Syst. Biol. 61(6):941–954, 2012

DOI:10.1093/sysbio/sys047
Advance Access publication on April 17, 2012

Testing of the Effect of Missing Data Estimation and Distribution in Morphometric Multivariate Data Analyses

CALEB MARSHALL BROWN[1,2,*], JESSICA H. ARBOUR[1], AND DONALD A. JACKSON[1]

[1] *Department of Ecology and Evolutionary Biology, University of Toronto, 25 Willcocks Street, Toronto, Ontario M5S 3B2, Canada; and* [2] *Department of Natural History - Palaeobiology, Royal Ontario Museum, 100 Queen's Park, Toronto, Ontario M5S 2C6, Canada;*
[*] *Correspondence to be sent to: Department of Natural History - Palaeobiology, Royal Ontario Museum, 100 Queen's Park, Toronto, Ontario M5S 2C6, Canada;*
E-mail: caleb.brown@utoronto.ca

Received 3 September 2011; reviews returned 6 January 2012; accepted 9 April 2012
Associate Editor: Norman MacLeod

Abstract.—Missing data are an unavoidable problem in biological data sets and the performance of missing data deletion and estimation techniques in morphometric data sets is poorly understood. Here, a novel method is used to measure the introduced error of multiple techniques on a representative sample. A large sample of extant crocodilian skulls was measured and analyzed with principal component analysis (PCA). Twenty-three different proportions of missing data were introduced into the data set, estimated, analyzed, and compared with the original result using *Procrustes* superimposition. Previous work investigating the effects of missing data input missing values randomly, a non-biological phenomenon. Here, missing data were introduced into the data set using three methodologies: purely at random, as a function of the Euclidean distance between respective measurements (simulating anatomical regions), and as a function of the portion of the sample occupied by each taxon (simulating unequal missing data in rare taxa). Gower's distance was found to be the best performing non-estimation method, and Bayesian PCA the best performing estimation method. Specimens of the taxa with small sample sizes and those most morphologically disparate had the highest estimation error. Distribution of missing data had a significant effect on the estimation error for almost all methods and proportions. Taxonomically biased missing data tended to show similar trends to random, but with higher error rates. Anatomically biased missing data showed a much greater deviation from random than the taxonomic bias, and with magnitudes dependent on the estimation method. [Crocodilia; deformation; fossil; incomplete; morphology; ordination; PCA; *Procrustes*; shape; taxonomy.]

Multivariate morphometric data analyses are becoming increasingly common in biological studies as a method of analyzing biological shape (e.g., Dodson 1975a, 1975b; Pimentel 1979; Bookstein 1985; Strauss 1985; Rohlf 1990; Zelditch et al. 2003, 2004). Within biological data sets, missing values are a common and systemic problem (Hadfield 2008). This is highlighted by the fact that many methods to analyze multivariate morphology require complete data sets, and cannot function with missing data (Strauss et al. 2003). Issues of missing data in very large data sets can often be addressed by deleting the incomplete observations or variables (list-wise deletion) from the data set (Beale and Little 1975; Rubin 1976; Legendre and Legendre 1998). This approach has the unfortunate result of decreasing the sample size and, as a result, the statistical power of the analysis (Nakagawa and Freckleton 2008). This approach is also not always possible or practical with morphometric data sets, especially those with already small sample sizes. This situation often leaves morphologists in the dilemma of either estimating missing data based on limited present data, or working around missing observation/variable pairs. Several approaches to estimating missing values from partial data sets exist, from simple approaches such as using the mean of the known observations to complex multivariate methods (Beale and Little 1975; Rubin 1976; Legendre and Legendre 1998; Strauss et al. 2003; Nakagawa and Freckleton 2008). In addition to these, multiple procedures exist that down-weight or ignore missing data occurrences (Gower 1966, 1971; Reyment 1991; Nakagawa and Freckleton 2008).

Despite these limitations, multivariate explorative data analyses are widely used in biological analyses, often without a full discussion of the effects of missing data or estimated data (but see Dodson 1975b, 1979, 1990; Hammer and Harper 2006). Additionally, the relative performance of different estimation methods have rarely been discussed, let alone tested. As a result their estimation accuracies as a function of the proportion of missing data and data set size are unknown.

On those few occasions when the effect of missing data in morphological analyses have been tested, the tests have been undertaken with missing data input randomly into the data set (Kramer and Konigsberg 1999; Strauss et al. 2003). This approach almost certainly stems from the simplicity of randomly assigning missing data within a data set for multiple replicates, compared with the relative complexity of inputting missing values with a specific bias in a quantitative and repeatable fashion. The distribution of missing values within morphological data sets is not random, however, and is a result of multiple biological and sampling factors (Hadfield 2008; Nakagawa and Freckleton 2008).

The most frequent cause of missing data in morphological data sets is the presence of partially incomplete, crushed, or distorted specimens in the sample. Incomplete or damaged specimens do not affect all measurements in that specimen, only those measurements that relate to anatomical region of the specimen that is missing or damaged. Additionally, not all regions have the same probability of being damaged or lost. Extremities, delicate regions, and elements subject to preferential taphonomic loss will have higher occurrence of missing data than central or robust regions. Missing data within a specimen are, therefore, not randomly distributed, but, rather, is often distributed in clusters based on the relative proximity of the measurements to each other, and the likelihood of those regions to be missing. It is unclear how missing data distributed regionally within specimens affects the analyses, or whether the error of estimation differs from the pattern seen in randomly distributed missing data.

In addition to the nonrandom distribution of missing data within a specimen, there is also a nonrandom distribution of missing data among specimens. The relative sample size of each taxon within a morphometric data set is often negatively correlated with the relative amount of missing data within specimens of that taxon, with specimen completeness proportional to the sample size for that taxon. This tendency results from a principal factor. When a sample of one species is represented by abundant specimens, the researcher can often be selective about which specimens to measure and can restrict inclusion to only complete specimens. Conversely, when a taxon is represented by a small number of specimens, there will be a desire to use all of the available specimens regardless of their completeness to increase sample size and ensure adequate taxonomic sampling. The lack of independence between the number of specimens for each taxon represented in the data set, and the relative amount of missing data in that sample is problematic when we consider that it is often these rare/incomplete taxa that are the primary subject(s) of interest in morphometric studies. Estimation of values from randomly input missing data has a higher degree of error for taxa with low sample sizes relative to the whole sample, because smaller samples reduce the accuracy of estimated values for missing data values. Therefore, it is necessary to test whether higher rates of missing data in taxa represented by low specimen numbers compound the problem seen in estimating randomly assigned missing data. This condition also represents the worst-case scenario in terms of relative distribution of missing data between taxa within the data set.

We herein capitalize on an existing and complete morphometric data set to test the effect that different proportions of missing data have on our ability to estimate the missing values. This approach allows for 1) a direct comparison of the accuracy of different methods of handling missing data; 2) a comparison of the relative accuracy of these methods across varying proportions of missing data; and 3) a direct comparison of the effect that the distribution of missing data (random, anatomic, and taxonomic) have given both the different proportions of missing data and the estimation method.

Institutional Abbreviations

CMN = Canadian Museum of Nature, Ottawa; FMNH = Field Museum of Natural History, Chicago; ROM = Royal Ontario Museum, Toronto; RTMP = Royal Tyrrell Museum of Palaeontology, Drumheller; UCMP = University of California Museum of Paleontology, Berkeley; UCMVZ = University of California Museum of Vertebrate Zoology, Berkeley; UCMZ = University of Calgary Museum of Zoology, Calgary; UM = University of Michigan Museum of Zoology, Ann Arbor.

Materials

This study uses a data set of crocodilian crania, from 226 specimens housed in eight museum collections. Crocodilians were chosen because they are relatively well-resolved, both taxonomically and phylogenetically, span a large range in body sizes to allow for testing the effect of body size and allometry, and large osteological collections exist in natural history museums to allow for ease of data collection (Romer 1956; Norell 1989; Densmore and White 1991; Harshman et al. 2003).

