Random correlated with phylogeny) (see below).

For the (ii) scenario, where the auxiliary trait is correlated with the target trait but uncorrelated with phylogeny, the auxiliary trait was simulated using equation 1 with **∑** having off-diagonal entries equal to zero (i.e. no covariance among species) and diagonal entries representing, for each species, the sum of all branch lengths from the root to the tip. We simulated MVN using the *mvrnorm* function in the R package MASS (Venables and Ripley 2002). When using this auxiliary trait to impute target trait values, we expected that using the phylogeny into the imputation methods would not improve our analysis (i.e. provide no information on missing data) since the probability of missing values would only be correlated with the auxiliary trait and not with the phylogeny.

*Missing data scenarios*

To create missing data, we used the target trait simulated above and deleted different percentages of its values following three scenarios of missing data: Missing Completely at Random (MCAR), Missing at Random but phylogenetically structured (MAR.PHYLO), and Missing at Random but correlated with the auxiliary trait not phylogenetically structured (MAR.TRAIT). We created the MCAR scenario by randomly sampling a percentage (see below) of species along each phylogeny and replacing their trait values with missing values. For the MAR.PHYLO scenario, we sampled a species in each phylogeny and selected a percentage of its closest species to replace their trait values with missing values, allowing a strong missing data pattern that was phylogenetically structured. For the last scenario, MAR.TRAIT, we used the auxiliary trait (see above) to replace values in the target trait. We ordered the values of the auxiliary trait in ascending order and replaced the first percentage of values of the target trait with missing values. This represented a missing data pattern correlated with another trait, different to the target one. For each scenario, we simulated different percentages of missing values in the target trait: 5, 10, 30, 50, 70 and 90% of missing data.

*Imputation methods*

We evaluated four methods often applied by researchers to handle missing data: choose the value of the phylogenetically closest species (hot-deck), no imputation but simply deleting missing values (Listwise), phylogenetic eigenvector maps (PEM), and multiple imputation by chained equations (MICE).

We used the hot-deck method by calculating the phylogenetic distance among species and choosing the value of the phylogenetically closest species to impute the missing values. Under the Listwise method, we did not impute values but simply pruned those species with missing values from the phylogenies before the analyses. The PEM method uses both phylogenetic eigenvectors (Diniz-Filho et al. 1998; Guénard et al. 2013) and traits to impute data considering different OU processes. We applied this method in two ways: first, using only the phylogenetic eigenvectors (PEM.notrait) and, second, using these eigenvectors and the auxiliary trait (PEM.trait). By applying the PEM method in these two ways allowed us to evaluate whether phylogenetic information alone could impute data well or auxiliary traits were necessary. Eigenvector selection and fitting of trait evolutionary models were performed using the *MPSEM* R package (Guénard et al. 2013) using forward selection based on the second-order Akaike Information Criterion (Burnham and Anderson 2002). The MICE method simulates several possible values for missing data from a posterior predictive distribution, then runs analysis and pools results over all simulated data (van Buuren et al. 2006). We chose this method because it is flexible and allows imputing categorical, continuous, and non-normally distributed data. We applied MICE by creating 10 datasets to run our analysis over them and pooled the results. The quantity of datasets created by MICE is dependent on the percentage of missing data and more datasets can provide higher accuracy and power in the analyses. However, because our objective was simply to estimate statistical bias instead of inference power (Graham et al. 2007; von Hippel 2018), 10 datasets can be considered appropriate (Graham et al. 2007). As with the PEM method, we applied MICE in two ways: only considering the auxiliary trait (MICE) and using this trait plus the phylogenetic eigenvectors selected as in PEM (MICE.phylo). We imputed data with MICE using the *mice* R package (Buuren and Groothuis-Oudshoorn 2011).

We simulated 540 scenarios representing each combination of missing data percentage, mechanism, OU’s α, and imputation methods. For each scenario, we simulated 100 replicates, thus producing 54000 independent results.

*Estimating phylogenetic signal*

We calculated the phylogenetic signal in our simulated phylogenies using two metrics: Blomberg’s K calculated with the *phylosig* function of *phytools* (Revell 2012) and Moran’s correlograms (Gittleman and Kot 1990). For calculating these correlograms, we created a phylogenetic distance matrix per phylogeny using the *cophenetic* function of *ape* *(Paradis et al. 2004)* and built the correlograms with the *lets.correl* function of the *letsR* R package (Vilela and Villalobos 2015). Then, we used the Moran’s I in the first distance class of the correlogram as indicative of phylogenetic signal, taking into account the non-linearity of correlograms generated under OU processes (Diniz-Filho 2001).

*Imputation effects on phylogenetic signal and common statistics*

Traditionally, performance evaluation of imputation methods have focused on common descriptive statistics such as mean, variance and regression coefficient, instead of phylogenetic patterns. Therefore, we also evaluated the effect of imputed data on the estimation of such descriptive statistics. We calculated the mean and variance of the target trait as well as the regression coefficient (Ordinary Least Square) between the target trait and the auxiliary trait, before producing missing data and after imputing such data. Next, we measured the percentage of estimation error (*p)* for these statistics and phylogenetic signal as below:

eqn2

where τ represents the statistics calculated over imputed traits, τ0 is the statistics calculated from original traits.

*Imputation error*

To measure the potential error introduced by imputation methods, that is the deviation between imputed and original data, we followed (Penone et al. 2014) and used the normalized root mean squared error (NRMSE):

eqn 3

where *y0* is the original trait value, *y*is the imputed value, max(*y0*) and min(*y0*) are the maximum and minimum values of the original trait, respectively. Lower NRMSE means high accurate imputation.

