

Incomplete specimens in geometric morphometric analyses

Jessica H. Arbour* and Caleb M. Brown

Department of Ecology and Evolutionary Biology, University of Toronto, 25 Wilcocks Street, Toronto, ON M5S 3B2, Canada

Summary

1. The analysis of morphological diversity frequently relies on the use of multivariate methods for characterizing biological shape. However, many of these methods are intolerant of missing data, which can limit the use of rare taxa and hinder the study of broad patterns of ecological diversity and morphological evolution. This study applied a multi-data set approach to compare variation in missing data estimation and its effect on geometric morphometric analyses across taxonomically variable groups, landmark position and sample sizes.

2. Missing morphometric landmark data were simulated from five real, complete data sets, including modern fish, primates and extinct theropod dinosaurs. Missing landmarks were then estimated using several standard approaches and a geometric-morphometric-specific method. The accuracy of missing data estimation was determined for each estimation method, landmark position and morphological data set. Procrustes superimposition was used to compare the eigenvectors and principal component scores of a geometric morphometric analysis of the original landmark data, to data sets with A) missing values estimated, or B) simulated incomplete specimens excluded, for varying levels of specimens incompleteness and sample sizes.

3. Standard estimation techniques were more reliable estimators and had lower impacts on morphometric analysis compared with a geometric-morphometric-specific estimator. For most data sets and estimation techniques, estimating missing data produced a better fit to the structure of the original data than exclusion of incomplete specimens, and this was maintained even at considerably reduced sample sizes. The impact of missing data on geometric morphometric analysis was disproportionately affected by the most fragmentary specimens.

4. Missing data estimation was influenced by variability of specific anatomical features and may be improved by a better understanding of shape variation present in a data set. Our results suggest that the inclusion of incomplete specimens through the use of effective missing data estimators better reflects the patterns of shape variation within a data set than using only complete specimens; however, the effectiveness of missing data estimation can be maximized by excluding only the most incomplete specimens. It is advised that missing data estimators be evaluated for each data set and landmark independently, as the effectiveness of estimators can vary strongly and unpredictably between different taxa and structures.

Key-words: fossils, geometric morphometrics, missing data estimation, principal component analysis, procrustes analysis, R statistical software, shape

Introduction

The study of morphological evolution and disparity can help to connect factors and processes such as adaptation, innovation, developmental constraint and performance tradeoffs to ecological opportunity and ecological interactions between species (Foote 1997; Wainwright 2007; Gavrilov & Losos 2009; Glor 2010; Mahler *et al.* 2010). Analyses of morphological disparity through evolutionary time and across ecological communities commonly rely on multivariate methods for summarizing biological shape. This has often included the use of principal component analysis (PCA) of traditional morphological measurements that have been previously correlated with performance or ecological variables (Adams & Rohlf 2000; Mahler *et al.* 2010; Montaña & Winemiller 2010;

López-Fernández *et al.* 2013). Additionally, geometric (landmark) morphometrics have been used to summarize body shape variation (Adams & Rohlf 2000; Cardini, Higgins & Road 2004; Claude *et al.* 2004; Price *et al.* 2010; Rijssel & Witte 2012). Geometric morphometrics analyses the geometry of coordinate data representing morphological features to determine significant patterns of variation in shape and may integrate important spatial components of shape variation not captured using traditional linear morphometrics (Bookstein 1991, 1997; Willis 2001; Adams, Rohlf & Slice 2004; Zelditch *et al.* 2004).

Missing data are an unfortunate and nearly unavoidable problem in many areas of biology, from field observations, to phylogenetic coding, to specimens measurements (Wiens 2003b; Strauss & Atanassov 2006; Nakagawa & Freckleton 2008). Current multivariate morphometric methods (both traditional and geometric morphometrics) are intolerant of missing data, which can preclude the analysis of poorly preserved or

*Correspondence author. E-mail: jessica.arbour@utoronto.ca

damaged specimens (Adams, Rohlf & Slice 2004; Mitteroecker & Gunz 2009; Brown, Arbour & Jackson 2012). In particular, this greatly reduces inclusion of fossilized specimens, which are frequently subject to fragmentation, distortion and erosion (Gunz *et al.* 2004, 2009; Neeser, Ackermann & Gain 2009). This limits the analysis of ancient taxonomic groups, or the incorporation of related fossil specimens into the analysis of modern taxa, which may otherwise help to elucidate broad evolutionary patterns (Sallan & Friedman 2011). Although the effects of missing data (i.e. incomplete character coding of taxa) within cladistic analyses has been studied in great depth with simulations of both empirical and simulated data sets (Wiens 1998, 2003a,b, 2005, 2006; Kearney 2002; Kearney & Clark 2003; Santini & Tyler 2004; Moen 2008 – and references therein), the effect of missing data in a morphometric context has received much less attention (Strauss 2010).