This data set contains 226 specimens representing 21 of the 23 extant taxa (online Appendix 1 at http://dx.doi.org/10.5061/dryad.m01st7p0). All specimens reside within four major groupings, two at the family level (Gavialidae and Crocodylidae) and two at the subfamily level within the family Alligatoridae (Alligatorinae and Caimaninae) (Romer 1956; Norell 1989; Densmore and White 1991; Poe 1996; Harshman et al. 2003). With a single exception, all skulls represent recent specimens, which are assigned taxonomically based on a combination of soft and hard tissue anatomy, and biogeography. One skull was measured from a fossil specimen—ROM 51011 from the Late Pleistocene (30 000 years old) of Florida, which is assigned to the extant genus *Alligator mississippiensis*.

The measured sample includes specimens from hatchling-sized (or near hatchling) individuals to large, old adults in all of the four major taxonomic groups, with nearly compete size series seen in many species, particularly *Alligator mississippiensis*. As a result, the data set shows a remarkable range of skull sizes. The skull length of the largest specimen (CMNAR 15792, *Crocodylus porosus*—749 mm) is >25 times larger (in linear dimensions) than that of the smallest (ROM R 7966, *Alligator mississippiensis*—29mm). As such, it represents as great a scaling problem as will likely be encountered in any morphometric analyses.

Variables

Twenty-three cranial measurements were taken from each skull (See Fig. 1 and online Appendix 1 at http://dx.doi.org/10.5061/dryad.m01st7p0). The measurements taken follow those of Dodson (1975a), with the measurements of the super temporal fossa omitted (due to poor preservation and observation of this feature in many specimens). These osteological measurements represent functional complexes as opposed to dimensions of individual bones, and have biomechanical and behavioral correlates (see Dodson

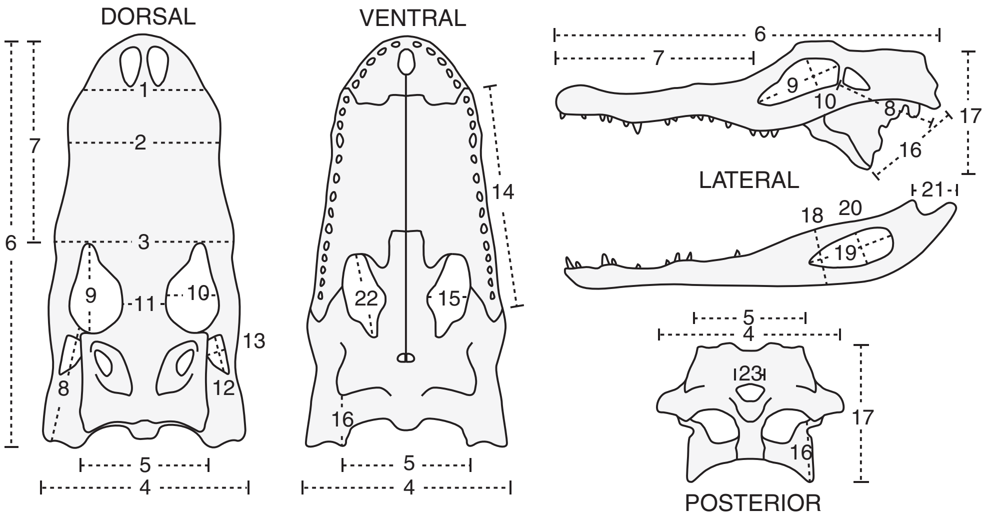

FIGURE 1. The 23 linear morphometric measurements used in this study illustrated on the skull of *Alligator mississippiensis*. For list of measurement see Online Appendix 1. For biological explanation of measurements see Dodson (1975a). Figure modified from Dodson (1975a).

1975a). Measurements were taken from the left side, unless this side was either incomplete or damaged, in which case the right side was used. Measurements of magnitude <150 mm were taken with digital calipers, between 150 and 300 mm with dial calipers, and >300 mm were taken using a fiberglass measuring tape. All measurements were taken to the nearest millimeter.

METHODS

All exploratory data analyses were carried out using the R software package (R Development Core Team 2011), using the packages "MASS" (Venables and Ripley 2002), "vegan" (Oksanen et al. 2010), "e1071" (Dimitriadou et al. 2010), "pcaMethods" (Stacklies et al. 2010), "psych" (Revelle 2010), "cluster" (Maechler et al. 2010), and "LOST" (cited herein) available on the CRAN (http://cran.r-project.org/) and/or Bioconductor (http://www.bioconductor.org/) websites. Figures and plots were created using the R software package (R Development Core Team 2011), Microsoft Excel (v 12.3.0), and Adobe Illustrator (v 15.1.0).

Missing Data Input

The complete data set of crocodilian measurements was subjected to artificial removal of data values to simulate the incompleteness of the fossil record, with differing proportions and configurations of missing data. The proportion of missing data is represented as a percentage of the total. Twenty-three different proportions of missing data—1% though 5% in intervals of 1, and 5% to 95% in intervals of 5—were tested. Although we do systematically test the error introduced by estimation of very high portions of missing data, this is done as a test of the performance of these methods, and should not be viewed as a realistic amount of missing data.

Three functions were developed in the R statistical language to introduce a specified number of missing values into the data set by substituting existing data values with "NA" following three criteria; random, anatomical bias, and taxonomic bias.

Random.—Data values were removed from the data set purely at random (henceforth the random missing data function or RMD), following the "missing completely at random" or MCAR model of Little (1988). "Random" data were generated by selecting values from the data matrix using the "sample" function from the "base" package in R. "Sample" uses a fast pseudorandom number generator called the "Mersenne-Twister" (Matsumoto and Nishimura 1998), which is commonly applied to Monte Carlo simulations because of its long period and equidistribution property (Matsumoto et al. 2006; Panneton et al. 2006). Although this random model does not represent the observed distribution of missing data in most morphometric data sets, it is the most basic simulation and, in most cases, would represent a best-case scenario.

Anatomic bias.—To replicate the pattern of regional missing data caused by incomplete or damaged specimens, the probability of a measurement being removed from a specimen was calculated as being proportional to the inverse of the minimum Euclidean distances of its landmarks to the landmarks of all other measurements previously removed from that specimen (online Appendix 2, Function "obliterator"). A single reference skull was used to establish the coordinates of each measurement landmark in three dimensions, and these were used to calculate the minimum absolute distances between measurements. While individual skull shape variation may have influenced the minimum distances between landmarks, it was not possible to generate a set of coordinates for each skull and such variations would not have significantly altered the effect of anatomical bias from that produced using the "obliterator" function.

To preserve the distribution of the number of missing data values across specimens (i.e., the number of specimens missing data and the number of measurements missing from each of those specimens), the function RMD was used to establish how many (and which) specimens would have measurements removed and how many measurements from each of those specimens would be removed. The first measurement removed from each "incomplete" skull (as well as which landmark of that measurement) was chosen at random. For subsequent removals, the minimum distance between each previously removed landmark and each remaining measurement's landmarks were calculated. The probability of measurement A being removed after measurement B had been removed (P(A|B)) was proportional to the inverse of the minimum distance between the landmarks of measurements A and B, divided by the sum of inverses of the minimum distances of all other measurement landmarks to point B [Equation (1)]. Therefore, the likelihood of a measurement being removed was greater for those positioned close to missing measurement's landmarks than those positioned far away; however, the strength of this effect lessened with distance from the missing measurement landmark (i.e., measurements very far from a missing structure had similar, low likelihoods of being removed). For a given specimen, the probability of a measurement being removed was proportional to the product of the probability of that measurement being removed given each of the measurements already removed [Equation (2)]. This process was iterated for each specimen, after the first data point, until reaching the number of data values to be removed for a given specimen.

$$P(i|j)=\frac{1/d_{i,j}}{\sum_k^n 1/d_{k,j}} \tag{1}$$

$$P(i)=\frac{P(i|1)*P(i|2)*P(i|3)*\ldots*P(i|m)}{\sum_k^n P(k|1)*P(k|2)*P(k|3)*\ldots*P(k|m)} \tag{2}$$

where: $d=$ the minimum distance between a measurement (i or k) and a missing measurement (j); $i=$ a particular measurement; $j=$ a missing measurement; $k=$ any non-missing measurement; $n=$ the number of measurements not missing; $m=$ the number of measurements already missing.