*Overall analyses*

We were also interested on evaluating the effects of percentage of data missing, missing data mechanism, OU’s α and imputation methods as factors influencing the *p* of abovementioned common statistics and phylogenetic signal. To do so, we grew regression trees (Hastie et al. 2009) with these factors as predictors and log(*p*), separately for each statistic, as response variables. We grew regression trees using the *rpart* R package (Therneau and Atkinson 2019) and variable importance was calculated as sum of improvements on the sum of squares in each node split by that variable. In addition, to evaluate the relationship between accuracy of imputation methods and *p* on phylogenetic signal, we also plotted the relationship between NRSME and *p* of Blomberg’s K and Moran’s Correlogram. All simulations and analysis were run in R 3.6.3 (R Core Team 2020).

**Results**

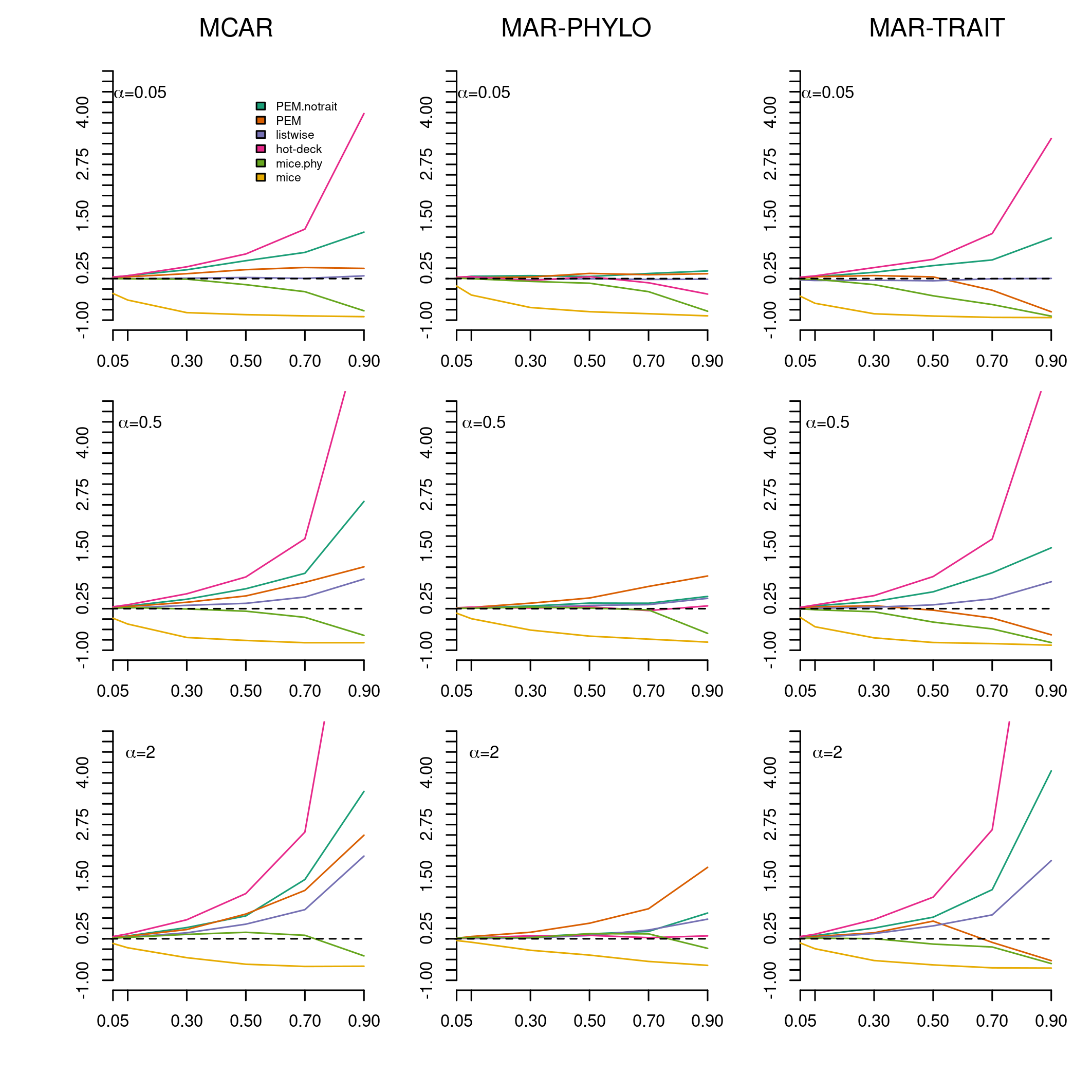
We found that differences in *p* for phylogenetic signal and common statistics, were mainly determined by the percentage of missing values, mechanism of missing data and imputation methods (Table 1). OU (α) and imputation error had little impact on estimation errors (Table 1), but they cannot be ignored as they interacted with the other factor, in different ways, according to each evaluated statistic (Online Resource Figure.S1-S5, Table 1). The first criterion to determine low estimation errors, for all statistics, was the presence of more than 90% of the data, or restricting the missing data to less than 50% of the data. Consecutively, missing data mechanisms and imputation methods were the factor that most contributed to *p*, but while mechanisms affected common statistic (mean, regression coefficient), imputation methods impacted phylogenetic signal metrics (Table 1). Variance was mostly impacted by missing data percentage.

