Syst. Biol. 63(2):203–218, 2014

DOI:10.1093/sysbio/syt100
Advance Access publication December 10, 2013

Missing Data Estimation in Morphometrics: How Much is Too Much?

JULIEN CLAVEL[1,*], GILDAS MERCERON[2], AND GILLES ESCARGUEL[1]

[1]*Laboratoire de Géologie de Lyon, UMR 5276, CNRS, UCB Lyon 1, ENS Lyon, Campus de la Doua, 2 rue Raphaël Dubois, 69622 Villeurbanne Cedex, France; and* [2]*IPHEP, UMR 7262, CNRS & Université de Poitiers, Bat. B35, 6 rue M. Brunet, 86022 Poitiers Cedex, France*
[*]Correspondence to be sent to: E-mail: julien.clavel@univ-lyon1.fr; julien.clavel@hotmail.fr.

Received 22 March 2013; reviews returned 4 June 2013; accepted 4 December 2013
Associate Editor: Norman MacLeod

Abstract.—Fossil-based estimates of diversity and evolutionary dynamics mainly rely on the study of morphological variation. Unfortunately, organism remains are often altered by post-mortem taphonomic processes such as weathering or distortion. Such a loss of information often prevents quantitative multivariate description and statistically-controlled comparisons of extinct species based on morphometric data. A common way to deal with missing data involves imputation methods that directly fill the missing cases with model estimates. Over the last years, several empirically-determined thresholds for the maximum acceptable proportion of missing values have been proposed in the literature, whereas other studies showed that this limit actually depends on various properties of the study data set and of the selected imputation method, and is by no way generalizable. We evaluate the relative performances of seven multiple imputation (MI) techniques through a simulation-based analysis under three distinct patterns of missing data distribution. Overall, Fully Conditional Specification and Expectation–Maximization algorithms provide the best compromises between imputation accuracy and coverage probability. MI techniques appear remarkably robust to the violation of basic assumptions such as the occurrence of taxonomically or anatomically biased patterns of missing data distribution, making differences in simulation results between the three patterns of missing data distribution much smaller than differences between the individual MI techniques. Based on these results, rather than proposing a new (set of) threshold value(s), we develop an approach combining the use of MIs with procrustean superimposition of principal component analysis results, in order to directly visualize the effect of individual missing data imputation on an ordinated space. We provide an R function for users to implement the proposed procedure. [Missing data; morphometrics; multiple imputation; ordination; Procrustes superimposition; R function; simulation.]

Missing data are relatively common in biological sciences such as morphometrics, most particularly when dealing with paleontological or archeological material due to its exposure to taphonomic processes (Holt and Benfer 2000; Strauss et al. 2003, Strauss and Atanassov 2006; Couette and White 2010; Brown et al. 2012). Fossil remains are often altered by post-mortem constraints as shearing, crushing, and breaking, which may dramatically affect extremities and delicate parts of otherwise fossilizable elements (Holt and Benfer 2000; Brown et al. 2012). Such a loss of information is highly penalizing when studying fossil material, where morphology is the only information available to distinguish systematic units and to describe evolutionary dynamics. Multivariate morphometrics (whether linear measurement-based or geometric coordinate-based) are now widely used to account for the morphological (size and shape) variations of homologous objects, but statistically-controlled systematic comparison is prevented when such data are affected by missing values. This limitation is due to the multivariate procedures themselves (e.g., ordination techniques such as Principal Component Analysis (PCA), Discriminant Factor Analysis (DFA), Multivariate Analysis of Variance (MANOVA), or clustering techniques such as Unweighted Pair Group Method with Arithmetic Mean (UPGMA) or Neighbor-Joining), which basically require the use of complete cases (Strauss et al. 2003). Unfortunately, in most studies even a small proportion of missing values can lead to a drastic reduction of the data set. For instance, in Rhode and Arriaza's (2006) study of human cranial measurements, as little as 5% missing data as a whole actually affected 50% of the sampled specimens.

The easiest and most common approach when dealing with missing data is to focus the multivariate analysis on complete cases, removing the missing values by specimen and/or variable deletion—the so-called available-case, or pairwise deletion method. Such an approach seems reasonable when deleted cases are not too abundant, making the resulting restricted sample large enough to reliably reflect the original data set structure and variability, a not-so-frequent situation when dealing with fossil material. Indeed, a variable with a high rate of missing values, leading to its removal from the analysed data set, may actually be highly relevant for the biological characterization of the morphometric space. The primary effects of these sample-size restrictions are a loss of statistical power in analyses performed on such reduced data sets (Cardini and Elton 2007; Nagawaka and Freckleton 2008), and the generation of substantial biases caused by truncating the data set's variability structure (e.g., ecologic or taxonomic). For instance, some fossil specimens (e.g., from smaller or slender species) are frequently more taphonomically altered than other ones, and thus less well represented in morphometric data sets (Behrensmeyer et al. 1979, 2000; Holt and Benfer 2000; Soligo and Andrews 2005; Andrews 2006; Le Fur et al. 2009, 2011; Brown et al. 2013). Accordingly, Strauss and Atanassov (2006) proposed a pairwise selection technique to determine the best complete subset of specimens that optimizes the statistical properties of the submatrix with respect to the original data set. This approach provides the most stable statistical results, but there is no guarantee that the selected data produce the most meaningful biological interpretation (Strauss and Atanassov 2006).

Alternatively, few multivariate methods allowing for the direct use of incomplete data sets are currently available, all of them showing low statistical power. For instance, Feldesman (2002) proposed to use a non-parametric method such as binary recursive partitioning, which does not require special treatment of the missing values, instead of linear discriminant analysis in classification/prediction problems. A distance-based approach using Gower's distance modified for missing data (Gower 1971) can also be worked out, either in an ordination or clustering problem, or in a non-parametric MANOVA-like design (Anderson 2001; Legendre and Legendre 2012). Nevertheless, this approach is known to give biased results when compared with estimation methods (Brown et al. 2012).

To overcome most of these problems, several imputation procedures have been proposed in order to replace the missing values by estimated scores. The basic techniques are single-imputation (SI) methods such as mean substitution or regression-based predictions. Unfortunately, these approaches tend to reduce descriptors' variances and covariances, and to erase inter-specimen relations as a consequence of data homogenization (Holt and Benfer 2000; Feldesman 2002; Nagawaka and Freckleton 2008).

Over the last three decades, more efficient and sophisticated computational methods of imputation, using Maximum Likelihood (ML) and Expectation–Maximization (EM) algorithms, have been developed (Dempster et al. 1977) and tested on morphometric data sets (Strauss et al. 2003; Couette and White 2010; Brown et al. 2012). As these methods directly estimate parameters by accounting for missing data distribution, and because of the highly inter-correlated nature of morphometric variables (Escarguel 2005), imputation methods are well-suited for dealing with missing data in morphometrics and clearly outperform non-estimation approaches in most cases (Brown et al. 2012). Unfortunately, they show sample specific, hardly generalizable characteristics. For instance, the upper limit for an acceptable percentage of missing values in a given morphometric data set (i.e., the threshold over which the estimation methods experience a dramatic error increase) ranges from 15% (Couette and White 2010) and 20% (Holt and Benfer 2000) up to 50% (Strauss et al. 2003), or to different values depending on the imputation technique and on the pattern of missing data distribution within the analysed data set (Brown et al. 2012).

Although ML and EM SI techniques show generally less biased estimates than mean or regression imputation and non-estimation approaches in simulation studies (e.g., Brown et al. 2012), the reliability of imputed specimens remains difficult to assess in real-case studies. This raises a major issue: how to know if an imputed value adds more information than noise in a given data set? Powerful and more flexible multiple imputation (MI) techniques were developed to address this issue (Rubin 1987; Schafer 1997; Schafer and Olsen 1998). Briefly, MI is operating as a combination of m (>1) SIs of the same data set through Monte Carlo procedures. The resulting variability of estimated values between the m imputed data sets (not available with SI techniques) accounts for the uncertainty in the missing data estimates with respect to the "true", but unknown, missing value. Whereas applications of MI in the medical and social sciences have been widely discussed (Little and Rubin 1989; Rubin and Schenker 1991; van Buuren et al. 1999; van Buuren 2007; Marshall et al. 2009, 2010; Spratt et al. 2010), applications in morphometrics remain scarce, with most of the few studies using MI focusing on an averaged imputed data set and not taking into account the specific information provided by the m imputed data sets (e.g., Rhode and Arriaza 2006; Botha and Angielczyk 2007; Glantz et al. 2009; Athreya and Raj 2010, Bernal et al. 2010).

Rather than providing a new (set of) threshold limit(s), the present study reports results of simulations that aimed to assess the performance of different MI techniques to estimate the error associated with the imputation procedure. Our approach relies on the fact that most multivariate morphometric analyses are done by ordination of individuals within reduced spaces rather than through modeling studies where MI is used to estimate the reliability of the model regression coefficients. We illustrate our approach by estimating individuals' uncertainty due to the imputed values within a reduced multivariate space, using seven distinct MI techniques, and based on an initially complete, artificially altered morphometric data set. The simulated patterns of missing data distribution were elaborated using random distribution as well as more realistic non-random processes (Brown et al. 2012). Thus, the present study aims at comparing distinct MI techniques in terms of accuracy and precision of the imputation-based individual coordinates in a reduced space. Finally, we discuss a graphical method to display individuals' confidence intervals in a reduced space, and thus to visually assess the error introduced by MIs on incomplete specimens.