In the special case where measurements shared a terminus and the distance between the two would have been zero, the distance was instead set as one half of the minimum distance from the missing measurement to the next closest measurement. In reality, if one such measurement was missing, both measurements would be missing; however, this situation would dramatically change the distribution of missing data values across specimens compared with RMD. By using a distance less than the minimum distance to the next nearest point, the overlapping measurement becomes the most likely to be removed next. This function "obliterator" is available on the CRAN website under the package "LOST".

Taxonomic bias.—To replicate the pattern of taxonomic bias (i.e., where taxa represented by only a few specimens tend to exhibit higher amounts of missing data), a function was developed (online Appendix 2, Function "by.clade") that weighted the probability of having the largest number of measurements removed from specimens belonging to taxa represented by few specimens. A sample matrix containing missing data was created using the RMD function. From this matrix, the number of specimens missing data was calculated. A vector was produced containing the number of measurements missing from each incomplete specimen, and this vector (**V**) was sorted into descending order.

Specimens were sampled without replacement, with a probability for each specimen relative to the sum of the entire sample sizes divided by the number of specimens in that respective specimen's taxon [Equation (3)]. Measurements were removed (randomly) from each sampled specimen based on the order in which they were sampled and corresponding number of measurements in **V** (i.e., the first specimen sampled had the largest number of measurements removed, the second specimen sampled had the second largest number of measurements removed, etc.). This meant that specimens being sampled from the least well-sampled taxon possessed the highest probability of being sampled first and thereby having the largest number of measurements removed. In this case, species was the taxonomic group on which this bias was based. This function "by.clade" is available on the CRAN website under the package "LOST".

$$P(i)=\frac{\sum(N_1+N_2+N_3\cdots+N_n)/N_i}{\sum_i^j\sum(N_1+N_2+N_3\cdots+N_n)/N_i} \tag{3}$$

where: $i=$ a given specimen; $j=$ the total number of specimens; $N=$ the sample size of a taxon; $n=$ the total number of taxa; $N_i=$ the sample size for the taxon to which specimen i belongs.

Missing Data Analysis

Diminished data sets were analyzed using non-estimation methods (Pairwise deletion and Gower's distance matrix), or missing data were estimated using multiple procedures [substitution of mean, *a priori* size regression, correlated variable regression, and Bayesian principal component analysis (PCA) missing value estimator].

Pairwise deletion.—Pairwise deletion is a common, often-recommended approach for dealing with missing data (Dodson 1975a), as no estimated values are input into the data set. In this method, a distance matrix is produced base on pairwise dissimilarity of observations. Observations missing from either variable used to calculate bivariate statistics are excluded from both variables. Here pairwise deletion was performed using the "principal" function in the package "psych" v. 1.0-92 (Revelle 2010).

Gower's distance matrix.—Gower's distance matrix handles missing data simply by weighting it as zero in the analysis, resulting in all the weight being associated with the present values (Gower 1966, 1971; Reyment 1991). Rather than estimating the missing data and returning a complete matrix, this method produces a distance matrix based on pairwise similarity (distance) between observations with missing values weighted as zero. Here a Gower's distance matrix was produced using the function "daisy" in the package "cluster" v. 1.32.3 (Maechler 2011).

Substitution of mean.—Mean substitution is a common estimation approach because the component axes of the ordination are centered on the grand mean and, as a result, substitution of the variable mean does not influence the axes. The mean of the observations for each variable was calculated from the existing data set and substituted for all missing values for that variable. In our analysis, this was performed using the function "impute" in the package "e1071" 1.5-25 (Dimitriadau et al. 2011).

A priori *size regression.*—When the sample is represented by a range of sizes, and variables are highly correlated with size, estimation of missing values may be based on their allometric equations relative to an *a priori* size variable (See discussion in Dodson 1975a; Little 1992; Strauss et al. 2003). A function was developed (online Appendix 2, Function "est.reg") to replicate the univariate estimation of missing data values based on a regression of all variables against an *a priori* determined variable that is a proxy for size. All variables (with the exception of Variable 6—skull length) were regressed independently against skull length (Variable 6). Missing values for each variable were then estimated using the allometry of that variable relative to skull length. Missing values for Variable 6 (skull length) were estimated using the regression line of the most correlated variable for that specimen. This function "est.reg" is available on the CRAN website under the package "LOST".

Correlated variable regression.—Rather than estimation based on the allometry of each variable against an *a priori* size variable, this method used the variable most highly correlated with the variable experiencing missing data. A second function was developed (online Appendix 2, Function "best.reg") to estimate missing data values based on a regression of each variable against its most correlated variable, and not an *a priori* determined size variable. All variables were regressed individually against all other variables. The regression of the variable with the highest correlation (the highest R^2 value) with that of the variable experiencing missing data was used to estimate the missing data values. This function "best.reg" is available on the CRAN website under the package "LOST".

Bayesian PCA missing value estimator.—This method combines the expectation maximization approach for PCA (Strauss et al. 2003) with a Bayesian model (Oba et al. 2003) and has been used for missing data estimation in morphometrics (Campione and Evans 2011). It is a highly complex method requiring multiple iterative matrix inversions and attempts to replace the missing value based via PCA by regressing the remaining values against the principal components for complete values (Strauss et al. 2003). This method was performed using the function "bpca" in the package "pcaMethods" (Stacklies et al. 2010).

Multivariate Explorative Data Analysis

PCAs were performed on the original unaltered crocodile data set, as well as independently derived replicates of each of the 23 proportions of missing data for all of the missing data handling techniques. Several methods were unable to effectively and consistently perform at relatively high proportions of missing data (with distribution of missing data values sometimes being a confounding problem), and as a result few replicates could be performed. Mean substitution was robust even at small sample size so 1000 replicates were performed for all proportions of missing data for random and taxonomic distribution. For the anatomic distribution of missing data, mean substitution is represented by 1000 replicates for proportions of 65% or lower, 300 replicates for 70%, 100 replicates for 75%, and 30 replicates for 80%. Gower's coefficient is represented by 1000 replicates for proportions <40%, 100 replicates at 40%, and 10 replicates at 45% for all distributions. Bayesian PCA (BPCA) is represented by 1000 replicates for proportions of 65% or lower, 300 for 70%, 100 for 75%, and 30 for 80%, for all distributions. Pairwise deletion is represented by 300 replicates for proportions <10% (all distributions), 200 replicates for 10–40%, and 50 replicates for 45–75% (for random), and 100 replicates for 10–35% and 20 replicates for 40–80% and 40–65 or 85% (for taxonomic and anatomic biases). The *a priori* size regression is represented by 1000 replicates for 1–60% (taxonomic and anatomic) and 1–65% (random), 200 replicates for 65–70% (taxonomic) and 70% (random), 50 replicates for 75–80% (taxonomic and random), and 30 replicates for 65% (anatomic). The correlated variable regression is represented by 1000 replicates for 1–60%, 300 replicates for 65–70%, 100 replicates for 75%, and 20 replicates for 80% for both taxonomic and random distributions, with anatomic distribution represented by 1000 replicates for 1–55% and 100 replicates for 60%.