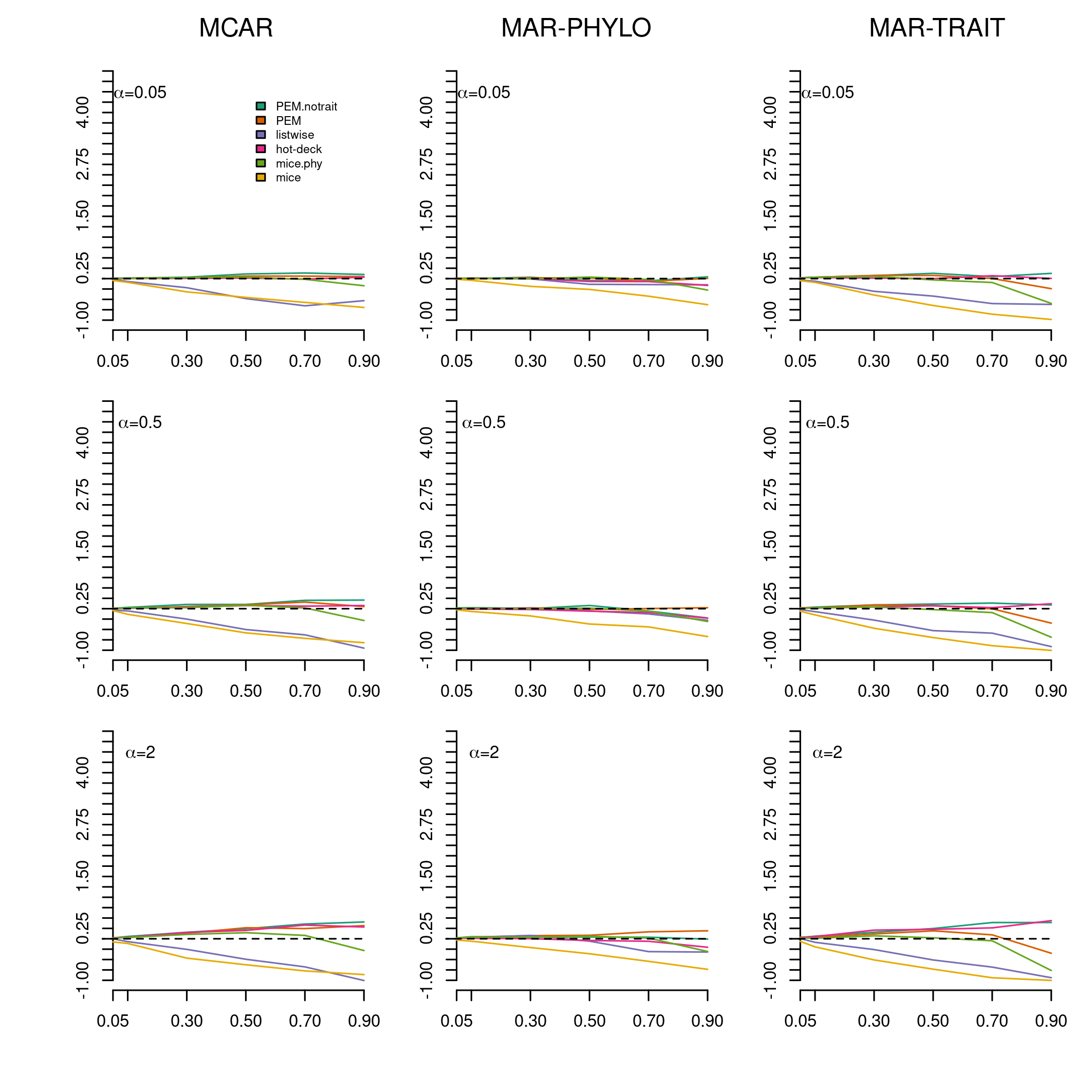
**Table 1.** The importance of each factor to explain percentage of error for common statistics and phylogenetic signal. Factor importance was calculated as model improvement in each node split related to the factor.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Imputation Error (NRMSE) | Methods | Mechanism | OU (α) | Percentage of missing values |
| Blomberg’s K | 0.005 | 0.241 | 0.000 | 0.000 | 0.754 |
| Moran’s Correlogram | 0.043 | 0.185 | 0.004 | 0.000 | 0.767 |
| Regression Coefficient | 0.040 | 0.009 | 0.181 | 0.002 | 0.769 |
| Mean | 0.005 | 0.000 | 0.163 | 0.177 | 0.654 |
| Variance | 0.039 | 0.030 | 0.034 | 0.000 | 0.898 |