Typically, the problem of missing data in morphometrics is accommodated by reducing the number of landmarks to exclude those missing from numerous specimens, or more frequently to exclude incomplete specimens from analyses (Strauss & Atanassov 2006). In the case of data sets with fossils or rare taxa, however, these are also typically represented by small sample sizes and the collection of additional, more complete, material may not be possible. These cases result in an unfortunate balancing act, where the number of specimens and overall completeness cannot both be simultaneously maximized, and a compromise must be reached (Strauss & Atanassov 2006). As a result, the removal of specimens from morphometric data sets can seriously hinder the power of statistical analyses (Nakagawa & Freckleton 2008; Brown, Arbour & Jackson 2012). For landmark data sets featuring bilaterally symmetrical structures, missing landmarks can be estimated effectively by methods such as mirroring along a midplane or aligning a specimen with its reflection (see Gunz *et al.* 2009 for a more thorough discussion of these methods). However, many data sets do not incorporate landmarks from both sides of a specimen/structure, and while several methods have been applied to estimate missing values in these cases, some of these possess serious weaknesses, including: underestimating the variability of a data set, which can affect test statistics (Little & Rubin 1987; Strauss, Atanassov & Alves de Oliveira 2003), sensitivity to the number of missing landmarks (Gunz *et al.* 2009; Neeser, Ackermann & Gain 2009), or requiring prohibitively large reference sets for accurate estimation (Neeser, Ackermann & Gain 2009).

The objective of this study was to evaluate the use of several missing data estimation methods, in particular how the use of estimated data influences the results of geometric morphometric analyses across data sets of varying size and structure. This study differs from many others by simulating missing data from real landmark sets of complete specimens spanning several taxonomic groups (both extant and extinct) and morphological structures, whereas many similar previous studies have typically focused on more narrow taxonomic groups and single structures (Gunz *et al.* 2004, 2009; Strauss & Atanassov 2006; Neeser, Ackermann & Gain 2009; Couette & White 2010; Brown, Arbour & Jackson 2012). This study also considered

not just the amount of missing data but also the resulting frequency of incomplete specimens, and the effect of removing these specimens from the data set. As a result, the information loss associated with specimen removal can be quantified and compared with the error introduced using different missing data estimators.

Materials and methods

MISSING DATA ESTIMATION

Several data sets of digitized landmarks were employed to evaluate the use of estimators for missing landmarks (Table 1). These data sets differed in terms of morphological structures examined, absolute size of these structures, taxonomic groups (and number of taxa included), variation in ontogeny as well as the number of specimens and landmarks included. All complete specimens within each data set were aligned using Procrustes superimposition ('procGPA' function, R package 'shapes'; Dryden 2009), after which incomplete specimens were aligned using Procrustes superimposition to corresponding landmarks in the consensus configuration of all complete specimens ('procOPA' function, R package 'shapes'). Both the consensus configuration and the final aligned landmarks were projected into the original Euclidean shape space of the data set.

Four missing data estimators previously examined in traditional linear or geometric morphometric analyses were applied to the five landmark data sets. Bayesian PCA (BPCA) incorporates principal component regression, Bayesian estimation and an expectation/maximization iterative algorithm to estimate missing data, the details of which are fully described by Oba *et al.* (2003). Bayesian PCA was implemented on aligned landmark data in R using the 'pca' wrapper function from the pcaMethods package v1.12.0 (Stacklies *et al.* 2007). A least-squares regression (REG), modified from the function 'best.reg' from the R package 'LOST' (Arbour & Brown 2012) and described in the study of Brown, Arbour & Jackson (2012), was used to independently estimate the x and y coordinate values for each missing landmark. Thin-plate spline interpolation (TPS), which uses a deformation grid to map the landmark locations of an incomplete specimen onto a reference configuration (in our case, the consensus configuration of complete specimens) to estimate missing landmark positions, was carried out using the function 'tps2d' from Claude (2008). Lastly, mean substitution (MS) was carried out using the mean coordinate position of each landmark based on the consensus configuration of the full data set.