The original matrix and the estimated matrices were then analyzed using PCA. Quantitative comparisons between the original and estimated PCA results were performed using *Procrustes* superimposition to compare the relative position of specimens in multivariate PC space (Gower 1975; Kendall 1989; Bookstein 1997). The overall set of differences (lack-of-fit) between the original and estimated scores, *Procrustes* sum of squares error, is due to error of estimation of data, and as such, the *Procrustes* sum of squares error is a metric of estimation accuracy of relative specimen positions in the multivariate ordination. Perfect matching in *Procrustes* results in zero sum of squares error, and a complete mismatch results in a sum of squares error of one. As such, *Procrustes* sum of squares error provides a scale on which performance of missing data estimation methods can be tested and compared in terms of the between-specimen distance in the ordination. It should be noted that this method only compares the change in the relative position of specimens in multivariate PC space, often the most important property for systematists, but does not consider other PCA properties such as change in loadings or variance in the vectors.

The sum of squares error for each replicate, as well as the means and standard deviations (SDs) of the entire replicate series were recorded for all amounts of missing data. The means and SDs were then plotted against the relative proportion of missing data. *Procrustes* superimposition was chosen for comparison as it uniquely allows for quantitative comparison of the effect missing data estimation has on the result of the PCA, not the effect the estimation has on changing the data set itself. Previous studies investigating the effect of missing data estimation on morphometric analyses have concentrated on comparing the original and estimated data sets, not the result (Strauss et al. 2003). It is the result that is interpreted, and estimations that cause error here will have real and meaningful effects on the interpretation of the findings. Additionally, this approach allows for the direct comparison of both estimation and non-estimation methods, a comparison not possible when analyzing the change in the data set itself.

Results

Relative Performance

Figure 2 illustrates the estimation error (*Procrustes* sum of squares) incurred by the various estimation methods across all values of missing data when the missing data are A, randomly; B, taxonomically; and C, anatomically distributed. For all proportions of missing data (and among all missing data distributions) mean substitution introduced the highest amount of missing data estimation error (Fig. 2). This method showed a linear increase in error with the proportion of missing data (Table 1). Although performing poorly in terms of estimating relative sample distances in the ordination, mean substitution is likely the estimation method with the least effect on the orientation of the PC axes.

Pairwise deletion of missing data resulted in less error than mean substitution, but more than the remaining methods and showed a non-linear relationship between error and increasing amount of missing data (Fig. 2). This relationship is best described by a power function (Table 1), which shows an increase in the error rate with a linear increase in proportion of missing data. These trends, however, are based on a smaller number of replicates than those of the other methods. The relative performance of this method lies between mean substitution and the other methods tested.

The *a priori* size regression estimation method shows a linear relationship with an increasing proportion of missing data from 1% to 75% with RMD, but a power function when the data have taxonomic or anatomic biases (Fig. 2 and Table 1). However, in >75% of taxonomic and random and 55% of anatomic biased data, both the error rate and variation increase disproportionally. This increase in the mean error and variance resulted from a small number of results with magnitudes of up to an order of magnitude greater than those of the mean. Between 1% and 75% of missing data, the relative performance of this function lies between that of BPCA and pairwise deletion.

Unsurprisingly, the correlated variable regression estimation method outperformed the *a priori* size regression method, but shows generally similar trends (Fig. 2). Unlike the *a priori* size method, the correlated variable shows a non-linear relationship with increasing missing data, best described as a power function, for all distributions of missing data (Table 1). As with the *a priori* size method, however, at >75% (and 55% for anatomic missing data) both the error rate and variation increased disproportionally. For random and taxonomic missing data, this method resulted in more estimation error than BPCA <35% and less >35%. When the missing data were distributed with an anatomic bias, this method consistently introduced more error than BCPA.

When the missing data were randomly or taxonomically distributed and in small proportions (<35%), BPCA introduced the lowest amount of error of any of the estimation methods (Fig. 2a,b). Above 35%, this method was outperformed by the correlated variable regression. When the missing data were anatomically distributed, however, BPCA consistently introduced less error than all other estimation methods including correlated variable regression (Fig. 2c), and all non-estimation methods, including Gower's distance matrix >15%. The increase in error of estimation with this BPCA is not linear to the increase in proportion of missing data, but rather is represented by a power or exponential function (Table 1), which indicates a disproportional increase in error as missing data increases. The most dramatic increases in error rate and variation occurred at 35% and 75% for random and

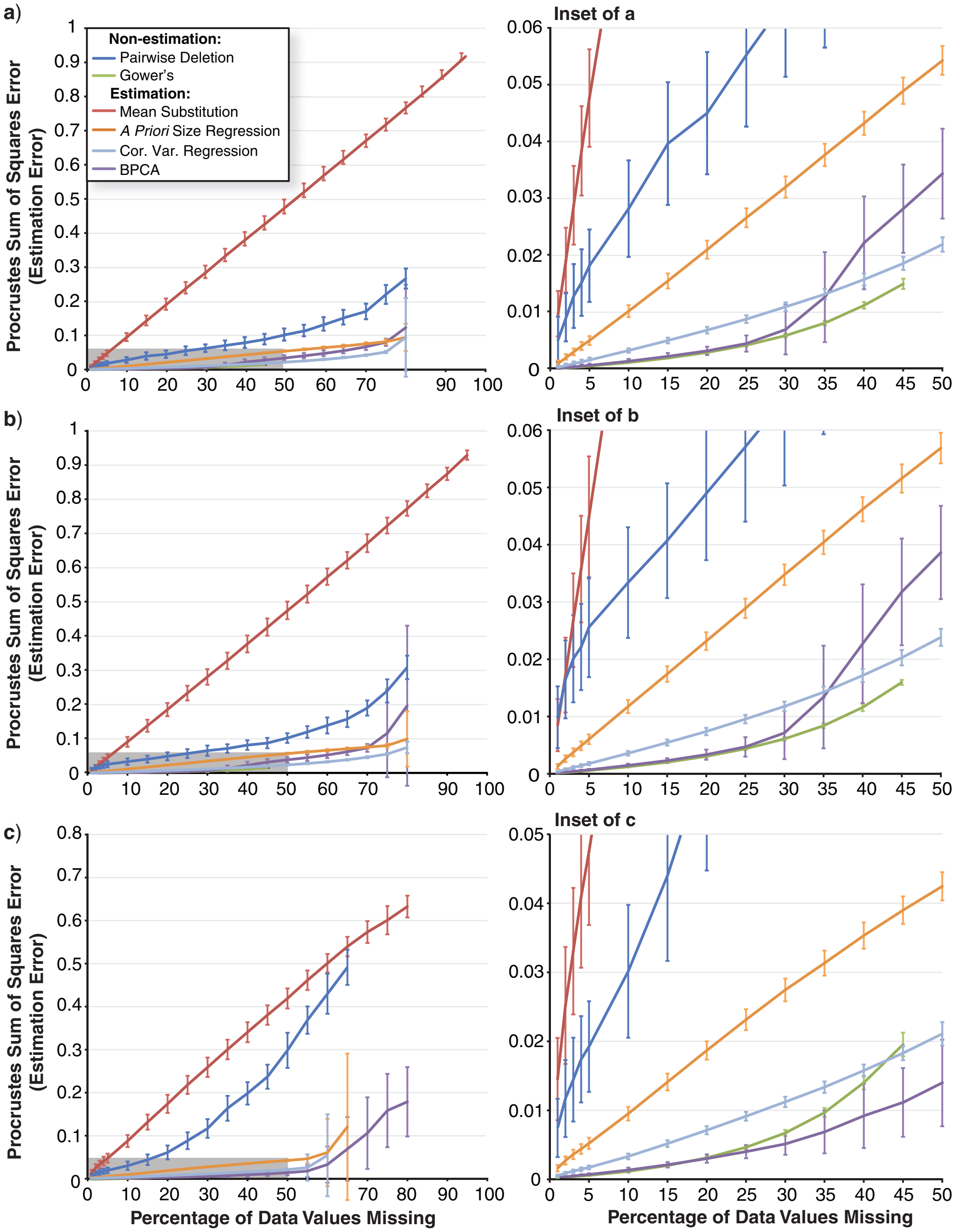

FIGURE 2. Estimation errors obtained by the methods as a function of the proportion of missing data in the data set introduced a) randomly, b) with a taxonomic bias, and c) with an anatomic bias. Points represent the replicate mean with error bars of ±1 SD. In all cases the horizontal axis is the percentage of missing data and the vertical axis is *Procrustes* sum of squares error.

taxonomic missing data, and 55% for anatomic missing data (Fig. 2).