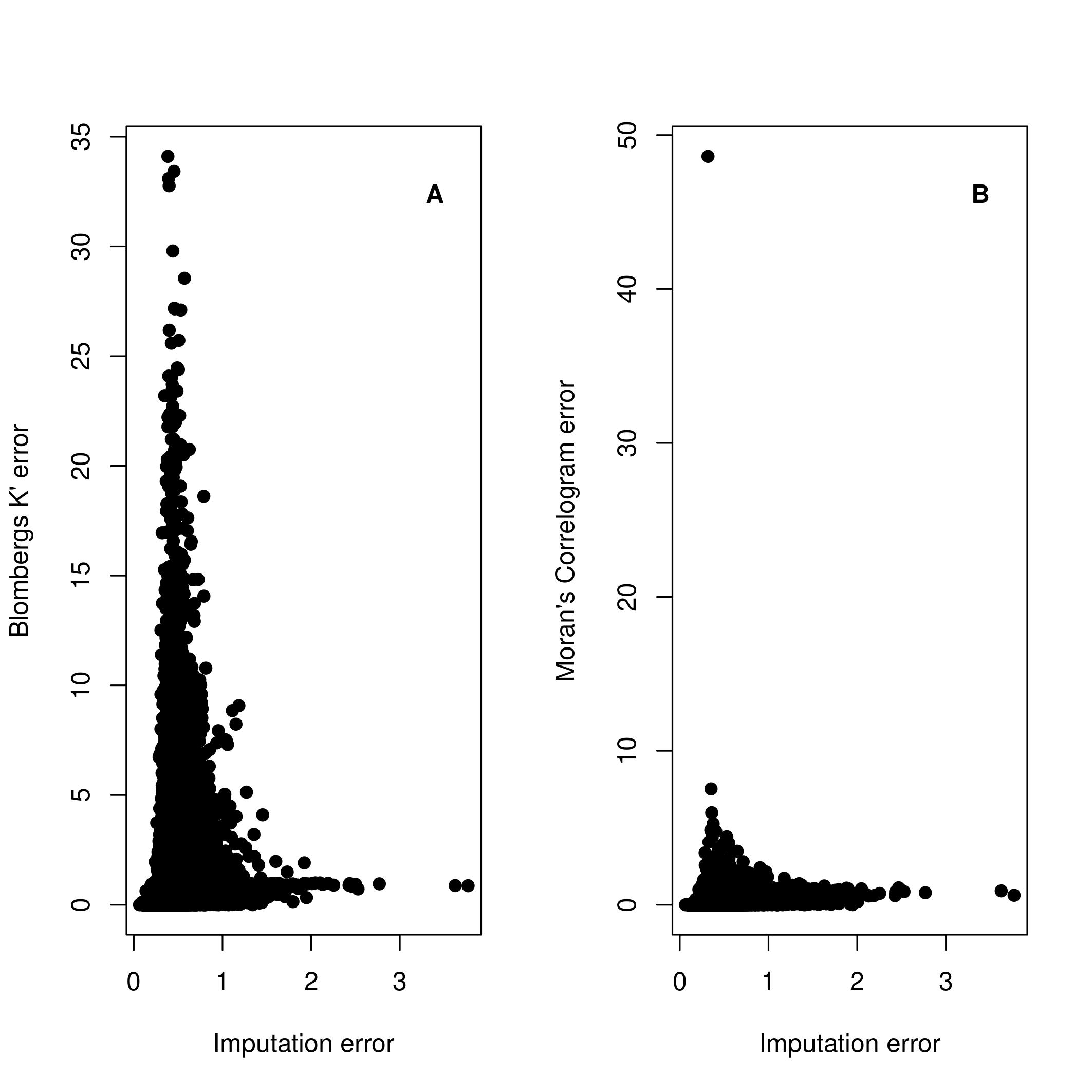
Looking specifically to phylogenetic signal metrics, Blomberg’s K was more sensitive to missing data percentage, mechanisms, imputation methods and phylogenetic signal than Moran’s Correlogram (Figure 2, 3). Blomberg’s K had better performance in MAR.PHYLO mechanism and under scenarios of trait evolution closer to Brownian motion (OU’s α = 0.05) (Figure 2). Imputation methods based only on phylogenetic information (PEM.notrait, hot-deck) caused overestimation of Blomberg’s K, specially hot-deck (Figure 2). Imputations combining phylogenetic information and auxiliary variable reduced phylogenetic signal overestimation, but it may also cause underestimation of phylogenetic signal when using MICE.phy method. Imputations without phylogenetic information (MICE) underestimated phylogenetic signal (Figure 2) and Listwise allowed accurate estimation of Blomberg’s K when the trait evolved with high phylogenetic signal, such as Brownian motion scenarios (OU’s α = 0.05) (Figure 2).

Moran’s Correlogram was more sensitive to missing data exclusion (Listwise method) than Blomberg’s K (Figure 2-3). Also, it is more sensitive to underestimation than overestimation of phylogenetic signal. As Blomberg’s K, MICE underestimated Moran’s Correlogram, but the other methods were more prone to cause errors in MAR-TRAIT scenarios with high percentage of missing data (Figure 3).

**Figure 2**. Median of the *p* for Blomberg’s K under different imputation methods, OU’s α, missing data percentage and mechanisms. Specie’s traits were deleted at random (MCAR), phylogenetically correlated (MAR-PHYLO) or correlated to a auxiliary trait without phylogenetic signal (MAR-TRAIT).

**Figure 3**. Median of the *p* for Moran’s Correlogram under different imputation methods, OU’s α, missing data percentage and mechanisms. Specie’s traits were deleted at random (MCAR), phylogenetically correlated (MAR-PHYLO) or correlated to a auxiliary trait without phylogenetic signal (MAR-TRAIT).

Finally, imputation error and phylogenetic signal errors (Fig. 4) had a non-linear relationship (Figure 4). It means that imputing trait values accurately can improve estimation of some statistics such as Moran’s correlogram, variance and regression coefficient, but at low percentage of missing values (Online Resource, Figure.S2,S4,S5). Nonetheless, small increases of imputation errors is prone to produce high error of phylogenetic signal estimation (Figure 4), but high imputation error may guarantee accurate phylogenetic signal, if other factor are ignored.

**Figure 4**. Scatterplot of imputation errors (NRMSE) and the absolute of the *p* for (A) Blomberg’s K and (B) Moran’s Correlogram.

**Discussion**

Ecologists and evolutionary biologists are increasingly creating, using, and sharing large trait databases that are inevitably sparse and often completed by imputing missing values (Guénard et al. 2013; Schrodt et al. 2015; Swenson 2014) . Here we re-force the statistical literature claims that we should be extremely careful when using imputed databases, even for the estimation of common statistics (i.e. means, variances and regression coefficients) (Enders 2010; Little and Rubin 2002; Rubin 1976). Our findings revealed that estimations based on imputed data depends on every aspect of data property and strategy of analysis, as percentage of missing data, mechanism causing data absence, trait evolution, methods for gap filling, and statistics or parameters to be estimated. This has commonly been acknowledged in statistical research and should begin to be so in the ecological and evolutionary research (Nakagawa 2015; Nakagawa and Freckleton 2008).