Materials

We focused our study on linear measurement rather than geometric coordinate data because: (i) linear measurement data sets including missing values have been extensively elaborated and published over the last few decades in both paleontological and archeological contexts, generating a huge amount of paleobiological information still waiting, in most cases, for proper imputation protocols prior to in-depth quantitative analyses; and (ii) more fundamentally, linear measurements allow the construction of logarithmic morphometric spaces where hypotheses

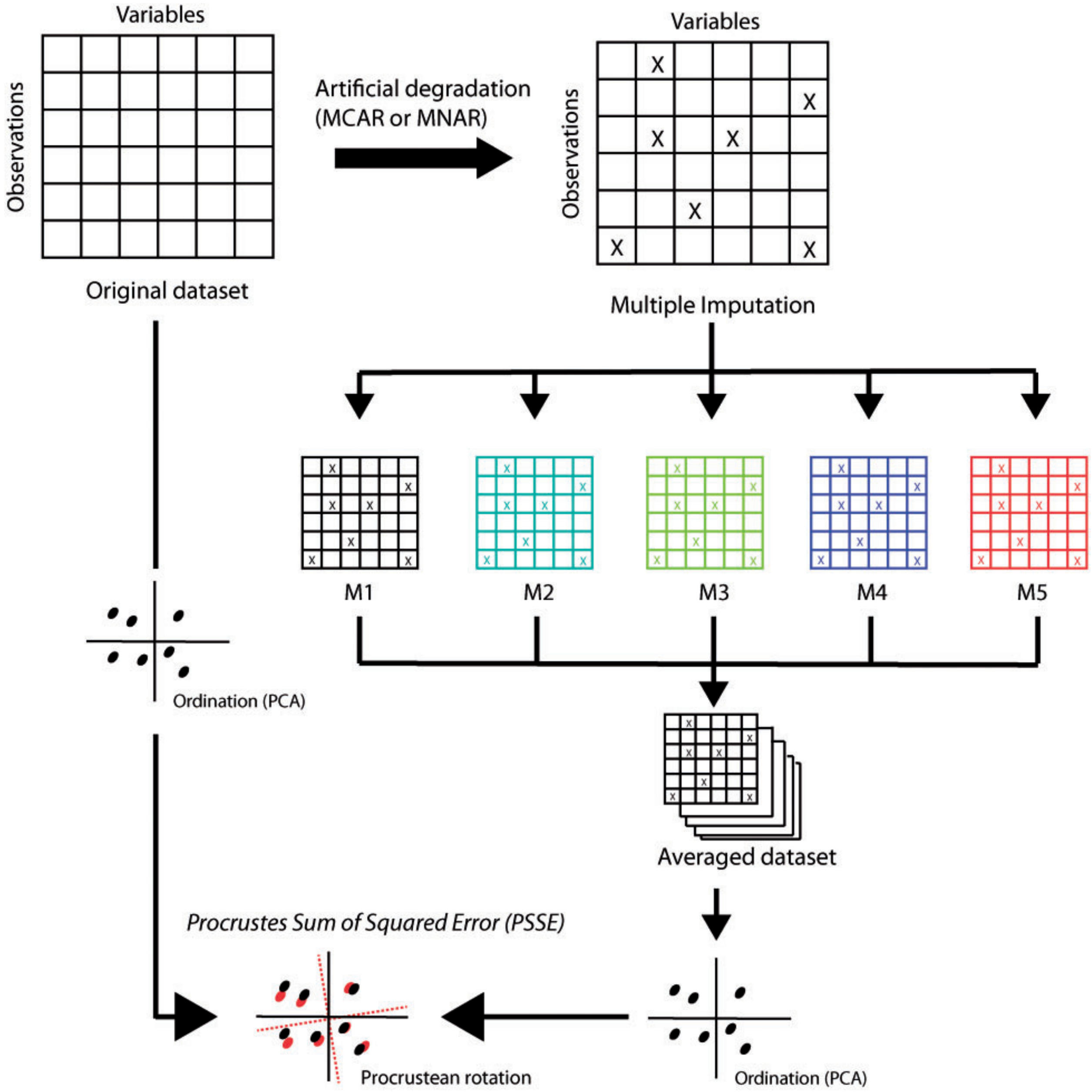

FIGURE 1. Flowchart of the procedure used to compute the PSSE between the original (complete) data set and the averaged MI-data set estimated by a given MI technique.

of proportional (isometric) or non-proportional (allometric) growth and evolutionary changes can be directly tested (Houle et al. 2011; Voje and Hansen 2013).

MI-based individual confidence intervals in a reduced space were generated using the recent crocodilian crania data set elaborated by Brown et al. (2012). This complete data set contains 226 specimens and 23 variables representing linear cranial measurements (Brown et al. 2012: fig. 1). As it is mainly composed of recent specimens, it is well-defined taxonomically, which allows for testing the effect of taxonomic bias in the pattern of data removal. In addition, sampled specimens and taxa span a large range of body size; this allows the testing of imputation method performances when allometry occurs. This makes this data set particularly well-suited to artificially introduce various distributional patterns of missing values and to evaluate the robustness of MI techniques to such data loss (Brown et al. 2012). Although Schafer and Olsen (1998) suggest to log-transform data before the imputation procedure in order to restore normality of right-skewed variables, we performed analyses on raw measurements: (i) for consistency with the previous study by Brown et al. (2012); and (ii) to assess the performances of the MI methods when allometry occurs. As a first attempt to evaluate the relative performance of MI methods with a smaller data set (i.e., lower specimen/variable ratio, making MI methods potentially less accurate due to the loss of among-specimen variability information), this initial data set was randomly reduced to 80 specimens to form a second data set. This size was empirically determined for this data set as the lowest possible one for which most MI techniques still produced results up to ~25% of missing values.

Methods

All simulations and statistical analyses were performed on a SGI Altix Xe 320 Cluster with a Linux platform, using a 64-bit version of the open access R-2.15.0 (Ihaka and Gentleman 1996; R Development Core Team 2005). R libraries used in the imputation simulations are: MICE (van Buuren and Groothuis-Oudshoorn 2011), Hmisc (Harrell 2012), Amelia II (Honaker et al. 2011), missMDA (Josse et al. 2011), Norm (Novo 2009), and pcaMethods (Stacklies et al. 2007). All these libraries propose different MI techniques listed below. The processes of missing data introduction prior to imputation simulations were performed using the "sample" function of the base package, and the LOST package (Brown et al. 2012). All these packages are freely available online from the CRAN (Comprehensive R Archive Network; http://www.r-project.org) and on Bioconductor for the pcaMethods package (http://www.bioconductor.org). The use of various methods combined in the R software provides an easy way for all researchers to use the imputation methods tested here in their own work.

Missing Data Mechanism

All the simulations were conducted under MCAR (missing completely at random) or MNAR (missing not at random) mechanisms of data loss. The MCAR mechanism was implemented using the "sample" function from the base package in R (Matsumoto et al. 2006). Two distinct MNAR mechanisms were implemented using the functions "obliterator" and "byclade" from the LOST package (Brown et al. 2012) in order to simulate more realistic, anatomically or taxonomically biased distribution patterns of missing data distribution. An anatomical bias ("obliterator" function) was generated by preferentially degrading anatomical regions as defined by their spatial organization, based on the 3D-coordinate's landmarks identifying the starting and end points of each morphometric measurement. These 3D coordinates were acquired on a skull of *Crocodylus porosus* (R6142) stored in the Cambridge University Museum of Zoology. A taxonomic bias was generated by primarily degrading taxa that are less represented in the sample. For each imputation method, data removal was independently repeated 1000 times for each missing data pattern (one MCAR and two MNAR) by degrading the complete sample by steps of 5% up to 50% missing values.

TABLE 1. Summary of imputation methods evaluated and compared in this study

Methods	Description	R function	R package
BPCA	SI with Bayesian PCA	pca (bpca)	pcaMethods
Mice-pmm	MI with PMM	mice (pmm)	mice
Mice-norm	MI with Bayesian linear regression	mice (norm)	mice
Hmisc-pmm	MI with PMM	aregImpute (pmm)	Hmisc
Hmisc-reg	MI with additive regressions and bootstrap	aregImpute (regression)	Hmisc
Amelia	MI with EM and bootstrap	amelia	Amelia
MI-PCA	MI with PCA and bootstrap	MIPCA (regularized)	missMDA
Norm	MI with DA, 50 iterations	imp.norm	norm

SI, single imputation; MI, multiple imputation; EM, Expectation–Maximization

Imputation Methods

One SI and seven MI methods were evaluated and compared (Table 1). Methods with different mathematical strategies and imputation models were selected in order to evaluate the relative efficiency of the main current ways to produce MIs. The MI techniques we used include:

- a data augmentation (DA) approach, assuming a joint multivariate normal distribution (Schafer 1997) as implemented in the Norm package;
- two techniques imputing incomplete multivariate data by Fully Conditional Specification (FCS), one with a Bayesian regression switching ("norm") and one with predictive mean matching after regression switching ("pmm"), both implemented in the MICE V2.0 package (van Buuren and Groothuis-Oudshoorn 2011);
- an approach using the bootstrap-based EM algorithm as available in the Amelia II package (King et al. 2001; Honaker et al. 2011);
- an approach using a PCA model with a bootstrap procedure on the residuals in order to reflect the uncertainty of the unknown parameters, using the missMDA package (Josse et al. 2011);
- the Hmisc package (Harrell 2001, 2012) which uses additive regression, bootstrapping, and a predictive mean matching (PMM) method with weighted probability sampling of donors.

In addition, we considered only one SI technique, the Bayesian PCA, using the function "bpca" in the pcaMethods package (Stacklies et al. 2007). This method gave the best results among three SI techniques (*a priori* size regression, correlated variable regression, and Bayesian PCA) in a previous study based on the same complete data set (Brown et al. 2012); it is used here as a baseline for comparison with MI techniques.

The number of dimensions considered when using the MI-PCA technique has a crucial impact on the accuracy of the estimates (Josse et al. 2011). Hence, simulations were performed from two (lowest value) up to six dimensions (Supplementary Material S1.1, available at http://dx.doi.org/10.5061/dryad.f0b50), but results are presented here only for two dimensions, representing the simplest parameterization of the imputation model. This choice is justified with respect to most real-case studies where the complete data set is unknown, preventing the use of a simulation-based approach to identify the number of principal components which simultaneously optimize accuracy and coverage probability.

For all MI techniques, simulations were run by setting the number of MIs to $m=5$, and then to $m=20$ in order to check outputs for stability. Simulation results from these two m-values were qualitatively similar, so we illustrate hereafter only the simulation outputs for $m=5$ (results for $m=20$ are given as Supplementary Material S1.3, available at http://dx.doi.org/10.5061/dryad.f0b50), a small value suggested by most previous publications on MI (e.g., Rubin 1987; Schafer 1997; Schafer and Olsen 1998; Little and Rubin 2002, and references therein; but see results below, suggesting that at least in some cases, $m=5$ may be too small to satisfactorily estimate the error introduced by the imputation procedure).