INCOMPLETE SPECIMEN SIMULATION

We sampled a percentage of specimens from each data set to have landmarks removed, rather than a percentage of total data points to be removed, as this is more reflective of how incomplete data are handled in real analyses. For example, Couette & White (2010) stated that when 3% of landmarks datapoints were removed to simulate missing data, 90% of specimens would be incomplete. More realistic missing data simulation would involve a mix of complete and variably incomplete specimens (ex: Neeser, Ackermann & Gain 2009). Incomplete specimens were generated by randomly sampling a subset of specimens (between 1 and 90% of the total sample size) from each data set and removing, at random, between 1 and ~50% of their landmarks. These randomizations were carried out to simulate real damage to specimens;

Table 1. Summary information for each of the landmark data sets used to evaluate the effectiveness of missing data estimators

Data sets	Taxa	#S	#L	Specimen Information	Variation Explained by PCA axes 1 and 2, %	References
AC	Arctic char, <i>Salvelinus alpinus</i> (Salmonidae)	121	13	Full lateral profiles of adult polymorphic char from Lake Hazen, Ellesmere Island.	43.9, 17.2	Arbour, Hardie & Hutchings (2011)
GD	<i>Guianacara dacrya</i> (Cichlidae)	73	16	Full lateral profiles of juvenile and adult specimens from the Essequibo and Amazon river basins.	42.5, 16.4	Arbour & López-fernández (2011)
TY	Sixteen Tyrannosauroid dinosaurs species including the following genera: <i>Albertosaurus</i> , <i>Alioramus</i> , <i>Bistahieversor</i> , <i>Daspletosaurus</i> , <i>Dilong</i> , <i>Gorgosaurus</i> , <i>Guanlong</i> , <i>Nanotyrannus</i> , <i>Proceratosaurus</i> , <i>Raptorex</i> , <i>Tarbosaurus</i> , <i>Tyrannosaurus</i> and <i>Xiongguanlong</i>	35	25	Lateral profiles of the skull of multiple taxa from the Jurassic to Late Cretaceous of North America, Europe and Asia	33.2, 13.6	Shychoski, Rayfield & Sakamoto (2006)
HS	Human, <i>Homo sapiens</i> (Hominidae)	28	13	Landmarks taken from the midline of the brain from MRI images	20.3, 18.5	(Bookstein 1996; Dryden 2009)
PR	Hominid primates including Gorilla (<i>Gorilla gorilla</i>), Chimpanzee (<i>Pan troglodytes</i>) and Orangutan (<i>Pongo pygmaeus</i>)	167	8	Landmarks from the skulls of both male and female specimens (~equal representation per species and sex).	37.5, 28.1	O'Higgins & Dryden (1993)

#S, number of specimens; #L, number of landmarks.

not all specimens would exhibit the same level of incompleteness, nor necessarily identical regions of damage. We did not incorporate the effect of anatomical bias (i.e. missing landmarks are clustered on closely located anatomical features) or taxonomic bias (i.e. rare taxa are more likely to be incomplete) into our missing data simulations; however, Brown, Arbour & Jackson (2012) showed that randomly distributed missing data can be an effective proxy for more realistic, nonrandom distributions.

VARIATION IN THE ACCURACY OF MISSING DATA RECONSTRUCTION

We compared the variation between data sets, landmarks and missing data estimators in the accuracy of missing landmark estimation. For 'incomplete' specimens, landmarks were estimated (independent of the original values for missing landmarks) using each of the four missing data estimators. For each real data set, the average deviation (Euclidean distance) between the estimated and the original landmark coordinates as a percentage of centroid size (Neuser, Ackermann & Gain 2009) was calculated for each landmark and scaled to the consensus configuration of each of 500 simulated data sets (see Fig. 2). We also compared the average deviation per incomplete specimen (as a percentage of centroid size) produced by each estimation technique within each data set.

MISSING DATA IN GEOMETRIC MORPHOMETRIC ANALYSIS

Principal component analysis was carried out on the original and estimated data sets using the 'procGPA' function. To simulate the typical practice of excluding damaged specimens from geometric morphometric analyses, we removed all incomplete specimens from each simulation and separately analysed the set of 'complete' specimens (COM).

The effect of the estimation techniques (or exclusion of incomplete specimens) on geometric morphometric analyses was assessed by contrasting the eigenvectors and PC scores of the BPCA, REG, MS, TPS and COM data sets with that of the actual landmark data (i.e. original landmark data, including incompletes with estimation, or excluding incomplete specimens). By comparing the variance structure (in terms of the eigenvectors of the landmark data, generated using the 'procGPA' function) within each data set, we are able to examine the point at which the inclusion of reconstructed specimens more closely matched the variation present within the original data set, than by analysing complete specimens alone.