The most pervasive obstacle from large datasets is simply the proportion of species lacking data. Previous studies found that reliable estimations from imputed data can be made when 60% or 90% of the values were missing (Barzi 2004; Marcondes 2019; Penone et al. 2014). It is congruent with our results, as we found that less than 50% of missing data could result in accurate estimates, but it should be taken with caution as missing data percentage interacted with other factors to make reliable estimations. Also, it is dependent on the total amount of information available in the database, it means imputations with several correlated variables, with complementary information, may improve analyses, even in high percentage of missing data (Graham et al. 2007; van Buuren 2012; von Hippel 2018). Therefore, it is more fruitful consider imputations as context dependent and measure the sensitivity of the methods and approaches to each study.

Researchers should hypothesize, as a first step, the causes of data absence (van Buuren 2012). The most common assumption in ecological and evolutionary studies is that data is MCAR (Nakagawa and Freckleton 2008). This is evident in the wide variety of functions of the most commonly used software (the R programming language) allowing deleting missing values (Listwise) indiscriminately. Indeed, if data were under MCAR, common statistics should be well estimated (Enders 2010; Marcondes 2019; Nakagawa and Freckleton 2010; van Buuren 2012), and as we found, in low percentage of missing data and the traits evolving by Brownian motion, estimations of phylogenetic signal would be accurate. However, biological data are rarely missing completely at random (Enders 2010; Hadfield 2008; Marcondes 2019; Nakagawa and Freckleton 2008). For instance, bias in ecological lack of data can be related to the fact that some taxa are most studied than others (Gonzalez-Suarez et al. 2012). Such biases can stem from body mass differences among species, where large species have a higher probability of being described first (Gaston and Blackburn 1994; Vilela et al. 2014) and have their data collected (Gonzalez-Suarez et al. 2012) compared to small species. Also, species present in easily accessible regions are better studied than those occurring in regions that are hard to access (Reddy and Dávalos 2003). If biases of missing data have a phylogenetic structure, we found that phylogenetic imputation and Listwise can result in accurate phylogenetic signal estimates, what is probably found in most databases. Otherwise, if missing data are not structured on phylogeny but to other variable, phylogenetic imputations can generate biases on phylogenetic signal estimation.

Our simulations revealed that imputation methods can recover the original phylogenetic signal, but methods that combine phylogenetic information and other variables such as PEM and MICE.phy are better than methods not doing so (hot-deck, PEM.notrait, MICE). Such findings support previous claims favoring “phylogenetic imputation” as a powerful tool in predicting missing species values (Diniz-Filho et al. 2015; Guénard et al. 2013; Penone et al. 2014; Swenson 2014; Swenson et al. 2017). More interestingly, our results showed that some phylogenetic imputation methods (PEM.notrait, MICE.phy) performed better than non-phylogenetic (MICE), even when missing data was not correlated to phylogeny but to an auxiliary trait (MAR.TRAIT). This result suggests that even when missing data is not correlated to phylogeny, phylogenetic information carry valuable informations about missing data.

PEM.notrait and PEM.trait are single imputation method, which imputes a single value for each missing datum, thus not accounting for uncertainty of the imputed value. Consequently, PEM methods may underestimate standard errors and increasing Type I error rates (Enders 2010; Schafer and Graham 2002; van Buuren 2012). To avoid such biases, the statistical literature suggests using multiple imputation methods, as we represented by MICE methods. In spite of we did not evaluated statistical power consequences, our results did not showed better estimation performance of MICE methods, even when including phylogenetic information, while estimating phylogenetic signal compared to PEM. Despite multiple imputation being one of the most suggested methods for handling missing data (van Buuren 2012), additional research is necessary to evaluate its performance on phylogenetically specific question such as phylogenetic signal. MICE and MICE.phy had a tendency to underestimate phylogenetic signal, that was likely caused by the procedure of sampling of values applied by Predictive Mean Matching (PMM) within MICE algorithm, but there are other options in MICE that could be further explored.

Filling missing values with the phylogenetically closer species value (hot-deck) generated poor estimates, except when missing values were phylogenetically structured. Therefore, it is advised to use other imputation methods. Listwise is known to disrupts the distribution of trait values, when not in MCAR, being prone to generate biased estimates (Enders 2010). However, this method performed well when estimating phylogenetic signal, mainly when there was low percentage of missing data. This results resemble Marcondes (2019) and Rabosky (2015), in other context of missing data, but concluding that pruning species without data from phylogenetic comparative analyses may result in accurate estimates. This is encouraging, given that researchers interested in phylogenetic signal usually delete missing values, potentially guaranteeing unbiased estimates.

Phylogenetic imputation is based on the assumption of traits being phylogenetically structured (i.e. showing phylogenetic signal; Swenson (2014)). However, how much a phylogeny can predict traits is dependent on how traits evolved (Diniz-Filho 2001; Diniz-Filho et al. 2015; Guénard et al. 2013), what may show very complex behaviors, such as multiple adaptive optima (Beaulieu et al. 2012b). Accordingly, despite trait evolution was not among the most important issue in our study, it impacted phylogenetic signal estimation. Across our simulated scenarios, estimation errors were smaller when target traits were simulated under Brownian motion than under strong OU processes. Better estimates under BM than OU processes may result from higher trait resemblance and lower variance among species generated when increasing selection strength under OU processes (Butler and King 2004; Hansen 1997b), which eliminates phylogenetic signal. To infer statistics from imputed data, imputations must be done by the a “congenial” model, imputation method must recover the data distribution and relationships to be tested (Meng 1994; van Buuren 2012; von Hippel 2009). This suggests that researchers need to be aware of the appropriate evolutionary model for their traits before using phylogenetic imputation methods. It should be noted, however, that fitting evolutionary models over incomplete data may also bias evolutionary model inference (Slater et al. 2012, but see Marcondes 2019). Thus, further methods have to be developed to impute values and infer evolutionary models under missing data, specially for MAR-TRAIT scenarios where it seems to perform worst as shown by our results.