Parameter estimates in the EM method worked out by the Amelia II package (Table 1) are heavily constrained by the ratio between the numbers of specimens and variables. Here, the maximum proportion of missing data reached in the full, 226-specimen data set was 25% for the MCAR mechanism and the MNAR mechanism under taxonomic bias, and 30% for the MNAR under anatomic bias. For the same reason, in the reduced, 80-specimen data set, this algorithm stopped before reaching a sufficient number of replicates, and results are thus not reported. Similarly, the algorithm used in Hmisc package stopped before reaching a reasonable number of replicates at a maximum proportion of missing data under MNAR/anatomic bias of 40% with "Hmisc-pmm" and 35% for "Hmisc-reg" in the full data set. With the reduced data set, "Hmisc-reg" reached 20% of missing data under all data removal mechanisms, whereas "Hmisc-pmm" reached 20% under MNAR/anatomic bias and 25% under MCAR and MNAR/taxonomic bias. We also get an insufficient number of replicates for "Mice-norm" and "Norm" on the reduced data set with more than 45% and 25% of missing data under MNAR/anatomic bias, respectively.

Accuracy of the Imputation Methods

In order to estimate the relative ability of the studied MI techniques to generate accurate estimates of the missing data, the m imputed data sets obtained by a given method were first averaged to form a single imputed data set. A PCA was then performed on this averaged imputed data set and on the original (unaltered) data set. The two PCA results were finally compared in order to evaluate the overall error produced by the imputation procedure. Rather than the mean squared error between the two sets of PCA coordinates (e.g., Strauss et al. 2003; Couette and White 2010), the PCA coordinates were compared through Procrustes transformation in order to obtain an optimal superimposition between the two multivariate configurations (Fig. 1). The Procrustes sum of squared error (PSSE) was then calculated as the sum of the Procrustes residuals to determine a goodness-of-fit statistic (Gower and Dijksterhuis 2004; Schneider and Borlund 2007). Such an approach is commonly used for transformed-matrix comparisons (e.g., Peres Neto and Jackson 2001; Rohlf 2003); it is particularly well-suited to quantitatively compare the effect on the PCA results of imputation procedures with different patterns of missing data distribution (Brown et al. 2012). Both original and imputed matrices were rescaled after initial translation to achieve a standardized PSSE ranging between 0 and 1 and directly reflecting the fit between the two PCA configurations (Schneider and Borlund 2007).

Estimating the MI-SE

To assess the accuracy of the MI estimates, that is, to reflect the uncertainty introduced in individuals' PCA coordinates by the imputation procedure, the combining rules of Rubin (1987) were adapted to the PSSE in order to calculate the standard error (SE) of each simulation. The between-imputation variance is first calculated as the sum of squared differences between the averaged MI-data set PCA coordinates and the PCA coordinates of each of the m MI-data sets. Based on MI basic principles (Rubin 1987; Schafer 1997; Schafer and Olsen 1998; Little and Rubin 2002), such between-imputation variance is positively related to the PSSE value between the PCA coordinates of the averaged imputed data set and the original (unaltered) data set.

First, a new Procrustes sum of squared error statistic (hereafter called PSSE′) is calculated between each of the m MI-data sets PCA configurations and the averaged MI-data set PCA configuration (Fig. 2). Then, the between-imputation variance is obtained as

$$B=\frac{1}{m-1}\sum_{i=1}^{m}\mathrm{PSSE}_i'^2,$$

and the within-imputation variance is

$$\overline{U}=\frac{1}{m}\sum_{i=1}^{m}U_i,$$

where $U_i=\frac{\mathrm{PSSE}_i'}{n-1}$, and n is the number of specimens in the data set. Hence, the total variance T is calculated as

$$T=\overline{U}+\left(1+\frac{1}{m}\right)B.$$

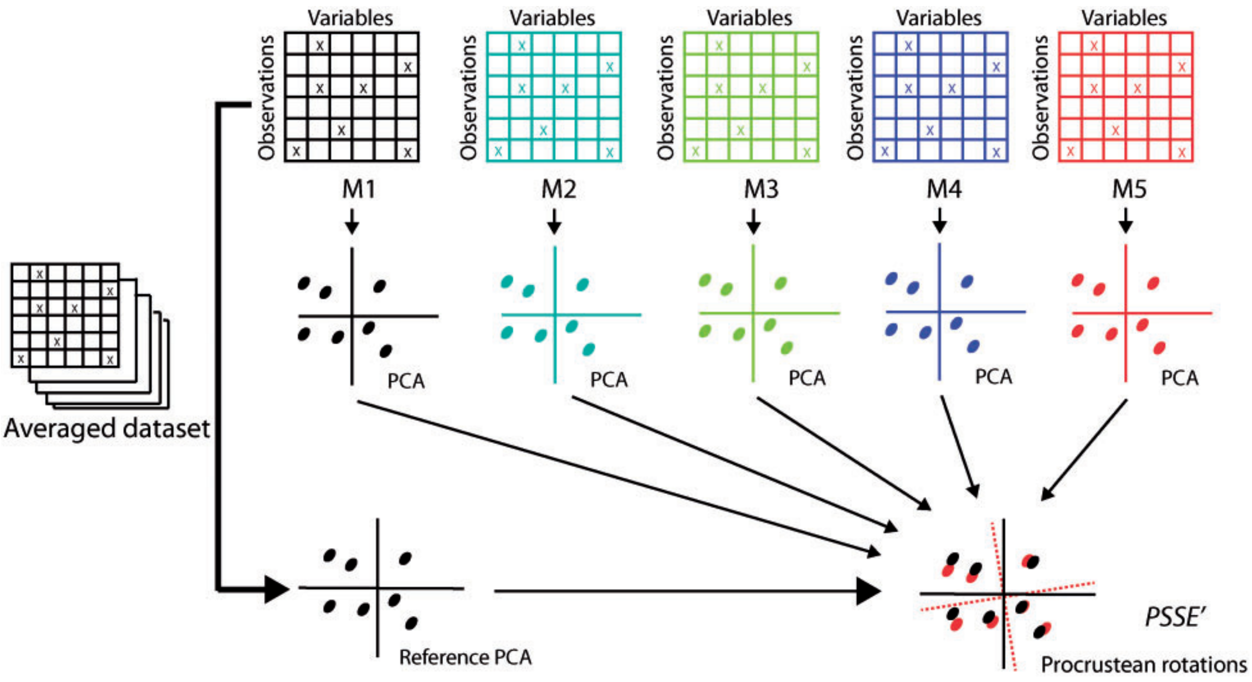

FIGURE 2. Flowchart of the procedure used to compute the PSSE′ between the averaged MI-data set and the m MI-data sets estimated by a given MI technique. This procedure allows estimating the SE of missing value estimates due to the MI technique.

Finally, the SE is calculated as SE $= \sqrt{T}$, and a ~95% confidence interval is obtained as $\pm 2\sqrt{T}$.

Analysis of Simulation Results

In all simulations, PSSE was recorded for each of the 1000 replicates, as well as the total variance for each MI technique. The median PSSE for each set of 1000 replicates and for each simulated proportion of missing data was reported, and plotted against the proportion of missing data for each imputation method under the three simulated patterns of missing data distributions.

To assess whether or not the 95% confidence intervals of individuals' PCA coordinates (based on an averaged MI-data set PCA) contain the expected individuals' PCA coordinates (computed from the original data set), we estimated the overall coverage probability, for each of the seven MI techniques and percentage of missing data, by counting the proportion of 95% confidence intervals over the 1000 replicates that include the expected individual's PCA coordinates. Such a coverage estimate is a mean relative proxy of the MI technique ability to provide accurate estimates of the missing values. Indeed, as further discussed below, a high coverage value will relate to an efficient MI technique only if associated with a low PSSE value. Conversely, a high coverage combined with a high PSSE will actually illustrate a poor MI result where 95% confidence limits around average imputed values encompass a large range of the data, including original values. Alternatively, a low coverage combined with a high PSSE will point to an MI technique which steadily returns imputed coordinates systematically departing from the expected ones.

Representation of the Estimated Error

Although the computation of the PSSE and coverage probability values associated with the PCA configuration of an averaged MI-data set allows for the comparison of the different MI techniques and an overall estimate of the reliability of the MI results, this approach does not identify those individual specimens with unreliable PCA coordinates due to MI inconsistencies. Indeed, in most paleontological or archeological morphometric analyses, populations are represented by few specimens for whom it may be interesting to evaluate the uncertainty introduced by the imputation procedure. In ordination techniques as currently used in morphometric studies (e.g., PCA), the level to which an introduced error can be considered acceptable directly depends on the proportion of total variance explained by the considered components, as well as the biological meaning of a potential shift in the reduced space. Procrustes superimposition enables a graphical assessment of the matching between two or more configurations (Peres Neto and Jackson 2001; Schneider and Borlund 2007). Here we propose to use ordinary Procrustes analysis (OPA) by matching the configuration of each of the m imputed data sets to the configuration of the averaged MI-data set (Peres Neto and Jackson 2001). In this way, it becomes possible to visualize the relative uncertainty introduced by the imputation procedure (Josse et al. 2011). An R function (Supplementary Material S2, available at http://dx.doi.org/10.5061/dryad.f0b50) was developed in order to display confidence ellipses representing the dispersion of the m imputed data sets around the configuration of the averaged MI-data set after