Procrustes superimposition was used to assess the estimation error for the first two principal component axes for the BPCA, REG, MS, TPS and COM (eigenvectors only) data sets from that of the original landmark data, using the 'protest' function in R ('vegan' package) (Mardia, Kent & Bibby 1979; Peres-Neto & Jackson 2001; Oksanen *et al.* 2010). Only the scores and eigenvectors from the first two principal component axes from each data set were examined to prevent the inclusion of noncritical axes of variation (Gauch 1982; Jackson 1993) and as these axes summarized the highest proportion of the variation in shape (see Table 2 for % variation explained by first two axes for each data set). The Procrustes sum of squares (PSS; the $\sum \text{distances}^2$ between two optimally translated and rotated sets of points) between the eigenvectors and PC scores of the original landmark data and of the estimated/excluded landmark data were examined across a range of frequencies (typically 1–90%) of incomplete specimens, with lower PSS indicating a stronger fit (reduced estimation error) to the original data. Any difference (PSS) between the estimated results and those of the original is error (i.e. lack of fidelity with the original data set) induced by estimation or exclusion.

The process of simulating landmark data with incomplete specimens and comparing the resulting BPCA, REG, MS, TPS and COM data to the original landmark data in each data set (AC, GD, TY, HS and PR,

see Table 1) was repeated 500 times for each frequency of incomplete specimens examined. Mean values and 95% confidence intervals for the PSS comparing BPCA/REG/MS/TPS/COM to the original landmark data, for both the eigenvectors and PCA scores, were calculated from log values to correct for skew (Bland & Altman 1996).

THE IMPACT OF SPECIMEN INCOMPLETENESS

The effect of increasing the number of landmarks on the fit between the eigenvectors and PCA scores was examined by constraining the number of landmarks removed to ~10, 25, 50 or 75% of the landmarks per specimen (with the exception of the PR data set due to the low number of landmarks) and maintaining the frequency of incomplete specimens at 20%. The above analysis was then repeated (using 500 iterations for each number of landmarks and percentage of incomplete specimens).

SAMPLE SIZE EFFECTS ON MISSING DATA ESTIMATION

The consistency of missing data estimation across various sample sizes within each of the original data sets was examined by randomly subsampling the data sets to 25, 50 and 75% of the total number of individuals in each data set (with the exception of the HS and TY data sets due to the low number of specimens). The number of landmarks removed was again randomized between 1 landmark and 50% of the total landmarks, with the frequency of incomplete specimens maintained at 20%. The original analysis was then repeated (using 500 iterations for each sample size and percentage of incomplete specimens). The PSS between the original and reconstructed PCA scores was scaled by the ratio between the full sample size to the subsample size to make the results directly comparable (otherwise fit would appear to increase as the sum of squared distances would automatically decrease with the number of data points being compared).

Results

LANDMARK RECONSTRUCTION

Within each data set, the average deviation from the original landmark position (i.e. residual) increased dramatically with the frequency of incomplete specimens up to ~10–30% and sometimes showed less marked increases at high frequencies (>80%), but did not differ significantly across moderate frequencies (Fig. 1a). **Regression-based estimation produced the lowest mean residuals in three out of the five data sets, while BPCA produced the lowest mean residuals in two data sets** (Fig. 1b). TPS performed the least consistently and in the AC and PR data sets resulted in markedly higher residuals than other estimation methods (56–172% higher residuals than the best estimator for each data set, respectively; Fig. 1b).

Relative to scale, the GD and AC data sets showed the highest estimation accuracy by landmark (Fig. 2). The accuracy of missing data estimation was, however, extremely variable across landmarks in each data set (Fig. 2). Some landmarks showed consistent patterns in estimation accuracy across each data set, for example the anterior most landmark in the AC data set (Fig. 2, furthest left landmark) was, on average, the most poorly estimated landmark across all methods, while caudal landmarks (Fig. 2, furthest right)