Gower's distance matrix introduced the least error of all methods tested when the missing data were either randomly or taxonomically distributed (Fig. 2a,b). Anatomically biased missing data, however, caused Gower's method to introduce the least error when the percentage was <15%, more error than BPCA when >15%, and more error than the correlated variable regression and BPCA >40% (Fig. 2c). Although it performs well, Gower's coefficient is limited in terms of the amount of missing data it is capable of handling.

TABLE 1. A table summarizing the best fit lines for the relationship between percent missing data and estimation error for each method and distribution of missing data

Estimation method	Distribution	Mode	Equation	R^2	Sample
Pairwise deletion	Random	Power	$0.0047x^{0.8163}$	0.97402	1–80%
	Taxonomic	Exponential	$0.0192e^{0.0344x}$	0.94863	1–80%
	Anatomic	Exponential	$0.0138e^{0.0615x}$	0.95776	1–65%
Gower	Random	Power	$8E\text{-}05x^{1.2838}$	0.98364	1–45%
	Taxonomic	Power	$9E\text{-}05x^{1.2663}$	0.98205	1–45%
	Anatomic	Exponential	$0.0003e^{0.0999x}$	0.96455	1–40%
Mean substitution	Random	Linear	$0.0096x$	0.99997	1–95%
	Taxonomic	Linear	$0.0096x$	0.99955	1–95%
	Anatomic	Linear	$0.0082x$	0.99739	1–65%
A priori size regression	Random	Linear	$0.0011x$	0.99980	1–75%
	Taxonomic	Power	$0.0013x^{0.96}$	0.99982	1–75%
	Anatomic	Power	$0.0014x^{0.8603}$	0.99742	1–55%
Correlated variable regression	Random	Power	$0.0003x^{1.1401}$	0.99073	1–75%
	Taxonomic	Power	$0.0003x^{1.1202}$	0.98891	1–75%
	Anatomic	Power	$0.0004x^{0.9755}$	0.99073	1–55%
BPCA	Random	Power	$6E\text{-}05x^{1.5721}$	0.95518	1–80%
	Taxonomic	Exponential	$0.0004^{0.0765x}$	0.94852	1–80%
	Anatomic	Exponential	$0.0005^{0.0803x}$	0.94852	1–80%

At values greater than ~30% missing, viable results of the analysis are inconsistent, and at values >45% nearly all data sets were unable to be estimated. This method showed a non-linear relationship between *Procrustes* sum of squares error and increasing amount of missing data (Fig. 2). It is best described by a power function (Table 1), which shows an increase in the rate of error with a linear increase in proportion of missing data.

Effect of the Distribution of Missing Data

Missing data estimation did not affect all specimens equally, even when the missing data were randomly distributed throughout the data set over multiple iterations. Specimens showing the largest movement (those with the largest introduced error) were those that were taxonomically underrepresented in the sample (here *Gavialis*, *Osteolaemus*, etc.) and those that lay on the periphery of the occupied morphospace (here *Gavialis*). This contrasts with the relative little movement that occurred in the abundantly sampled and tightly associated *Alligator* cluster. Additionally, within taxonomic samples the outlying specimens were brought closer to the main cluster, and the tight cluster of central specimens spread out when the missing data were estimated.

The experiments contrasting the effect of both taxonomically and anatomically biased data with randomly distributed missing taxa allow for a quantitative test of the effect of random and non-randomly distributed missing data. The results of the introduced error for the various methods of estimation, at both different proportions and distributions of missing data, are summarized in Figure 3.

Taxonomic bias.—Pairwise deletion of taxonomically biased missing data resulted in higher error than for RMD across all proportions of missing data. These overall trends are similar and the absolute differences between the two methods are small (Fig. 3a).

Although the overall trends were similar (Fig. 3b), Gower's distance matrix showed higher error in taxonomic biased distributions than random distributions for all proportions. Similarity the trends between random and taxonomic can be seen by comparing the equations describing their slope (Table 1).

For mean substitution, when the proportion of missing data was <70%, missing data with a bias toward rare taxa systematically resulted in a lower rate of estimation error than missing data introduced at random (Fig. 3c). There is also a gradual decrease in the relative difference in error between the two distributions as the amount of missing data increases.

The regression estimation methods, both *a priori* size and correlated variable, showed similar response to the distribution of missing data. In both cases, taxonomically biased missing data result in greater error than is encountered when the missing data are randomly distributed (Fig. 3d,e). Both methods show a sharp increase in the error at 80% of missing data, matching the spike at the same proportion in the random sample (Fig. 3d,e). As with the random data sets, this increase in the mean error and variance seen at 80% resulted from a small number of results with magnitudes of up to an order of magnitude greater than those of the mean.

Taxonomic missing data estimated using BPCA showed a more complex relationship (Fig. 3f). As with the other methods, the patterns between random and the taxonomic biased samples were closer than either was to the regional bias, and taxonomic missing data incurred more estimation error than random across all proportions of missing data. Additionally, for reasonable amounts of missing data (<70%) the absolute differences between these two methods was quite small compared