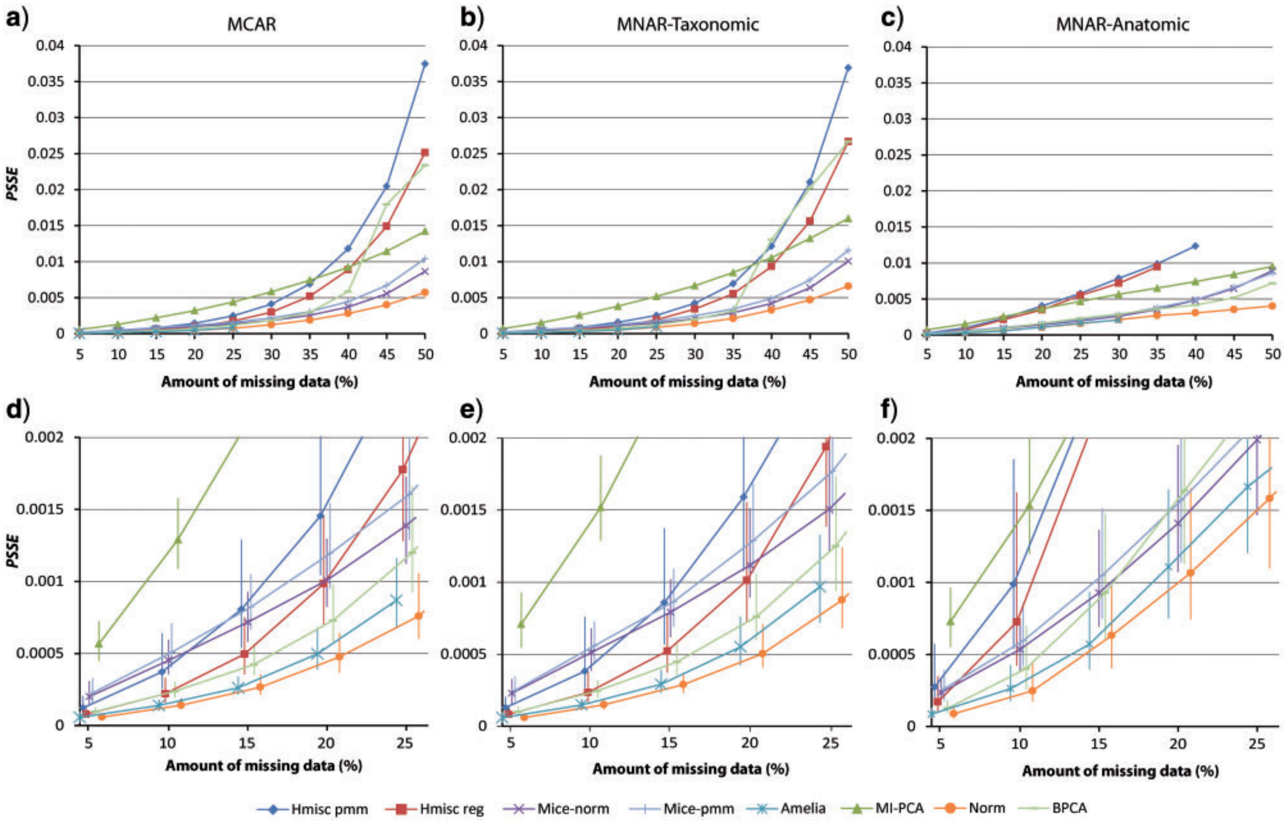

FIGURE 3. Median of the distributions of the PSSE based on 1000 simulation replicates of the full (226 specimens) data set for each MI technique ($m=5$) and percentage of missing data under the 3 simulated patterns of missing data distributions: MCAR (a), MNAR under taxonomic bias (b), and MNAR under anatomic bias (c); (d–f) insets of (a–c) for better visualization, including 1st–3rd quartile confidence intervals (decays between symbols for each percentage of missing data for graphical convenience only).

Procrustes superimposition. For illustration purposes, simulations were done on the full (226 specimens) and reduced (80 specimens) data sets by introducing 10% of missing data under MNAR/anatomic bias in both cases.

Testing the Effect of the Number m *of MIs*

As suggested by some authors (e.g., Horton and Lipsitz 2001; Graham et al. 2007), more than $m=5$ imputations as classically proposed (e.g., Rubin 1987) should be performed. Stability of MI estimates for PCA configurations of morphometric data was evaluated by further simulations for different number of MIs. An arbitrary level of 10% missing data was first randomly removed. Then, the incomplete data set was imputed 1000 times with a new starting seed value at each iteration for $m=5, 10, 15, 20, 30,$ and 50 imputations. This simulation was repeated for all the MI techniques, and associated PSSE were reported in each case.

RESULTS

Relative Performances of the Imputation Methods

The simulation results using the different imputation methods are presented over the three patterns of missing data distribution (MCAR, MNAR/anatomic bias, and MNAR/taxonomic bias) for the full (226 specimens) and reduced (80 specimens) data sets. All the imputation methods show similar results with missing data introduced at random (MCAR) or under MNAR/taxonomic bias, indicating that the between-specimen distribution of missing values does not affect the behavior of these MI techniques (Fig. 3). On the other hand, results based on an MNAR/anatomic bias pattern of missing data distribution show a different relation between PSSE and the percentage of missing data, with steeper relations below 25% of missing values and flatter relations above 30%, leading to slightly larger PSSE values below 40% of missing values, and markedly smaller PSSE values above.

Based on the full data set, the MI techniques “Amelia”, “Norm”, and “Mice”, show a relatively similar response for the three patterns of missing data distribution, whereas the SI BPCA technique shows a lower error level under MNAR/anatomic bias than under MCAR or MNAR/taxonomic bias from 35% of missing data onward (Fig. 3). The two “Hmisc” techniques show an important amount of error when compared with other MI techniques under MNAR/anatomic bias from 10% of missing data onward (Fig. 3c,f). Under this mechanism of missing data distribution, they fail to estimate the error values beyond 35% and 40% of missing data

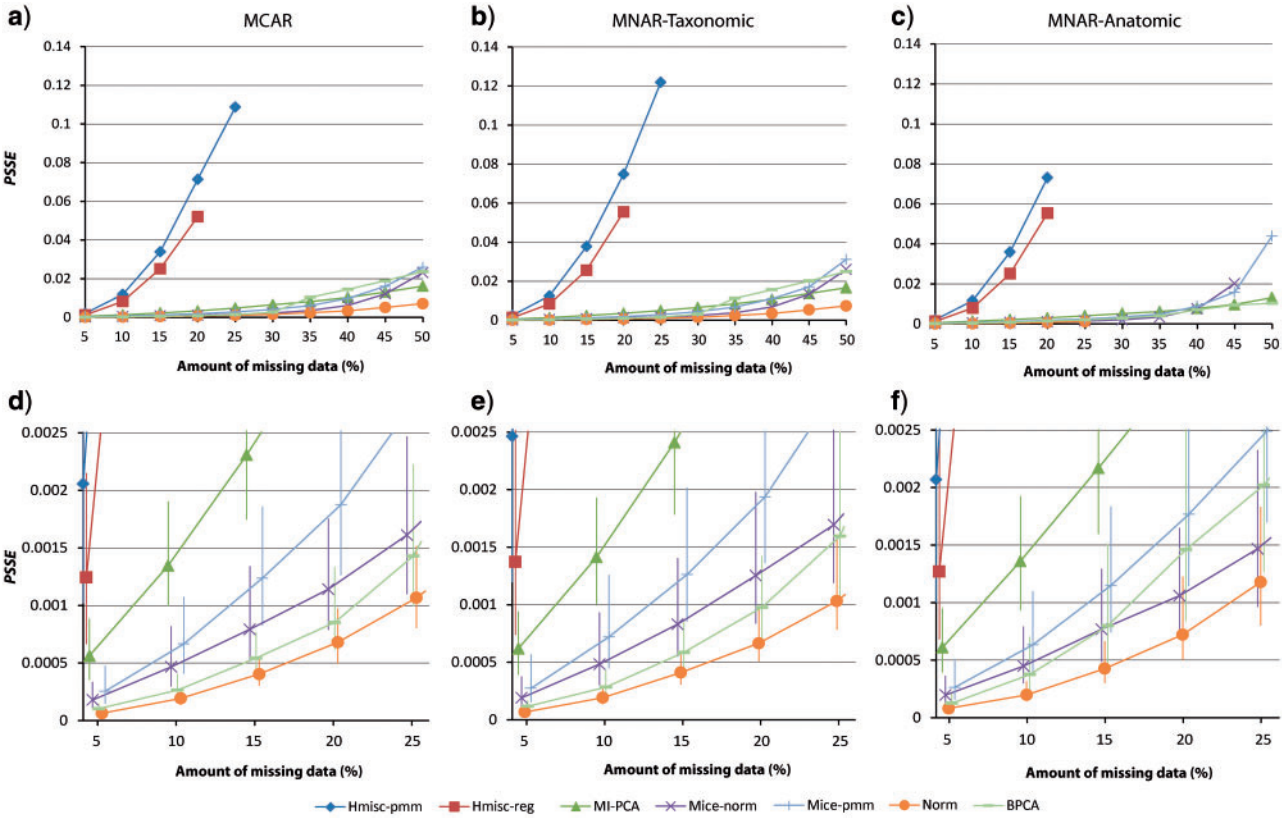

FIGURE 4. Median of the distributions of the PSSE based on 1000 simulation replicates of the reduced (80 specimens) data set for each MI technique ($m=5$) and percentage of missing data under the 3 simulated patterns of missing data distributions: MCAR (a), MNAR under taxonomic bias (b), and MNAR under anatomic bias (c); (d–f): insets of (a–c) for better visualization, including 1st–3rd quartile confidence intervals (decays between symbols for each percentage of missing data for graphical convenience only).

for “Hmisc-reg” and “Hmisc-pmm”, respectively (see Materials and Methods section). Moreover, they show a dramatic increase in error from 35% of missing data onward under MCAR (Fig. 3a,d) and MNAR/taxonomic bias (Fig. 3b,e) missing data distributions. Using only the two first components, the MI-PCA technique generate the largest amount of errors, while estimated errors when more than four principal components are used show values in the same range as “Amelia”, “Norm”, and “Mice” (Supplementary Material S1.1, available at http://dx.doi.org/10.5061/dryad.f0b50). Different numbers of DA iterations were tested for the “Norm” technique, with indistinguishable error differences. Reduced data set-based simulations show similar relative efficiencies between imputation methods over the three patterns of missing data distribution (Fig. 4; see Methods section for computations using the “Amelia” technique), except for the two “Hmisc” techniques which show highly increased errors, and for the two “Mice” techniques above 40% of missing values.

Coverage

For both full (226 specimens) and reduced (80 specimens) data sets, the coverage probability, that is, the probability that any individual’s expected PCA coordinate (based on the analysis of the complete data set) falls within the 95% confidence interval associated with its averaged MI-data set PCA coordinate, was calculated for each MI technique and proportion of missing data under the three missing data distribution patterns (Fig. 5). In all cases, all MI techniques show a coverage probability well above the nominal 95% level when dealing with a very low amount of missing data (5%; in all cases corresponding to very low PSSE values, thus indicating a high fit between original and average imputed values).