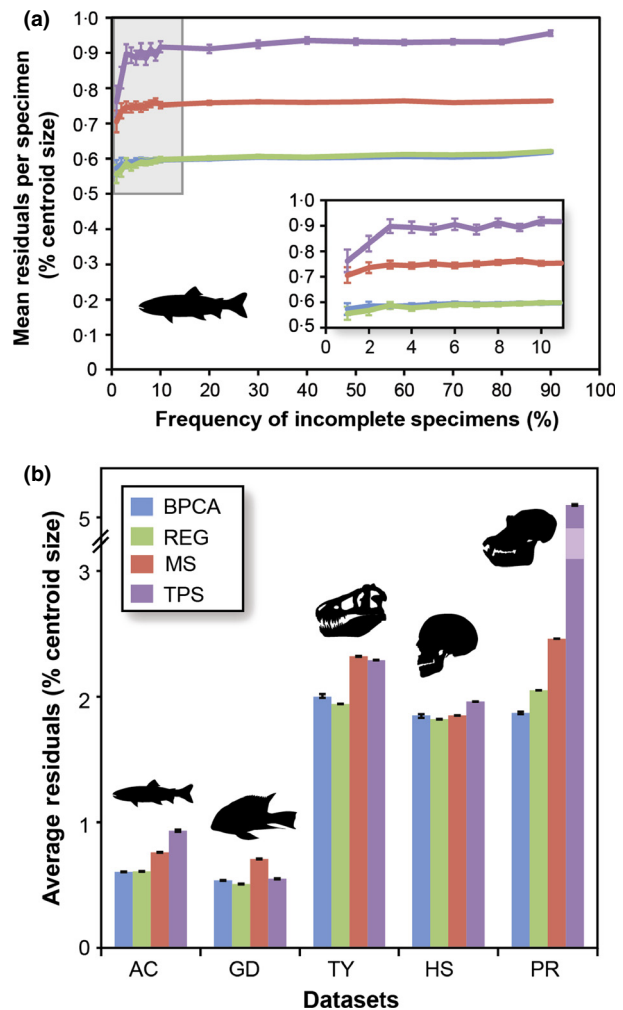


Fig. 1. Deviation from original landmark positions for each of four missing data estimators. (a) Mean residual per incomplete specimen (AC data set) between the estimated and actual landmark position for four missing data estimators (see legend in lower panel) standardized to the centroid size of the individual. (b) The average residuals between the reconstructed missing landmarks and the actual landmark coordinates per incomplete specimen, expressed as a percentage of the centroid size of each specimen. Residuals were determined from data sets containing 50% incomplete specimens. Means were calculated from 500 simulations and error bars indicate 95% confidence intervals for the mean residuals.

showed high estimation accuracy. In many cases, however, the most variable landmarks differed between estimation techniques. **For example, the most anterior (furthest left in Fig. 2) landmarks in the PR and TY data sets were much more poorly estimated by TPS than any other estimation method.**

THE EFFICIENCY OF MISSING DATA ESTIMATION

In all five data sets, most or all missing data estimators produced eigenvectors with a better fit (lower PSS) to the original data than those produced when incomplete specimens were excluded from analysis (COM). **In all data sets, either BPCA**