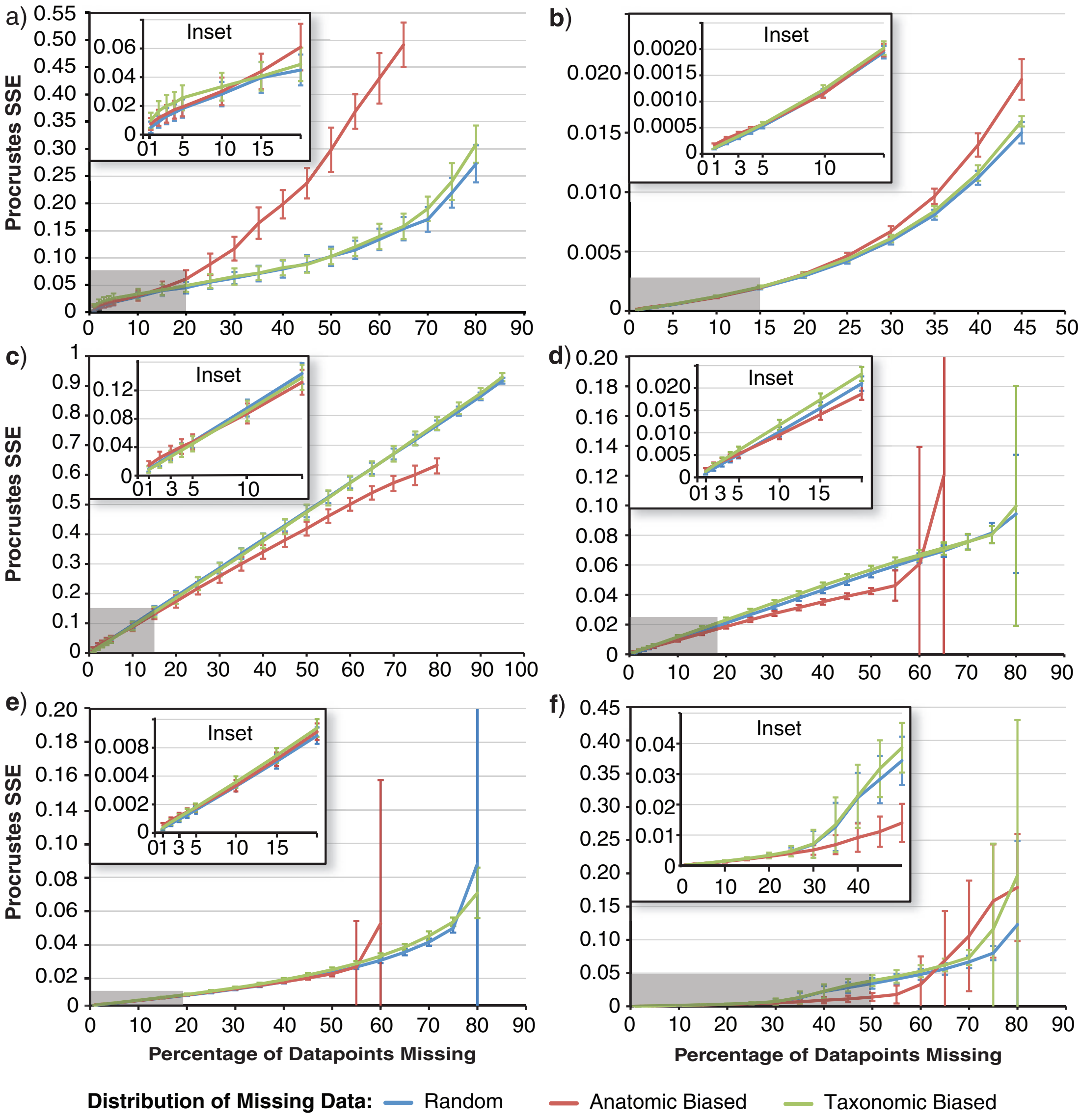

FIGURE 3. Estimation error introduced by a) pairwise deletion, b) Gower's distance matrix, c) mean substitution, d) *a priori* size regression, e) correlated variable regression, and f) BPCA as a function of both proportion of missing data and distribution of that missing data. Points represent the replicate mean with error bars of ±1SD. In all cases the horizontal axis is the percentage of missing data and the vertical axis is *Procrustes* sum of squares error.

with the total error. For proportions of missing data >70%, the rate of error for the taxonomically biased sample increased dramatically more than the same proportions in the random sample, resulting in a large difference in error values (Fig. 3f and Table 1).

Although the estimation errors derived from taxonomically biased missing data were almost always higher than derived from RMD, the overall trends were similar. In almost all cases, the distribution of the missing data, whether random or taxonomically biased, had less of an effect on the resultant estimation error than either the estimation method or the proportion of missing data.

Anatomic bias.—Pairwise deletion of anatomically biased missing data resulted in a very large increase in error compared with the randomly introduced missing data (Fig. 3a). All error means obtained, with the exception of that at 10%, were greater than those obtained with random distribution.

As is the case with the majority of the methods, Gower's distance matrix showed a stronger deviation from the random distribution with anatomically biased missing data than with taxonomically biased (Fig. 3b). The obtained result for the anatomic biased data showed higher error for all proportions with the exception of that for 10%.

Estimation error from mean substitution of missing data introduced to mimic regional loss of data is shown in Figure 3c. For small amounts of missing data (<10%), the error rate of random distributions was lower than that of the regional bias, with the relative difference between the two increasing with lower amounts of missing data. For portions of missing data >10%, however, the regional bias systematically resulted in a lower rate of estimation error than missing data introduced at random (increasing from around 5% to 20%). The differences in the error rate were much greater than those seen between random and the taxonomically biased samples. In addition to having lower amounts of error at given proportions of missing data (>10%), the rate of error increase with increasing proportion of missing data was less than those of both random and taxonomically biased missing data.

As with the taxonomically biased missing data, the two regression estimation methods again showed similar results to each other. For proportions of <50% anatomic biases showed a steady increase in error. Above 50% the error rate increased greatly, matching the pattern seen at 80% for both the taxonomic and random samples (Fig. 3d,e). For the *a priori* size method, between 5% and 55% missing data, the error for the anatomic samples was less than those findings seen for the random samples. For the correlated variable method, however, all proportions, except for 45%, had a higher error rate for the anatomically biased sample (Fig. 3d,e). The dramatic increase in the mean error and variance seen >50% for both the *a priori* size and correlated variable method was a result of a small number of results having differences up to an order of magnitude greater than that of the mean.

For BPCA, proportions of missing data between 10% and 60% resulted in anatomically biased samples experiencing less estimation error than those of a random distribution. Below 10%, although the difference is very small, results from regional data showed more error than random. Above 65% the anatomic error rate increased dramatically to exceed both those of the random and of the taxonomic sample (Fig. 3f).

With the possible exception of the correlated variable regression method, the estimation errors resulting from anatomically biased missing data were consistently more disparate from those obtained at random than are the taxonomic results. The anatomic distribution tended to have lower error rates than random for small proportions of missing data, but in many methods (BPCA, *a priori* size regression, correlated variable regression) this rate spiked when analyses were conducted above a certain threshold of missing data. Unlike taxonomically biased missing data, our simulation of anatomically biased missing data suggested that this regionalized distribution can have an equal or greater effect on the amount of introduced error than either the proportion of missing data or the method of estimation used.

Discussion

This analysis has attempted to answer questions pertaining to three main ideas. 1) Which of the many possible approaches to either working around missing data, or estimating them from the existing data, introduces the least amount of error into the analysis? 2) How do the relative performances of these methods change as the amount of missing data increases? Is there a specific proportion or threshold of missing data estimation that should not be exceeded? 3) Does missing data having either taxonomic or anatomic biases result in differences in the results of the analysis than randomly allocated missing data? If the distribution of missing data does have an effect, do these biases result in a greater degree of estimation error?

Relative Performance of Missing Data Estimation Methods

Our study finds that approaches based on the substitution of the mean and pairwise deletion are likely to introduce the largest amount of error into the analysis, regardless of the proportion or distribution of the missing data. The use of these methods in estimation of missing data values is not recommended. The poor performance of mean substitution is neither surprising nor a new result. The literature testing its performance is well documented (Gornbein et al. 1992; Little 1992; Strauss and Atanassov 2006). Its inclusion is this study is mainly 2-fold, first to provide a context in which the relative performance of the other methods can be measured and, second to add to the investigation of the effect of non-randomly distributed missing data—even in poorly performing methods.

Gower's distance matrix consistently results in the lowest difference (error) between the result of the original data set and the results of data sets with missing data. Between 1% and 30% of missing data, the performance of BPCA is very similar, though slightly worse than that of Gower's distance matrix. Although Gower's distance matrix consistently introduced the least amount of error, it possesses two significant drawbacks. First, Gower's distance matrix does not return a data set comparable with the original data set, rather it returns a distance matrix with the dimensions being equal to the number of observations. Results of the analysis, therefore, cannot illustrate the effect of the variables and can only illustrate the similarity between specimens in regards to all variables. This approach is useful for analyses interested in the relative clustering of specimens (irrespective of the variables), but may not be practical when the effect of individual variables is of interest. Second, although it performs well at low amount of missing data, Gower's distance matrix cannot handle larger amounts of missing data. For this data set, it performed consistently <30%, inconsistently between 30% and 40%, and very rarely was able to handle 45% or more missing data. Preliminary results suggest that this limited range of performance is related to data set size and shape, and that its range of performance is more limited with smaller data sets.

If estimation of the missing data values is required, rather than generating a distance matrix, and the proportion of missing data is low, then BPCA is the most reliable method. At low proportions of random and taxonomically biased missing data (<35% missing, representative of most morphometric analyses), this study found that BPCA significantly outperforms all other estimation methods, and consistently results in the lowest amounts of error. Above 35% of missing data, BPCA shows an increase in error and is surpassed by correlated variable regression as the best performing estimation method. When the missing data are anatomically distributed, however, BPCA is the best performing estimation method across all proportions of missing data.

Interestingly, even when the range of specimen sizes is very large, and the scaling patterns between the variables fairly constrained, the correlated variable regression consistently outperforms the *a priori* size regression. This outcome suggests that regression estimation approaches, when possible, should not depend on the regression against an *a priori* size value, but rather, should use the variable that is most correlated with each other variable.