Based on the full data set, the MICE (“pmm” and “norm”) and Amelia imputation methods show coverage probabilities above the 95% level for the three distribution patterns of missing data (Fig. 5a–c). The “Hmisc-pmm” technique remains around the 95% nominal level for MCAR and MNAR/taxonomic bias distribution patterns (Fig. 5a,b), but falls dramatically under 90% with >5% of missing data under MNAR/anatomic bias (Fig. 5c). On the other hand, the “Hmisc-reg” technique shows the worst coverage probabilities, with coverage under 95% with only 15% of missing data under MCAR and MNAR/taxonomic bias distribution patterns, and 10% under an MNAR/anatomic bias pattern. The “Norm” technique shows a coverage probability below 95% with >20% of missing data under an MNAR/anatomic bias distribution pattern, and with >30% for MCAR and

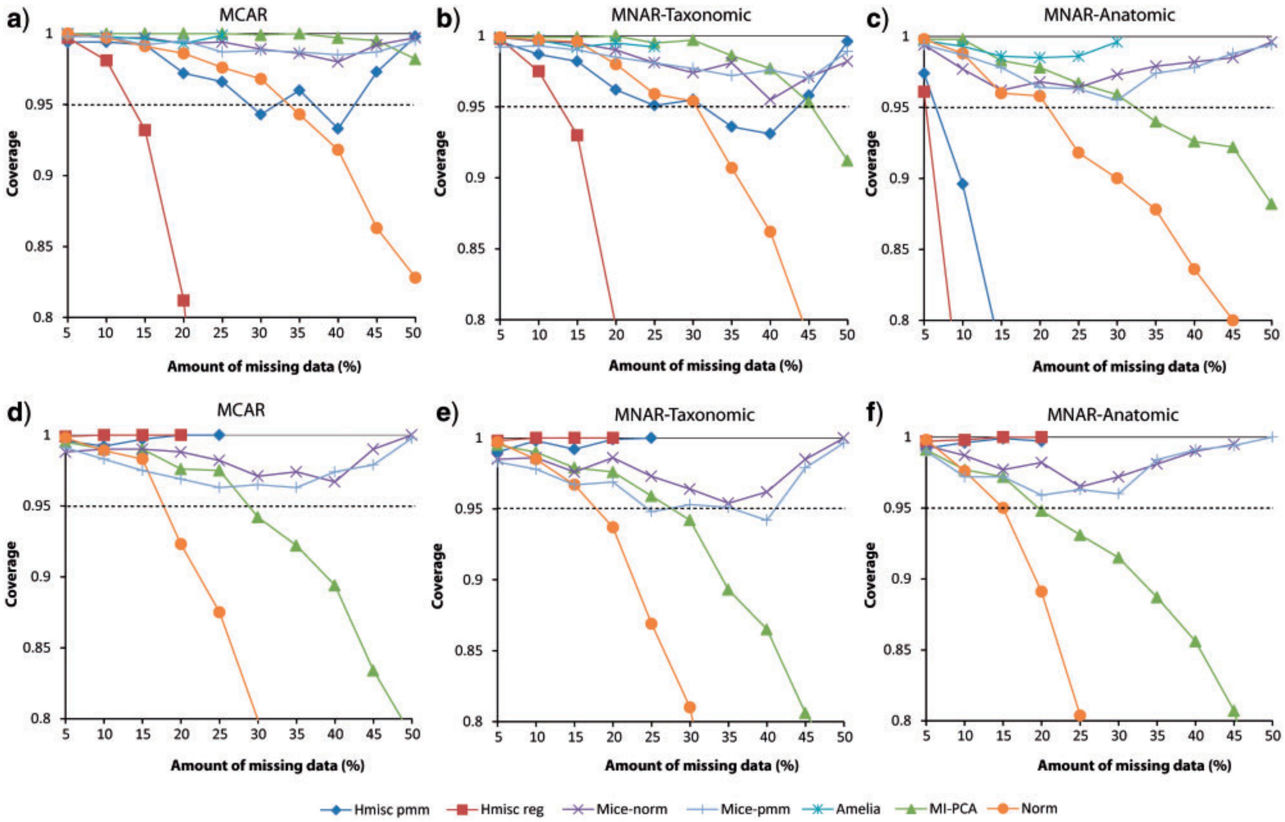

FIGURE 5. Evolution of the coverage probability in relation to the amount of missing data, based on 1000 simulation replicates of the full (a–c, 226 specimens) and reduced (d–f, 80 specimens) data sets for each MI technique ($m=5$) and percentage of missing data under the 3 simulated patterns of missing data distributions: MCAR (a, d), MNAR under taxonomic bias (b, e), and MNAR under anatomic bias (c, f).

MNAR/taxonomic distribution patterns. Finally, the MI-PCA technique has a good coverage probability under an MCAR mechanism, but falls below 95% with >40% and >30% of missing data under MNAR/taxonomic and MNAR/anatomic biases patterns, respectively.

Coverage probabilities calculated for the reduced data set (Fig. 5d–f) show that the “Norm” technique falls below 95% from 20% of missing data onward under the three missing data distribution patterns. “MI-PCA” coverage probability falls below 95% from 30% of missing data onward under an MCAR and MNAR/taxonomic bias pattern and from 20% with missing data onward under an MNAR/anatomic bias pattern. Coverage probabilities obtained for the “Mice-pmm” and “Mice-norm” techniques are similar to those obtained for the full data set (Fig. 5a–c); however, “Hmisc-pmm” and “Hmisc-reg” techniques show here very high coverage probabilities (100%), contrary to those obtained for the full data set.

Superimposition of MI Configurations

As an example of the use of MI to assess the reliability of imputed specimens in a multivariate analysis (here, a PCA), Figure 6 shows the 95% individual confidence ellipses calculated from Procrustes superimposition of 20 imputed configurations obtained with the “Mice-pmm” and “Hmisc-pmm” techniques on the reduced and complete data sets with 10% of missing data under an MNAR/anatomic bias distribution pattern of missing values. For both data sets, the “Hmisc-pmm” technique shows larger 95% confidence ellipses relatively to the “Mice-pmm” technique, leading to a very high overall coverage probability (Supplementary Materials S1.3 and S1.4, available at http://dx.doi.org/10.5061/dryad.f0b50). When combined with a larger median PSSE value for this MI technique, this indicates that “Hmisc-pmm” actually does a relatively poor job in estimating missing values in this example, generating large uncertainties in absolute and relative specimens’ locations within the ordinated space. Conversely, the “Mice-pmm” technique returns small 95% confidence ellipses for most of the imputed specimens, associated with a relatively small median PSSE value and large overall coverage probability (Supplementary Materials S1.3 and S1.4, available at http://dx.doi.org/10.5061/dryad.f0b50). This indicates that this MI technique estimates missing values leading to accurate absolute and relative specimens’ locations within the ordinated space. Ultimately, for any given imputed data set and MI technique, this simple graph allows the identification of the specimens with uncertain PCA coordinates due to MI, making possible an iterative removal procedure in order to improve the

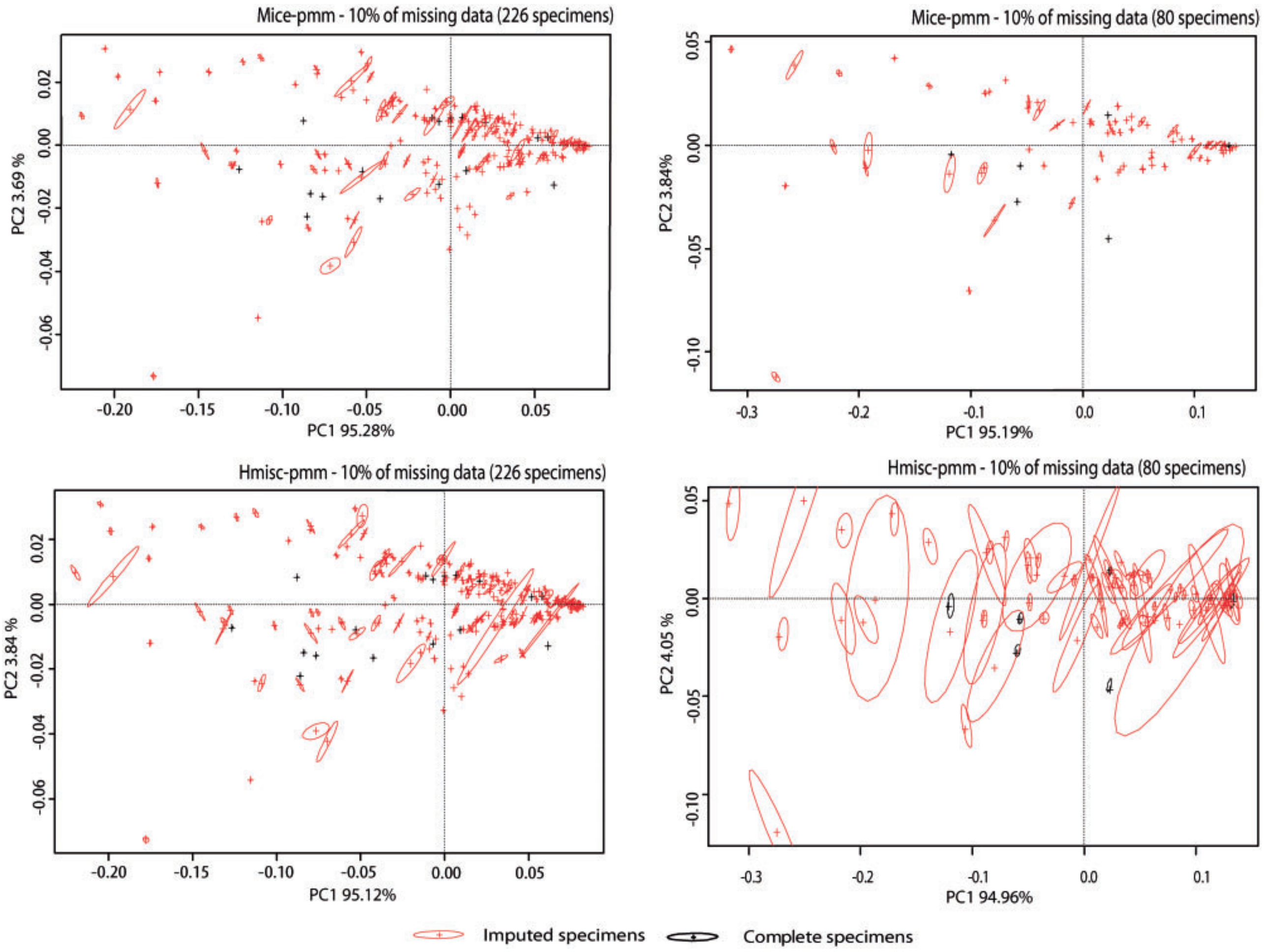

FIGURE 6. Examples of averaged MI-data set PCA results (first principal planes) showing the projections of the 95% confidence ellipses associated with each imputed specimen for the full (left) and reduced (right) data sets with 10% missing data introduced under an MNAR-anatomic bias distribution pattern and missing values estimated using the MICE-pmm (top) and Hmisc-pmm (bottom) techniques with $m = 20$ imputations.

overall reliability of the PCA result with a minimum loss of individuals.