In addition, estimation errors were dependent on the phylogenetic signal metric. Regardless of the simulated scenario, estimation errors were lower for Moran’s Correlogram than Blomberg’s K. Similarly, (Münkemüller et al. 2012) showed that Moran’s I is less sensible than Blomberg’s K to changes in trait phylogenetic structure even when random noise is added. Blomberg’s K measures a global pattern along a phylogeny, based on observed and expected total trait variance under Brownian motion (Simon P Blomberg et al. 2003), whereas Moran’s Correlogram measures the correlation of trait values within phylogenetic distance classes (Gittleman and Kot 1990). Therefore, changes in total trait variance caused by imputation may not have strong impacts in the first class of the correlogram used here to calculate phylogenetic signal, suggesting Moran’s Correlogram more robust than Blomberg’s K to imputations.

Imputation methods have been proposed to fill sparse databases concerning about their degree of imputation error, meaning how much imputed values deviate from the original trait values (Guénard et al. 2013; Penone et al. 2014; Schrodt et al. 2015). We found a non-linear relationship between imputation error and phylogenetic signal errors, thus small increases in imputation errors can generate high errors of statistic estimation. It is in accordance to (Kim et al. 2018) that found more accurate machine learning method does not mean more accurate functional diversity estimates. This is worth to highlight as machine learning is often used to fill missing data in macroecological analyses (Oliveira et al. 2016; Penone et al. 2014). In fact, applying imputation to database should be concerned to recover the original distribution of the data and the relationship among variables, guaranteeing unbiased statistics estimates, not predicting accurately the missing values (Schafer and Graham 2002; van Buuren 2012).

**Concluding remarks**

Instead of providing imputed trait databases, researcher should explore the raw data, hypothesize about what caused the missing values and evaluate which method works better for the data features and research goals. We have shown that phylogenetic imputation methods can be a tool to deal missing data on phylogenetic signal studies, but it should combine phylogenetic information and other variables to improve imputation outcomes. We encourage researchers to use simulations of their data and imputation methods to find the appropriate solution for their study goals, but focusing on the accuracy of statistics used to test hypothesis, not on imputation accuracy.

**Funding**

This work was supported by CAPES doctoral fellowship, BJT “Science without borders” CNPq grant, MCTIC/CNPq (proc. 465610/2014-5) and FAPEG (proc. 201810267000023).

**Conflict of Interest**

The author declare no conflict or competing interest.

**Ethics approval**

Not applicable

**Consent to participate**

The authors all consent to participate.

**Consent to participate**

The authors all consent to publication.

Availability of data and material

Not applicable

**Code availability**

All scripts of simulations, analyses and graphics can be found at ……

**References**

Barzi, F. (2004). Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology*, *160*(1), 34–45. https://doi.org/10.1093/aje/kwh175

Beaulieu, J. M., Jhwueng, D. C., Boettiger, C., & O’Meara, B. C. (2012a). Modeling stabilizing selection: Expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution*, *66*(8), 2369–2383. https://doi.org/10.1111/j.1558-5646.2012.01619.x

Beaulieu, J. M., Jhwueng, D. C., Boettiger, C., & O’Meara, B. C. (2012b). Modeling stabilizing selection: Expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution*, *66*(8), 2369–2383. https://doi.org/10.1111/j.1558-5646.2012.01619.x

Blomberg, Simon P, Garland, T., & Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, *57*(4), 717–745. https://doi.org/10.1111/j.0014-3820.2003.tb00285.x

Blomberg, Simone P., Rathnayake, S. I., & Moreau, C. M. (2020). Beyond brownian motion and the ornstein-uhlenbeck process: Stochastic diffusion models for the evolution of quantitative characters. *American Naturalist*, *195*(2), 145–165. https://doi.org/10.1086/706339

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference* (2nd ed.). New York, NY: Springer-Verlag. http://linkinghub.elsevier.com/retrieve/pii/S0304380003004526

Butler, M. A., & King, A. A. (2004). Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution APP. *The American Naturalist*, *164*(6), 683–695. https://doi.org/10.1086/426002

Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice : Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3). https://doi.org/10.18637/jss.v045.i03

Cavender-Bares, J., Kozak, K. H., Fine, P. V. A., & Kembel, S. W. (2009). The merging of community ecology and phylogenetic biology. *Ecology letters*, *12*, 693–715. https://doi.org/10.1111/j.1461-0248.2009.01314.x

Diniz-Filho, J. A. F. (2001). Phylogenetic autocorrelation under distinct evolutionary process. *Evolution*, *55*(6), 1104–1109.

Diniz-Filho, J. A. F., Sant’Ana, C. E. R., & Bini, L. M. (1998). An Eigenvector Method for estimating Phylogenetic Inertia. *Evolution*, *52*(5), 1247–1262.

Diniz-Filho, J. A. F., Villalobos, F., & Bini, L. M. (2015). The best of both worlds : Phylogenetic eigenvector regression and mapping. *Genetics and Molecular Biology*, *38*(3), 396–400.

Enders, C. K. (2010). *Applied Missing Data Analysis* (1st ed.). New York, NY.

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist*, *125*(1), 1–15.

Freckleton, R. P., Harvey, P. H., & Pagel, M. (2002). Phylogenetic analysis and comparative data : a test and review of evidence. *The American naturalist*, *160*(6), 712–726.