Although for random and taxonomically biased distributions, the correlated variable regression method outperforms BPCA >35%, the performance of this method (as well as the *a priori* size regression) is closely tied to the nature of the data, specifically the size range of the specimens and the isometric/allometric scaling patterns of the variables. The data set tested here, with its large range of specimen sizes, and the constrained scaling patterns of the variables, likely represents the best conditions for regression estimation techniques. Data sets with more restricted ranges of specimen sizes or less constrained variable scaling would almost certainly return poorer results. Further testing of these methods with differing data set types, sizes, and shapes is required to fully understand the general performance of these methods.

Threshold of Missing Data Estimation

The second question addressed in our study related to how estimation error is related to the proportion of missing data, and if there is a significant increase in error beyond a certain proportion, indicating a threshold of missing data beyond which it should not be estimated. Dodson (1975a) cited Pilbeam (1969) as suggesting that principal coordinates analysis can tolerate up to one-third of missing data. Strauss et al. (2003) found that the expected maximization and PC approach were reliable up to 50% of missing data for very small number of characters. But, their reliability was strongly dependent on size of the data set. It is unclear, however, on what they based their cut off for maximum estimation.

The majority of the methods and distributions tested herein showed either a linear or exponential/power correlation of increasing estimation error with increasing proportions of missing data. Simple methods, such as mean substitution, show a strong linear increase in estimation error in response to a linear increase in percent missing data, with no distinct increase in error rate at any one portion of missing data. Because of this, no mathematically derived maximum threshold of missing data that can, or should, be estimated can be easily derived or justified.

For methods showing a disproportional increase in estimation error with increased missing data (Gower's distance matrix, pairwise deletion, BPCA, and the two regressions), the increase in error is not constant, and the rate of error per unit missing data increases as the percentage of missing data values increases. Visual inspection of the graphs permits identification of regions where certain methods experience dramatic increases in the error (e.g., 30–40% and 75–80% for BPCA and 75–80% for the two regressions—Fig. 2). As illustrated in Figure 3, however, the distribution of missing data can dramatically affect these potential thresholds. Additionally, preliminary unpublished data, as well as previous studies (Strauss et al. 2003), suggest that the size and the shape (i.e., number of variable vs. number of specimens) of the data set can also have dramatic effect on the position of these potential thresholds. This finding suggests that there may not be a universal and distinct maximum amount of missing data that can or should be estimated in morphometric data sets. Rather, there is likely to be a combination of factors including, but potentially not limited to, the size and shape of the data set, the distribution of the missing data, and the choice of estimation method. Researchers should be aware of the issues involved with missing data, as well as the degree of error of estimation that would be associated with their particular data set. Further work is required to investigate this question.

Distribution of Missing Data

The disproportionate effect of the missing data estimation on underrepresented and morphological disparate specimens in the analysis, even when the missing data are randomly introduced, is an issue of concern. This is because it is often these very specimens (those of small sample size and those that are morphological disparate) that are 1) most likely to have missing data in the first place, and 2) are often of most interest to the scientist. The question of how a bias against poorly represented specimens in the distribution of missing data would confound this effect is, therefore, of great interest. The pattern seen when poorly represented specimens are subject to the highest amount of missing data is relatively constant across all estimation methods, with increased, but not dramatically, higher error rates. This finding indicates that estimates based on randomly input missing data will be reasonable proxies (will show similar trends), but will consistently underestimate the estimation error when there are systemic biases against completeness of rare taxa in data sets.

It is important to keep in mind that the measure of estimation error in these tests is the *Procrustes* sum of squares, which measures the total difference (error) of all values between both results (original and estimated). This, in effect, averages the error seen across all values. If the error of estimation is concentrated on a few outlying values, the majority of the values will match closely and have little error. When the missing data (and resulting error) are biased against these rare specimens, allowing for the specimens of common taxa to be free of missing data, the result is disproportionate, with greater sum of squares error across the entire result. If the total error is greater, and the common specimens are complete, the relative movement of the rare specimens must be very large.

In addition to biases in the distribution of missing data among specimens, there are also systemic biases in the distribution of missing data within specimens. Interestingly, when missing data are biased toward anatomical regions the resulting estimation error deviates much more from the random pattern than does the taxonomically biased sample. Not only is the average deviation greater, but the anatomically biased samples often show trends that distinguish them from those seen commonly with the taxonomic or random samples. This outcome suggests that whether or not the missing data are distributed at random within the specimens has a greater effect on the result than whether it is distributed at random between the specimens.

Not only is the magnitude of the effect intriguing, but its direction is as well. For most estimation methods, the anatomic distribution results in reduced estimation error—at least for small proportions of missing data. This is likely partly a result of the metric used to compare the estimation results, *Procrustes* sum of squares, but may indicate a more general phenomenon. As discussed previously, *Procrustes* will compare all data values in the two outputs, create the best match for all the data, and then summarize the total difference with the sum of squares metric. Regional biases in the distribution of missing data will result in certain anatomical regions having the majority of the missing values, whereas the remaining areas will encounter relatively little data loss. If the variables experiencing the loss are relatively consistent between specimens, as may be expected in morphological data sets with extremities and/or fragile areas, or as can be encountered in our simulation with measurements close together, systematic loss of information on the co-variance of these variables will occur. The unaffected variables, however, will likely perform with similar accuracy as seen in the original data set. The result of this pattern may be similar to what would be expected if the variables experiencing data loss were removed (listwise deletion) from the data set altogether.

As mentioned above, BPCA and Gower's distance estimation perform very similarly when the missing data are randomly assigned to the data set, with Gower's distance matrix showing slightly better performance. When the missing data are allocated with an anatomic bias, however, the gap in the performance of these two methods increases and BPCA shows less error than that seen in Gower's distance matrix if >15% of the data are missing.

The dramatically higher rate of error seen when anatomical missing data are handled using pairwise deletion is of particular interest. This large deviation from the pattern seen with the random simulation, and the increased rates of error, emphasizes the importance that both the distribution of missing data and its amount have on the resulting estimation error. Not only do the relative performance results suggest that pairwise deletion should be used with caution if used at all, but that its performance where the missing data show anatomic biases should be questioned.

It is almost certain that most biological data sets will have missing data with combined biases both among and within specimens, and that combining these scenarios may be a better proxy for biological data. Testing the effect of non-random distributions both among and within specimens in a single data set was not performed here, however, due to complexity of the model involved. It is noted that taxonomic bias (between specimens) and anatomic bias (within specimens) tested independently, may not be representative of the combined effects that may result when co-occurring in the same data set. Despite this possibility, we believe that the incorporation of testing of non-random distributions of missing data, within and among specimens, has increased our understanding of the effect these phenomena have on our analysis, and represents a marked improvement from testing of data missing at random.

Conclusion

BPCA and Gower's distance measure introduce the least amount of error when handling missing data, and are herein recommended when dealing with missing values in data sets. Pairwise deletion of missing data, and mean substitution introduced the greatest amount of estimation error, and are not recommended in general. Future work systematically testing the effect of data set size, and testing additional data sets is required to address this issue.

Mean substitution shows a linear relationship between the error of estimation and the proportion of missing data, does not show a disproportionate increase in error at a specific proportion of missing data, and, therefore, has is no obvious mathematical upper limit to missing data estimation. Most methods do show a disproportionate increase in estimation error as the percentage of missing data increases and some show dramatic increases at specific amounts of missing data. These dramatic increases may superficially be seen as logical upper limits to missing data estimation (and they may indeed be for this data set), but these sharp increases in error are a result of a combination of factors including data set size and shape, which have not been tested systematically here. General conclusions regarding the upper limit of the different estimation techniques await systematic testing of multiple independent data sets and subsets, to quantify the effect of data set size and shape.

Missing data input with biases both toward rare specimens and anatomical regions can dramatically affect both the amount of error at specific proportions of missing data, and the patterns of how estimation error responds to increases in missing data. In general, biases inputting more missing data in specimens that are represented by fewer numbers of specimens increase the amount of error seen in the overall results compared with randomly assigned missing data. Although consistent, the absolute difference is relatively small and the pattern of response to increasing missing data is consistent in direction. Biases distributing missing data to mimic loss of whole anatomical regions result in much greater deviation from the pattern seen with RMD. With smaller amounts of missing data, anatomical biases tend to show less error than with randomly assigned data, but as the proportion increases, this type of distribution experiences rapid increases sooner than is seen in the randomly generated samples. These results underscore the importance that the distribution pattern of the missing observations across the data set, not merely the proportion, has on the estimation error and the interpretation of the result.