Number of MI

The stability of the estimated missing values was tested depending on m, the number of MIs used for all MI techniques (Fig. 7). Simulations with 10% of missing data under an MCAR pattern of missing data distribution show that the variability of the PSSE decreases with increasing m. In most cases, stability is reached with a minimum of ~20 imputed data sets. Using $m = 5$ shows increased errors in average, except with techniques using predictive mean matching ("Hmisc-pmm" and "Mice-pmm"), where the median PSSE value for each set of simulations appears independent of m (Fig. 7).

Based on this result, we re-run all the simulations with $m = 20$ in order to check the relations between the percentage of missing values, the PSSE and the coverage probability for stability (Supplementary Material S1.3, available at http://dx.doi.org/10.5061/dryad.f0b50). Results show that increasing m generates more accurate estimates (i.e., lower PSSE and higher coverage probability), but does not modify the relative behavior and quality of the various MI techniques.

DISCUSSION

Missing data are a relatively common phenomenon in morphometrics, particularly with archeological and paleontological material, leading a few authors to challenge missing data estimation through simulation-based studies (Holt and Benfer 2000; Strauss et al. 2003; Couette and White 2010; Brown et al. 2012). These previous works focused on the error increase in relation to the proportion of missing data, in order to propose some rule-of-thumb thresholds (e.g., an increase of more than three times in magnitude of the recorded error is used to decide at which proportion of missing data the estimations became unreliable; Strauss et al. 2003). As noted by Brown et al. (2012), error thresholds were empirically derived, but in turn, effects of such errors on multivariate (e.g., ordination) analyses were never assessed. Moreover, such thresholds actually

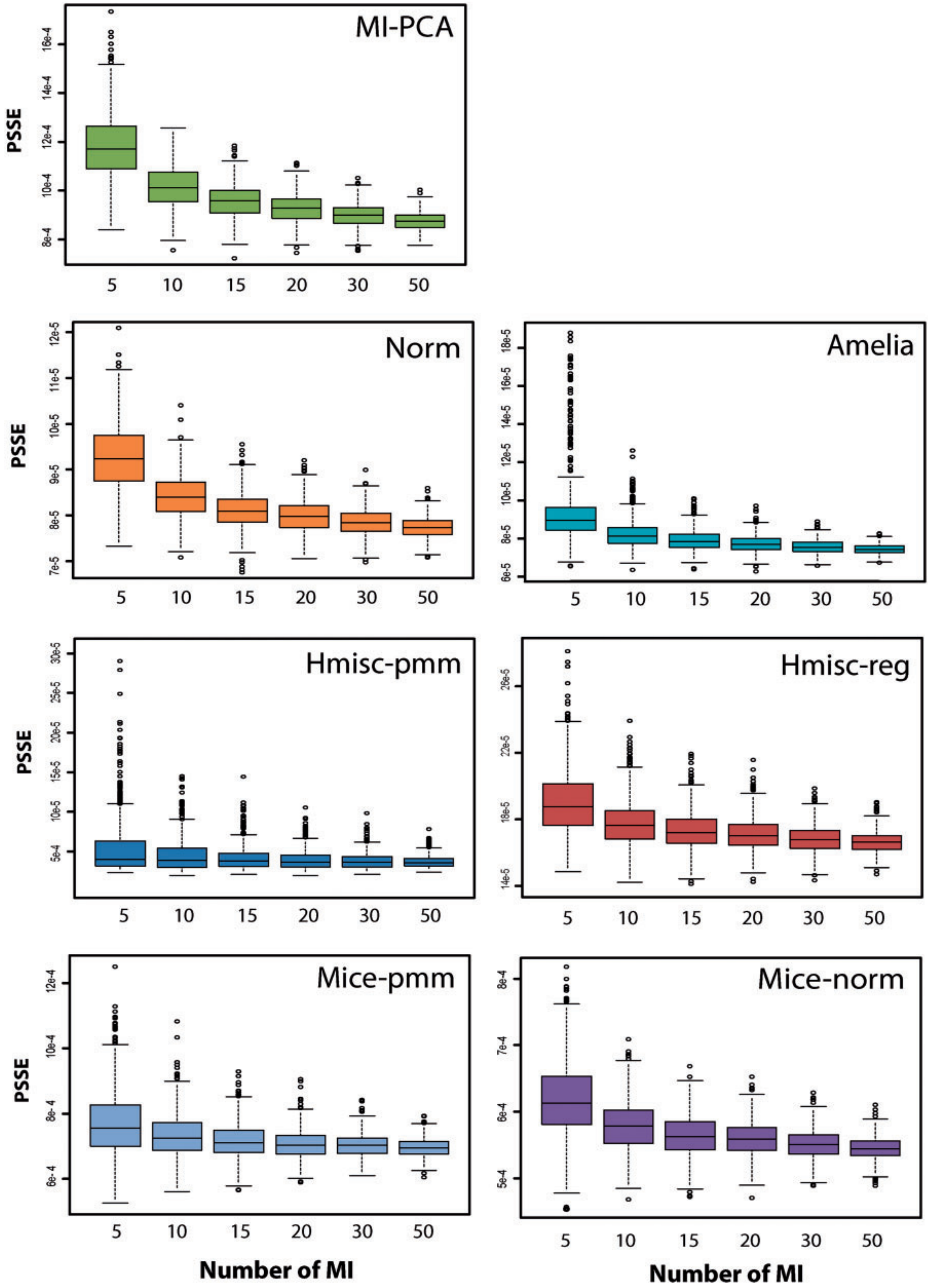

FIGURE 7. Distributions of the PSSE depending on the number of MIs (m), based on 1000 simulation replicates of the full (226 specimens) data set for each MI technique and 10% missing data under an MCAR pattern of missing data distribution.

depend on the data set configuration over which they have been estimated, since the numbers of variables and specimens, the size and shape-range covered, the distribution pattern of missing data, as well as the imputation method itself, have dramatic effects on these limits (Strauss et al. 2003; Brown et al. 2012). Instead of defining some new, more or less arbitrary threshold limit(s), our study focuses on the use of MIs as a tool to evaluate the uncertainty in missing data estimation due to the relative abundance and distribution pattern of missing information in an incomplete data set. Hence, our objectives are to determine: (i) the relative accuracy and precision (coverage probability) of different MI methods; (ii) the effect of the number of MI used; and (iii) how to assess individual's reliability of missing data estimates.

MI Accuracy

As also evidenced by Brown et al. (2012) based on the same crocodile data set analysed through three SI techniques, our simulation results show important differences in estimation accuracy, most particularly depending on the imputation method, and to a lesser extent on the distribution pattern of missing data (MCAR or MNAR with a taxonomical or anatomical bias; Figs. 3 and 4; Supplementary Materials S1.2–S1.4, available at http://dx.doi.org/10.5061/dryad.f0b50). Remarkably, the two MNAR processes impact the MI accuracy differently. On the one hand, MNAR/taxonomic bias shows similar effects to the MCAR process (as also observed for SI methods by Brown et al. 2012), leading to exponential-like relations between the percentage of missing values and PSSE. On the other hand, the MNAR/anatomic bias pattern of missing data distribution generates more linear relations, corresponding to PSSE values markedly lower than MCAR and MNAR/taxonomic bias values above ~40% of missing data, for all MI techniques but regression-based ones (the two "Hmisc" techniques). These results are consistent with Brown et al.'s (2012) observations, and are the logical consequence of a high level of among-variable covariation within the analysed data set. As covariation relates to information redundancy between variables, on average a taxonomical bias more heavily and non-randomly affects variables' distributions, and thus imputation accuracy than an anatomical bias.

Indeed, in all three distribution patterns of missing data, the two "Hmisc" techniques return quickly increasing PSSE with increasing percentage of missing values, making it impossible to provide missing value estimates above 20–25% missing values when working on the reduced (80 specimens) data set, and above 35–40% missing values under MNAR/anatomic bias when working on the full (226 specimens) data set. This illustrates the much greater sensitivity (lower robustness) of these two regression-based MI techniques to the amount of missing values and level of anatomic bias in the pattern of missing data distribution, precluding an accurate estimation of the overall structure of covariance in relatively small and/or incomplete data sets.

Overall, in both the full and reduced data sets, our simulations show that all MI techniques but the two "Hmisc" algorithms perform as well as, or even better than "BPCA" (the SI technique generating the best estimation results *fide* Brown et al. 2012) for all three distribution patterns of missing data. "BPCA" achieves accuracy levels (PSSE values) comparable to the two "Mice" techniques in all cases but highly degraded data sets (≥40% of missing values) under MCAR and MNAR/taxonomic bias, where the MI techniques clearly outperform "BPCA". Better estimations are reached by the "Amelia" and "Norm" MI techniques, both using an EM algorithm, than by "BPCA" for all amounts and distribution patterns of missing data. Nevertheless, "Amelia" appears more sensitive than "Norm" to the number of specimens relative to the number of variables, as well as to the percentage of missing data. Hence, "Amelia" works well for the full data set with <25% of missing data, but crashes before reaching a substantial number of simulation replicates on the reduced data set.

MI Coverage of the Expected Values

Although it is easy to derive from several MI-data sets a unique (averaged) imputed data set that can be analysed as an SI data set, the goal of MI is not primarily to produce values as close as possible to the original. Instead, MI is a simulation-based approach which handles missing data in a way allowing for valid statistical inference (Rubin 1987). From that point of view, MI is particularly appropriate to assess the uncertainty introduced by missing values into a multivariate analysis of the imputed data set. This is done through combination of the MI-data set results, producing confidence intervals that incorporate this uncertainty. The main goal of this study was to explore the coverage of the expected values by the confidence intervals generated by MI techniques. Simultaneous comparisons of the relative accuracies and coverage probabilities of the seven studied MI techniques allow us to determine which technique displays the best compromise of statistical performances (Supplementary Materials S1.2 and S1.3, available at http://dx.doi.org/10.5061/dryad.f0b50).