Freckleton, Robert P, & Jetz, W. (2009). Space versus phylogeny: disentangling phylogenetic and spatial signals in comparative data. *Proceedings of the Royal Society B*, *276*(1654), 21–30. https://doi.org/10.1098/rspb.2008.0905

Gaston, K. J., & Blackburn, T. M. (1994). Are Newly Described Bird Species Small-Bodied ? *Biodiversity Letters*, *2*(1), 16–20.

Gaston, K. J., Chown, S. L., & Evans, K. L. (2008). Ecogeographical rules: elements of a synthesis. *Journal of Biogeography*, *35*, 483–500. https://doi.org/10.1111/j.1365-2699.2007.01772.x

Gillespie, D. (1996). Exact numerical simulation of the Ornstein-Uhlenbeck process and its integral. *Physical Review E*, *54*(2), 2084–2091. https://doi.org/10.1103/PhysRevE.54.2084

Gittleman, J. L., & Kot, M. (1990). Adaptation: statistics and a null model for estimating phylogenetic effects. *Systematic Zoology*, *39*(3), 227–241.

Goldberg, E. E., Kohn, J. R., Lande, R., Robertson, K. a., Smith, S. a., & Igic, B. (2010). Species selection maintains self-incompatibility. *Science*, *330*(493), 493–495. https://doi.org/10.1126/science.1194513

Gonzalez-Suarez, M., Lucas, P. M., & Revilla, E. (2012). Biases in comparative analyses of extinction risk: mind the gap. *The Journal of Animal Ecology*, *81*, 1211–22. https://doi.org/10.1111/j.1365-2656.2012.01999.x

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*, 206–213. https://doi.org/10.1007/s11121-007-0070-9

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*(12), 1120–1131. https://doi.org/10.1111/2041-210X.12111

Hadfield, J. D. (2008). Estimating evolutionary parameters when viability selection is operating. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1635), 723–734. https://doi.org/10.1098/rspb.2007.1013

Hansen, T. F. (1997a). Stabilizing selection and the comparative analysis of adaptation. *Evolution*, *51*(5), 1341–1351.

Hansen, T. F. (1997b). Stabilizing selection and the comparative analysis of adaptation. *Evolution*, *51*(5), 1341–1351.

Hansen, T. F., & Martins, E. P. (1996). Translating between microevolutionary process and macroevolutionary patterns: correlation structure of interspecific data. *Evolution*, *50*(4), 1404–1417.

Harmon, L. J., Losos, J. B., Jonathan Davies, T., Gillespie, R. G., Gittleman, J. L., Bryan Jennings, W., et al. (2010). Early bursts of body size and shape evolution are rare in comparative data. *Evolution*, *64*(8), 2385–2396. https://doi.org/10.1111/j.1558-5646.2010.01025.x

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M., & Ladle, R. J. (2015). Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, *46*, 523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400

Jetz, W., & Freckleton, R. P. (2015). Towards a general framework for predicting threat status of data-deficient species from phylogenetic, spatial and environmental information. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *370*, 20140016. https://doi.org/http://dx.doi.org/10.1098/rstb.2014.0016

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*(9), 2648–2648. https://doi.org/10.1890/08-1494.1

Kattge, J., Ogle, K., Bönisch, G., Díaz, S., Lavorel, S., Madin, J., et al. (2011). A generic structure for plant trait databases. *Methods in Ecology and Evolution*, *2*, 202–213. https://doi.org/10.1111/j.2041-210X.2010.00067.x

Kim, S. W., Blomberg, S. P., & Pandolfi, J. M. (2018). Transcending data gaps: a framework to reduce inferential errors in ecological analyses. *Ecology Letters*. https://doi.org/10.1111/ele.13089

Landis, M. J., & Schraiber, J. G. (2017). Pulsed evolution shaped modern vertebrate body sizes, *0*. https://doi.org/10.1073/pnas.1710920114

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd editio.). New Jersey: John Wiley & Sons.

Lukas, D., & Clutton-Brock, T. H. (2013). The evolution of social monogamy in mammals. *Science*, *341*, 526–30. https://doi.org/10.1126/science.1238677

Marcondes, R. S. (2019). Realistic scenarios of missing taxa in phylogenetic comparative methods and their effects on model selection and parameter estimation. *PeerJ*, *7*(10), e7917. https://doi.org/10.7717/peerj.7917

Meng, X.-L. (1994). Multiple-Imputation inference with Uncogenial Sources of Input. *Statistical Science*, *9*(4), 538–573.

Münkemüller, T., Lavergne, S., Bzeznik, B., Dray, S., Jombart, T., Schiffers, K., & Thuiller, W. (2012). How to measure and test phylogenetic signal. *Methods in Ecology and Evolution*, *3*, 743–756. https://doi.org/10.1111/j.2041-210X.2012.00196.x

Nakagawa, S. (2015). Missing data: mechanisms, methods, and messages. In G. A. Fox, S. Negrete-Yankelevich, & V. J. Sosa (Eds.), *Ecological Statistics: Contemporary Theory and Application* (1st editio., pp. 81–105). Oxford, UK: Oxford University Press.