Supplementary Material

Supplementary material, including data files and/or online-only appendices, can be found in the Dryad data repository (http://dx.doi.org/10.5061/dryad.m01st7p0).

Funding

This work was supported by two National Science and Engineering Research Council of Canada—Alexander Graham Bell Canada Graduate Scholarships; an NSERC Discovery Grant; a Doris O. and Samuel P. Welles Fund Travel Grant (UCMP); and a Dinosaur Research Institute Student Project Grant; and by the University of Toronto.

Acknowledgments

We would like to thank multiple individuals at the University of Toronto and Royal Ontario Museum for their contributions (both theoretical and practical) to the development and undertaking of this research project. D. Evans and R. Reisz provided guidance and support for the project. K. Brink, N. Campione, D. Larson, L. O'Brien, M. Ryan, R. Schott, J. Theodor, and M. Vavrek provided assistance, and discussion for this project, as well as computer processing power. Access to specimens was facilitated by D. Evans and K. Seymour (ROM—Vertebrate Palaeobiology), R. MacCulloch (ROM—Herpetology), K. Khidas and M. Steingerwald (CMN), W. Fitch (UCMZ), B. Strilisky (RTMP), A. Resetar (FMNH), G. Schneider (UM), J. McGuire (UCMVZ), and P. Holroyd (UCMP). P. D. Polly and an anonymous reviewer, as well as N. MacLeod and R. DeBry (editors), provided useful revisions and suggestions that greatly improved this article.

References

Beale E.M.L., Little R.J.A. 1975. Missing values in multivariate analysis. J. R. Soc. Ser. B 37:129–145.

Bookstein F.L. 1985. Morphometrics in evolutionary biology: The geometry of size and shape change, with examples from fishes. Philadelphia (PA): Academy of Natural Sciences of Philadelphia.

Bookstein F.L. 1997. Morphometric tools for landmark data. New York: Cambridge University Press.

Campione N.E., Evans D. C. 2011. Cranial growth and variation in *Edmontosaurs* (Dinosauria: Hadrosauridae): Implications for latest Cretaceous megaherbivore diversity in North America. PLoS One 6:e25186.

Densmore L.D. III, White P.S. 1991. The systematics and evolution of the crocodilia as suggested by restriction endonuclease analysis of mitochondrial and nuclear ribosomal DNA. Copeia 1991: 602–615.

Dimitriadou E., Hornik K., Leisch F., Meyer D., Weingessel A. 2010. e1071: Misc functions of the Department of Statistics (e1071) TU Wien. R Package, Version 1.5-23. Available from http://cran.R-project.org (last accessed May 26, 2012).

Dodson P. 1975a. Functional and ecological significance of relative growth in *Alligator*. J. Zoology 175:315–355.

Dodson P. 1975b. Taxonomic implications of relative growth in lambeosaurine hadrosaurs. Syst. Zoology 24:37–54.

Dodson P. 1979. Quantitative aspects of relative growth and sexual dimorphism in *Protoceratops*. J. Paleontology 50:929–940.

Dodson P. 1990. On the status of the ceratopsid *Monoclonius* and *Centrosaurus*. In: K. Carpenter Currie P.J., editors. Dinosaur systematics: approaches and perspectives. Cambridge (MA): Cambridge University Press, p 231–243.

Gornbein J.A., Lazaro C.G., Little R.J.A. 1992. Incomplete data in repeated measures analysis. Stat. Methods. Med. Res. 1:275–295.

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. 1975. Generalized Procrustes analysis. Psychometrika 40:33–51.

Hadfield J.D. 2008. Estimating evolutionary parameters when viability selection is operating. Proc. R. Soc. B Biol. Sci. 275:723.

Hammer Ø., Harper D. 2006. Paleontological data analysis. Malden (MA): Blackwell Publishing.

Harshman J., Huddleston C.J., Bollback J.P., Parsons T.J., Braun M.J. 2003. True and false gharials: a nuclear gene phylogeny of crocodylia. Syst. Biol. 52:386–402.

Kendall D.G. 1989. A survey of statistical theory of shape. Stat. Sci. 4:87–120.

Kramer A., Konigsberg L.W. 1999. Recognizing species diversity among large-bodied hominoids: a simulation test using missing data finite mixture analysis. J. Hum. Evol. 36:409–421.

Legendre P., Legendre L. 1998. Numerical ecology. 2nd ed. New York: Elsevier.

Little R.J.A. 1988. A test of missing completely at random for multivariate data with missing values. J. Am. Stat. Assoc. 83:1198–1202.

Little R.J.A. 1992. Regression with missing X's: a review. J. Am. Stat. Assoc. 87:1227–1237.

Maechler M., Rousseeuw P., Struyf A., Hubert M., Hornik K. 2010. cluster: cluster analysis basics and extensions. R Package, Version 1.12.3. Available from http://cran.R-project.org (last accessed May 26, 2012).

Matsumoto M., Nishimura T. 1998. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans. Model Comput. Simul. 8:3–30.

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

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

Norell M.A. 1989. The higher level relationships of the extant Crocodylia. J. Herpetology 23:325–335.

Oba S., Sato M.-a., Takemasa I., Monden M., Matsubara K.-i., Ishii S. 2003. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics 19:2088–2096.

Oksanen J., Blanchet F. G., Kindt R., Legendre P., O'Hara R. B., Simpson G. L., Solymos P., Henry M., Stevens H., Wagner H. 2010. Vegan: Community Ecology Package 1.17-2. Ordination methods, diversity analysis and other functions for community and vegetation ecology. R Package, Version 1.17-2. Available from http://cran.R-project.org (last accessed May 26, 2012).

Panneton F., L'Ecuyer P., Matsumoto M. 2006. Improved long-period generators based on linear recurrences modulo 2. ACM Trans. Math. Software 32:1–16.

Pilbeam D.R. 1969. Tertiary Pongidae of East Africa: evolutionary relationships and taxonomy. Bull. Peabody Mus. Nat. Hist. 31:1–185.

Pimentel R.A. 1979. Morphometrics, the multivariate analysis of biological data. Dubuque (IA): Kendall/Hunt Pub. Co.

Poe S. 1996. Data set incongruance and the phylogeny of crocodilians. Syst. Biol. 45:393–414.

R Core Development Team. 2011. R: A Language and Environment for Statistical Computing Vienna (Austria): R Foundation for Statistical Computing Available from: URL http://www.R-project.org/.

Revelle W. 2010. psych: Procedures for Personality and Psychological Research Version 1.01.9. Evanston (Il): Northwestern University.

Reyment R.A. 1991. Multidemenional palaeobiology. New York: Pergamon Press.

Rohlf F.J. 1990. Morphometrics. Ann. Rev. Ecol. Syst. 21:299–316.

Romer A. S. 1956. Osteology of the reptiles. Chicago (Il): University of Chicago Press.

Rubin D.B. 1976. Inference and missing data. Biometrika 63:581–592.

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. 1985. Evolutionary allometry and variation in body form in the South American catfish genus *Corydoras* (Callichthyidae). Syst. Zoology 34:381–392.

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. Linnean 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. Vertebrate Paleontology 23:284–296.

Venables W.N., Ripley B.D. 2002. Modern applied statistics with S 4th edn. New York: Springer.

Zelditch M.L., Sheets H.D., Fink W.L. 2003. The ontogenetic dynamics of shape disparity. Paleobiology 29:139–156.

Zelditch M.L., Swiderski D.L., Sheets H.D., Fink W.L. 2004. Geometric morphometrics for biologists. San Diego (CA): Elsevier Academic Press.