Coverage of the expected values by the MI techniques was calculated for the full and reduced data sets (Fig. 5). Results indicate that "Hmisc" methods are highly sensitive to sample size and percentage of missing values, and tend to generate very large SEs in highly corrupted situations—the reason why a ~100% coverage is maintained with the reduced data sets until the algorithms collapse in spite of large, exponentially increasing associated PSSE. These results may be due to: (i) the inappropriateness of the regression-model assumptions; (ii) the non-monotonic pattern of missing data distribution when using sequential regression-based methods; (iii) a diffusion effect of the Procrustes error on specimens with no missing data (Siegel and Benson 1982); or (iv) a combination of these three explanations. Our simulation study shows more extreme results for the "Hmisc-reg" technique since missing values are imputed by randomly adding a residual quantity that can dramatically affect the variability range. For instance, with only 10% missing data in the reduced data set, "Hmisc-pmm" MI configurations show large confidence ellipses when compared with MI configurations obtained with "Mice-pmm" on the same altered data set (Fig. 6).

For the "Norm" and "MI-PCA" techniques, the coverage probability decreases in all cases when the percentage of missing data increases, especially under MNAR/anatomic pattern of missing data distribution, where the PSSE also more strongly increases with the percentage of missing data (Figs. 3 and 5; Supplementary Materials S1.2–S1.4, available at http://dx.doi.org/10.5061/dryad.f0b50). The loss of parameters with the reduction of the sample size tends to decrease the coverage probabilities of these techniques (Fig. 5d–f). The "Norm" technique is a joint modeling approach that involves specifying a multivariate distribution for the missing data (Schafer 1997). For simulation convenience, a "non-informative" prior distribution was used since in the vast majority of data analyses, prior ignorance about model parameters works well (Schafer and Olsen 1998). However, prior assumption of multivariate normal distribution is frequently violated with small samples, sparse data, or a high percentage of missing data (Schafer 1997; Schafer and Olsen 1998). Coverage results obtained here are consistent with this, and highlight the need to choose an informative prior distribution in such cases in order to avoid a spurious underestimation of MI-based variability reflecting the estimate uncertainty.

Similarly, the "MI-PCA" technique provides solutions that are dependent on the number of dimensions chosen a priori and the uncertainty is reflected by a bootstrap procedure on the residuals of the last dimensions of the PCA (Josse et al. 2011). Accordingly, the number of dimensions to be used is crucial and should be estimated a priori by cross-validation for a specific data set although this approach is computationally expensive. An important drawback with this technique is the need for a compromise between gain in accuracy and decrease in coverage probabilities, which is associated with overfitting (particularly with sparse data) when too many parameters have to be estimated from the observed data (Ilin and Raiko 2010). In contrast, the two "Mice" techniques show relatively high coverage probabilities for all percentages of missing data in the full as well as reduced data sets. "Mice" use a FCS approach which specifies on a variable-by-variable basis the multivariate model for each incomplete variable by a set of conditional densities (van Buuren and Groothuis-Oudshoorn 2011). This approach may perform better than joint modeling approaches such as "Norm" when no suitable prior distributions can be found. However, "Mice" techniques experience greater increase in error (PSSE) than "Norm" when dealing with large percentages of missing data (Figs. 3 and 4).

Number of MIs

Noting the effect of the number of imputations on the stability of missing value estimates is not a new result (e.g., Horton and Lipsitz 2001; Graham et al. 2007). Our simulations show that not only the variability of the estimate is affected by the number of MI, but also the median value of the error estimate (Fig. 7). This means that the number of MI not only impacts the variability of missing data estimates, but also that a low number of MI may affect the accuracy of the estimates. Interestingly, PMM-based methods "Hmisc-pmm" and "Mice-pmm" display relative stability of the mean error estimates with respect to the number of MI. This relative stability is probably due to the semi-parametric nature of the PMM algorithm, which imputes only values that are actually observed, making this approach particularly useful to preserve nonlinear relations within the data set. One of the interesting properties of using PMM is that no "illegal" values are created (e.g., negative values when extrapolation is done as with regression methods). In contrast, PMM produces artificially low SEs by reducing the between-imputation variability when the number of predictors is small.

In our study, results show that more (say, at least 20 with as few as 10% of missing data; Fig. 7) than $m=5$ imputations must be used in MI procedures to reach reasonably stable inferences. Nevertheless, our simulation results (relative accuracy and coverage probability) remain qualitatively unchanged with higher m-values (Supplementary Material S1.3, available at http://dx.doi.org/10.5061/dryad.f0b50), suggesting that this parameter does not affect the relative behavior and quality of the compared MI techniques. Thus, instead of proposing here a threshold value for m, we rather suggest the use of a relative high number of MI (e.g., $m=100$), because computational time is no longer a heavy constraint, unless one is working with huge data sets (Marshall et al. 2009).

Graphical Representation of the MI Error

Procrustes analysis is a classical tool for measuring and comparing the match between multivariate data sets (e.g., Gower 1966; Peres Neto and Jackson 2001; Schneider and Borlund 2007). It is often used to derive a goodness-of-fit statistic that summarizes the overall resemblance between two configurations (here, the sum of squared errors between a complete and an imputed configuration). However, one of the unique advantages in using Procrustes analysis is its ability to enable a graphical assessment of individual's matching between the two data set configurations. This particularity is useful in order to graphically display the confidence ellipse associated with each imputed specimen by scaling the PCA configurations of the m MI-data set to the averaged MI-data set PCA configuration by procrustean rotations (Fig. 6). Such an approach can be seen as a way to display the effects of measurement error biases on a given analysis (e.g., measurement of the same data set by different observers). Displaying the MI-data set configurations on the reference data set allows visualization of how an individual is affected by the imputation procedure in the reduced space relative to the variation explained by each principal component. The use of MI also allows depiction of the relative magnitude in change of an individual position depending on the uncertainty of the imputation. In other words, as estimation's variability in MI depends on the missing information rate contained in the imputed data set, projections of the MI-data sets give information about the confidence one can attribute to an individual point estimate. Large dispersion of imputed values for a given specimen with missing data can occur, especially when the amount of missing data in that specimen is large and/or the sample is small, or when missing values are located in variables that explain most of the variation in the principal component(s) under scrutiny.

The least-square fitting criterion used for matching the PCA configurations can allocate changes in all objects, even those without imputed data (Siegel and Benson 1982). As a result, SE could be inflated and the coverage probability overestimated in simulations with large amount of missing data, resulting in large incongruence in estimations and matching of the configurations (as possibly observed for the "Hmisc-pmm" results). In such a case, overestimation is a confounding factor when estimating coverage probabilities and makes it difficult to assess graphically the impact of imputed data with several overlapping confidence ellipses. It may be noted, however, that uncertainty around imputed individuals may remain larger than uncertainty around complete specimens. Instead of using Procrustes fitting, Josse et al. (2011) proposed to consider the projection of the MI-data sets onto the reference data set as supplementary elements in order to assess the stability of the individuals due to missing values. However, this approach does not allow assessment of the stability of eigenvalues and vectors due to the uncertainty of the estimates. Caution must also be taken when using projection of Procrustes superimposition of the m configurations when some principal components share close eigenvalues, due to axis reordering (Jackson 1995; Peres-Neto et al. 2003). Nevertheless, Peres-Neto et al. (2003) have provided a method to address this issue.

Finally, when dealing with a large amount of missing data, it may be useful first to "optimize" the data set, when possible, using preliminary stepwise procedures that minimize the loss of data and the number of missing values to estimate (Strauss and Atanassov 2006). Indeed, considerable errors appear unavoidable when about half of the data or more need to be estimated. On the other hand, the benefit of using imputed data over complete case analysis may be questionable when there are few missing data in a large data set, since the analysis of an imputed data set may not give any additional information in that case (Marshall et al. 2010). Nevertheless, even if not abundant in the analysed data set, the specimens with missing values may be of special interest, thus legitimating data set imputation. In such a case, the graphical approach discussed here, by allowing the assessment of individual's variability due to the imputation procedure, may be particularly useful, all the more when dealing with relatively small data sets, as is frequently the case in archeological or paleontological studies.

Conclusion

As discussed at length in the literature about SI or MI techniques, the goal of imputation is not to estimate the exact value of the missing measurements, but rather to estimate these values in such a way that descriptive parameters of the data set are preserved, and thus subsequent complete-data analytical methods are minimally affected by the missing data points. Whereas imputation accuracy necessarily decreases when the percentage of missing values increases, this relation directly depends on the selected imputation technique and the study data set (Brown et al. 2012; this study). Indeed, most cutoffs already proposed in the morphometric literature are affected by several parameters such as the sample size, the distribution of missing cases, and even the range of allometric/isometric scaling (Strauss et al. 2003; Brown et al. 2012). In addition, most paleontological data sets are basically incomplete, and frequently include specimens and data without extant relatives, thus precluding any simulation-based threshold estimation.

We show here that MI techniques can be used in morphometric studies to display the uncertainty of the estimation due to the rate and distribution of missing values in the analysed data set. Moreover, Procrustes superimposition (Gower 1971; Schneider and Borlund 2007) allows the computation and graphical visualization of this uncertainty into the multivariate analysis result (Josse et al. 2011). Ultimately, visualization of the MI-induced individual confidence intervals on the ordination axes allows assessment of the reliability of the relative location of imputed specimens in the ordinated space, based on individual's distribution of this MI-induced error on the ordination axes, and on the biological relevance of these axes. In this way, MI can be seen as a simulation method that generates itself, for any given analysed data set, its own estimation threshold by directly identifying those imputed data points with unreliable locations in the ordinated space.

Our simulation results show that MI techniques are as efficient as, if not slightly more efficient than, SI techniques in producing accurate estimates of missing values, and are also remarkably robust to the violation of some basic assumptions such as the occurrence of taxonomically or anatomically biased patterns of missing data distribution. Clearly, differences in simulation results among the three compared patterns of missing data distribution are much smaller than differences among the seven compared MI techniques. From this point of view, each MI technique can be seen as an element of a sensitivity analysis, altogether representing a panel of compromise between relative accuracy in parameter estimates and coverage probability. In terms of accuracy, the "Norm" (for both full and reduced data sets) and "Amelia" (for the full data set with <25–30% of missing values) techniques show the best results (lowest PSSE values), whereas in terms of coverage probability, the two MI techniques from the "Mice" library produce the best results with the reduced data set, but are outcompeted by the "Amelia" technique below 25–30% of missing values when applied to the full data set.