Nakagawa, S., & Freckleton, R. P. (2008). Missing inaction: the dangers of ignoring missing data. *Trends in Ecology & Evolution*, *23*(11), 592–596. https://doi.org/10.1016/j.tree.2008.06.014

Nakagawa, S., & Freckleton, R. P. (2010). Model averaging, missing data and multiple imputation: a case study for behavioural ecology. *Behavioral Ecology and Sociobiology*, *65*(1), 103–116. https://doi.org/10.1007/s00265-010-1044-7

Oliveira, B. F., Machac, A., Costa, G. C., Brooks, T. M., Davidson, A. D., Rondinini, C., & Graham, C. H. (2016). Species and functional diversity accumulate differently in mammals. *Global Ecology and Biogeography*, *25*(9), 1119–1130. https://doi.org/10.1111/geb.12471

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*(2), 289–290. https://doi.org/10.1093/bioinformatics/btg412

Penone, C., Davidson, A. D., Shoemaker, K. T., Marco, M. Di, Rondinini, C., Brooks, T. M., et al. (2014). Imputation of missing data in life-history traits datasets: which approach performs the best? *Methods in Ecology and Evolution*, *5*(9), 961–970. https://doi.org/10.1111/2041-210X.12232

Purvis, A., Gittleman, J. L., Cowlishaw, G., & Mace, G. M. (2000). Predicting extinction risk in declining species. *Proceeding of the Royal Society B*, *267*, 1947–1952. https://doi.org/10.1098/rspb.2000.1234

R Core Team. (2020). R: A Language and Environment for Statistical Computing. Vienna, Austria. https://www.r-project.org/

Rabosky, D. L. (2015). No substitute for real data : A cautionary note on the use of phylogenies from birth – death polytomy resolvers for downstream comparative analyses. *Evolution*, *62*(12), 3207–3216. https://doi.org/10.1111/evo.12817

Reddy, S., & Dávalos, L. M. (2003). Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, *30*, 1719–1727.

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and Opportunities of Open Data in Ecology. *Science*, *331*(6018), 703–705. https://doi.org/10.1126/science.1197962

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, *3*(2), 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x

Revell, L. J., Harmon, L. J., & Collar, D. C. (2008). Phylogenetic signal, evolutionary process, and rate. *Systematic biology*, *57*(4), 591–601. https://doi.org/10.1080/10635150802302427

Rosado, B. H. P., de S. L. Figueiredo, M., de Mattos, E. A., & Grelle, C. E. V. (2015). Eltonian shortfall due to the Grinnellian view: functional ecology between the mismatch of niche concepts. *Ecography*, *39*(11), 1034–1041. https://doi.org/10.1111/ecog.01678

Rubin, D. . (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Schafer, J. L., & Graham, J. W. (2002). Missing Data: our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.

Schrodt, F., Kattge, J., Shan, H., Fazayeli, F., Joswig, J., Banerjee, A., et al. (2015). BHPMF - a hierarchical Bayesian approach to gap-filling and trait prediction for macroecology and functional biogeography. *Global Ecology and Biogeography*, *24*(12), 1510–1521. https://doi.org/10.1111/geb.12335

Slater, G. J., Harmon, L. J., Wegmann, D., Joyce, P., Revell, L. J., & Alfaro, M. E. (2012). Fitting models of continuous trait evolution to incompletely sampled comparative data using approximate bayesian computation. *Evolution*, *66*, 752–762. https://doi.org/10.1111/j.1558-5646.2011.01474.x

Swenson, N. G. (2014). Phylogenetic imputation of plant functional trait databases. *Ecography*, *37*(July 2013), 105–110. https://doi.org/10.1111/j.1600-0587.2013.00528.x

Swenson, N. G., Weiser, M. D., Mao, L., Araújo, M. B., Diniz-Filho, J. A. F., Kollmann, J., et al. (2017). Phylogeny and the prediction of tree functional diversity across novel continental settings. *Global Ecology and Biogeography*, *26*(5), 553–562. https://doi.org/10.1111/geb.12559

Therneau, T., & Atkinson, B. (2019). rpart: Recursive Partitioning and Regression Trees. https://cran.r-project.org/package=rpart

van Buuren, S. (2012). *Flexible Imputation of Missing Data* (1st ed.). Boca Raton, Fl: Chapman and Hall/CRC. https://doi.org/10.1201/b11826

van Buuren, S., Brands, J. P. L., Groothuis-Oudshoorn, K., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049–1064.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). New York: Springer. http://www.stats.ox.ac.uk/pub/MASS4

Vilela, B., & Villalobos, F. (2015). letsR: a new R package for data handling and analysis in macroecology. *Methods in Ecology and Evolution*, *6*, 1229–1234. https://doi.org/10.1111/2041-210X.12401

Vilela, B., Villalobos, F., Rodríguez, M. Á., & Terribile, L. C. (2014). Body Size, extinction risk and knowledge bias in New World snakes. *PloS one*, *9*(11), e113429. https://doi.org/10.1371/journal.pone.0113429

von Hippel, P. T. (2009). How to impute squares, interactions, and other transformed variables. *Sociological Methodology*, *39*(Rubin 1987), 265–291. https://doi.org/10.1111/j.1467-9531.2009.01215.x

von Hippel, P. T. (2018). How Many Imputations Do You Need? A Two-stage Calculation Using a Quadratic Rule. *Sociological Methods & Research*, 004912411774730. https://doi.org/10.1177/0049124117747303

Webb, C. O., Ackerly, D. D., Mcpeek, M. A., & Donoghue, M. J. (2002). Phylogenies and community ecology. *Annual Review of Ecology, Evolution, and Systematics*, *33*, 475–505. https://doi.org/10.1146/annurev.ecolsys.33.010802.150448

Wiens, J. J., & Graham, C. H. (2005). Niche Conservatism: integrating evolution, ecology, and conservation biology. *Annual Review of Ecology, Evolution, and Systematics*, *36*, 519–539. https://doi.org/10.1146/annurev.ecolsys.36.102803.095431

Wilman, H., Belmaker, J., Simpson, J., de la Rosa, C., Rivadeneira, M. M., & Jetz, W. (2014). EltonTraits 1.0 : Species-level foraging attributes of the world ’s birds and mammals. *Ecology*, *95*(October 2013), 2027.