Based on these results, an approach can be suggested through graphical investigation of the effects of missing-value imputation on the PCA results, looking at, and iteratively removing, those imputed specimens with excessively large 95% confidence ellipses. In a few iterations, a minimal number of badly imputed specimens should be eliminated from the analysed data set in order to optimize the overall accuracy of the imputation procedure. To facilitate this procedure, we provide an R function automatically computing: (i) the average MI-data set from *m* MI-data sets; (ii) the Procrustes superimposition of the *m* MI-data sets on the average MI-data set PCA result; and (iii) the 95% confidence ellipses associated with each analysed specimen (Supplementary Material S2, available at http://dx.doi.org/10.5061/dryad.f0b50).

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.f0b50.

Acknowledgments

We thank Fabien Dubuffet for his help on computational procedures with the cluster under his care. We are grateful to Caleb M. Brown for discussions and data sharing. We thank Lionel Hautier for 3D data acquisition on a crocodile skull and Mathew Lowe, curator of the zoological museum of Cambridge. We thank Philippe Fortin for his technical help, and Allowen Evin, Raphaël Cornette and the "Plateforme morphometrie" (UMR 5202, National Museum of Natural History, Paris, France), for the loan of a MicroScribe. We thank Caleb M. Brown, Donald A. Jackson and an anonymous reviewer, as well as Norman MacLeod (Associate Editor) and Frank E. Anderson (Editor in Chief) who provided suggestions that improved this article.

References

Anderson M.J. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 26:32–46.

Andrews P. 2006. Taphonomic effects of faunal impoverishment and faunal mixing. Palaeogeogr. Palaeoclimatol. Palaeoecol. 241:572–589.

Athreya S., Raj R. 2010. A rare tribal (adivasi) burial from the lower Narmada River valley at Rampura, Gurajat, western India. Anthropol. Sci. 118:151–158.

Behrensmeyer A.K., Kidwell S.M., Gastaldo R.A. 2000. Taphonomy and Paleobiology. Paleobiology 26:103–147.

Behrensmeyer A.K., Western D., Dechant Boaz D.E. 1979. New perspectives in vertebrate paleoecology from recent bone assemblage. Paleobiology 5:12–21.

Bernal V., Perez I.S., Gonzalez P.N., Sardi M.L., Pucciarelli H.M. 2010. Spatial patterns and evolutionary processes in southern South America: a study of dental morphometric variation. Am. J. Phys. Anthropol. 142:95–104.

Botha J., Angielczyk K.D. 2007. An integrative approach to distinguishing the Late Permian dicynodont species *Oudenodon bainii* and *Tropidostoma microtrema* (Therapsida: Amonodontia). Palaeontology 50:1175–1209.

Brown C.M., Arbour J.H., Jackson D.A. 2012. Testing of the effect of missing data estimation and distribution in morphometric multivariate data analyses. Syst. Biol. 61:941–954.

Brown C.M., Evans D.C., Campione N.E., O'Brien L.J., Eberth D.A. 2013. Evidence for taphonomic size bias in the Dinosaur Park Formation (Campanian, Alberta), a model Mesozoic terrestrial alluvial-paralic system. Palaeogeogr. Palaeoclimatol. Palaeoecol. 372:108–122.

van Buuren S. 2007. Multiple imputation of discrete and continuous data by fully conditional specification. Stat. Methods Med. Res. 16:219–242.

van Buuren S., Boshuizen H.C., Knook D.L. 1999. Multiple imputation of missing blood pressure covariates in survival analysis. Stat. Med. 18:681–694.

van Buuren S., Groothuis-Oudshoorm K. 2011. Mice: multivariate imputation by chained equations in R. J. Stat. Softw. 45:1–67.

Cardini A., Elton S. 2007. Sample size and sampling error in geometric morphometric studies of size and shape. Zoomorphology 126:121–134.

Couette S., White J. 2010. 3D geometric morphometrics and missing-data. Can extant taxa give clues for the analysis of fossil primates? C. R. Palevol. 9:423–433.

Dempster A.P., Laird N.M., Rubin D.B. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). J. Roy. Statist. Soc. Ser. B. 39:1–38.

Escarguel G. 2005. Mathematics and the lifeway of *Mesopithecus*. Int. J. Primatol. 26:801–823.

Feldesman M.R. 2002. Classification trees as an alternative to linear discriminant analysis. Am. J. Phys. Anthropol. 119:257–275.

Le Fur S., Fara E., Mackaye H.T., Vignaud P. 2009. The mammal assemblage of the hominid site TM266 (Late Miocene, Chad Basin): ecological structure and paleoenvironmental implications. Naturwissenschaften 96:565–574.

Le Fur S., Fara E., Vignaud P. 2011. Effect of simulated faunal impoverishment and mixture on the ecological structure of modern mammal faunas: implications for the reconstruction of Mio-Pliocene African palaeoenvironments. Palaeogeogr. Palaeoclimatol. Palaeoecol. 305:295–309.

Glantz M., Athreya S., Ritzman T. 2009. Is Central Asia the eastern outpost of the Neandertal range? A reassessment of the Teshik-Tash child. Am. J. Phys. Anthropol. 138:45–61.

Gower J.C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325–338.

Gower J.C. 1971. A general coefficient of similarity and some of its properties. Biometrics 27:857–871.

Gower J.C., Dijksterhuis G.B. 2004. Procrustes problems. Oxford: Oxford University Press.

Graham J.W., Olchowski A.E., Gilreath T.D. 2007. How many imputations are really needed? Some practical clarifications of Multiple Imputation theory. Prev. Sci. 8:206–213.

Harrell F.E. 2001. Regression modeling strategies with applications to linear models, logistic regression, and survival analysis. New York: Springer.

Harrell F.E. 2012. Hmisc: Harrell Miscellaneous library for R statistical software. Version 3.12-2.

Holt B., Benfer R.A. 2000. Estimating missing data: an iterative regression approach. J. Hum. Evol. 39:289–296.

Honaker J., King G., Blackwell M. 2011. Amelia II: a program for missing data. Version 1.7.2.

Horton N.J., Lipsitz S.R. 2001. Multiple Imputation in practice: comparison of software packages for regression models with missing variables. Amer. Statist. 55:244–254.

Houle D., Pélabon C., Wagner G.P., Hansen T.F. 2011. Measurement and meaning in biology. Q. Rev. Biol. 86:3–34.

Ihaka R., Gentleman R. 1996. R: a language for data analysis and graphics. J. Comput. Graph. Stat. 5:299–314.

Ilin A., Raiko T. 2010. Practical approaches to principal component analysis in the presence of missing values. J. Mach. Learn. Res. 11:1957–2000.

Jackson D.A. 1995. PROTEST: a procrustean randomization test of community environment concordance. Ecoscience 2:297–303.

Josse J., Pagès J., Husson F. 2011. Multiple imputation in principal component analysis. Adv. Data Anal. Classif. 5:231–246.

King G., Honaker J., Joseph A., Scheve K. 2001. Analysing incomplete political science data: an alternative algorithm for Multiple Imputation. Am. Pol. Sci. Rev. 95:49–69.

Legendre P., Legendre L. 2012. Numerical ecology. Amsterdam: Elsevier Science.

Little R.J.A., Rubin D.B. 1989. The analysis of social science data with missing values. Sociol. Methods Res. 18:292–326.

Little R.J.A., Rubin D.B. 2002. Statistical Analysis with missing data. 2nd ed. New York: J. Wiley & Sons.

Marshall A., Altman D., Holder R., Royston P. 2009. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med. Res. Methodol. 9:57.

Marshall A., Altman D.G., Royston P., Holder R.L. 2010. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. BMC Med. Res. Methodol. 10:1–16.

Matsumoto M., Saito M., Haramoto H., Nishimura T. 2006. Pseudorandom number generation: impossibility and compromise. J. Univ. Comput. Sci. 12:672–690.

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

Novo A.A. 2009 based on Schafer, 1997. norm: analysis of multivariate normal datasets with missing values. Version 1.0-9.5.

Peres-Neto P.R., Jackson D.A. 2001. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. Oecologia 129:169–178.

Peres-Neto P.R., Jackson D.A., Somers K.M. 2003. Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. Ecology 84:2347–2363.

R Development Core Team 2005. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rhode M.P., Arriaza B.T. 2006. Influence of cranial deformation on facial morphology among prehistoric south central Andean populations. Am. J. Phys. Anthropol. 130:462–470.

Rohlf F.J. 2003. Bias and error in estimates of mean shape in geometric morphometrics. J. Hum. Evol. 44:665–683.

Rubin D.B. 1987. Multiple imputation for nonresponse in surveys. New York: J. Wiley & Sons.

Rubin D.B., Schenker N. 1991. Multiple imputation in health-care databases: an overview and some applications. Stat. Med. 10:585–598.

Schafer J.L. 1997. Analysis of incomplete multivariate data. New York: Chapman & Hall.

Schafer J.L., Olsen M.K. 1998. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. Multivariate Behav. Res. 33:545–571.

Schneider J., Borlund P. 2007. Matrix comparison, Part 2: Measuring the resemblance between proximity measures or ordination results by use of the mantel and procrustes statistics. J. Am. Soc. Inf. Sci. Tec. 58:1596–1609.

Siegel A.F., Benson R.H. 1982. A robust comparison of biological shapes. Biometrics 38:341–350.

Soligo C., Andrews P. 2005. Taphonomic bias, taxonomic bias and historical non-equivalence of faunal structure in early hominin localities. J. Hum. Evol. 49:206–229.

Spratt M., Carpenter J., Sterne J.A.C., Carlin J.B., Heron J., Henderson J., Tilling K. 2010. Strategies for multiple imputation in longitudinal studies. Am. J. Epidemiol. 172:478–487.

Stacklies W., Redestig H., Scholz M., Walther D., Selbig J. 2007. pcaMethods – a Bioconductor package providing PCA methods for incomplete data. Bioinformatics 23:1164–1167.

Strauss R.E., Atanassov M.N. 2006. Determining best complete subsets of specimens and characters for multivariate morphometric studies in the presence of large amounts of missing data. Biol. J. Linn. Soc. 88:309–328.

Strauss R.E., Atanassov M.N., De Oliveira J.A. 2003. Evaluation of the principal-component and expectation-maximization methods for estimating missing data in morphometric studies. J. Vert. Paleontol. 23:284–296.

Voje K.L., Hansen T.F. 2013. Evolution of static allometries: adaptive change in allometric slope of eye span in stalk-eyed flies. Evolution 67:453–467.