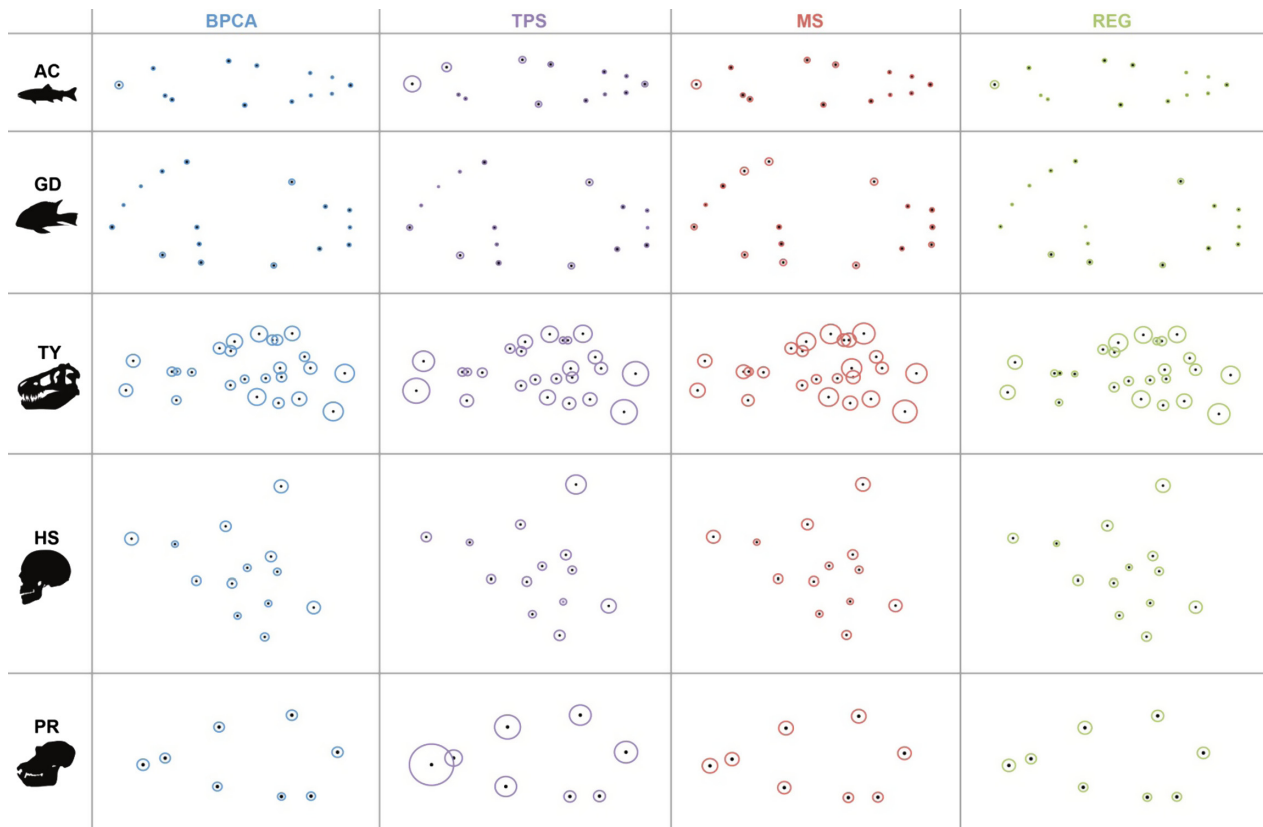


Fig. 2. Accuracy of landmark estimation by data set and estimation method. Radius of circles indicate the mean deviation of reconstructed landmarks from the original landmark position scaled to the consensus configuration for each data set using MS (red), BPCA (blue), REG (green) and TPS (purple). Residuals were determined from data sets containing 50% incomplete specimens. See Table 1 for descriptions of data sets.

or regression-based estimation produced the best fit between the original and estimated eigenvectors and PC scores (Fig. 3). Mean substitution produced a better fit to the original eigenvectors and PC scores than excluding incomplete specimens in all cases, and in some cases showed similar average performance to either BPCA or REG (TY and HS). TPS produced a better fit to the original data than using only fully complete specimens in three data sets (HS, GD, TY). However, up to 90% incomplete specimens in the AC data set and across all frequencies the PR data set, it produced a fit (to the original eigenvectors) that was not different from, or poorer than, using only complete specimens (Fig. 3, left). TPS produced a significantly poorer fit between the original and estimated PC scores than all other estimators in the PR, HS and AC data sets, as well as at moderate to high (>50%) frequencies of incompleteness in the TY data set (Fig. 3, right), and in all other cases performed significantly better than mean substitution only.

THE IMPACT OF SPECIMEN INCOMPLETENESS

The fit between the eigenvectors and PCA scores from the original and reconstructed data sets declined (higher PSS) with increasing number of missing landmarks per specimen for all missing data estimators (Fig. 4). BPCA, REG and MS exhibited similar patterns of increasing PSS for both eigenvectors and PCA scores with increasing number of missing landmarks,

and the percentage of specimen incompleteness did not alter which estimator showed the optimal fit to any given data set (Fig. 4). The impact of increasing the number of landmarks missing from incomplete specimens had a comparatively larger impact on TPS than other missing data estimators across all data sets (Fig. 4).

For all estimation methods, the fit between the original and reconstructed data sets for the initial simulations (up to 50% of landmarks per specimen were removed) were most similar to simulations in which the number of landmarks removed was constrained to 50% rather than when constrained to 25% (approximate average for initial simulations). For example, across all sampled frequencies of incomplete specimens in the HS data set, the regression coefficient between the mean PSS (for eigenvectors of BPCA vs. original data) from the initial analyses (removal of 1–7 landmarks) and the mean PSS from when 23% of landmarks were removed was 2.24 ($r = 0.99$), versus when 54% of landmarks were removed and the regression coefficient was 0.96 ($r = 0.99$).

SAMPLE SIZE EFFECTS ON MISSING DATA ESTIMATION

As the sample size decreased in each data set, the PSS between the eigenvectors resulting from only complete specimens and those of the original data set increased (Fig. 5, left). With the exception of TPS in the PR data set, the PSS

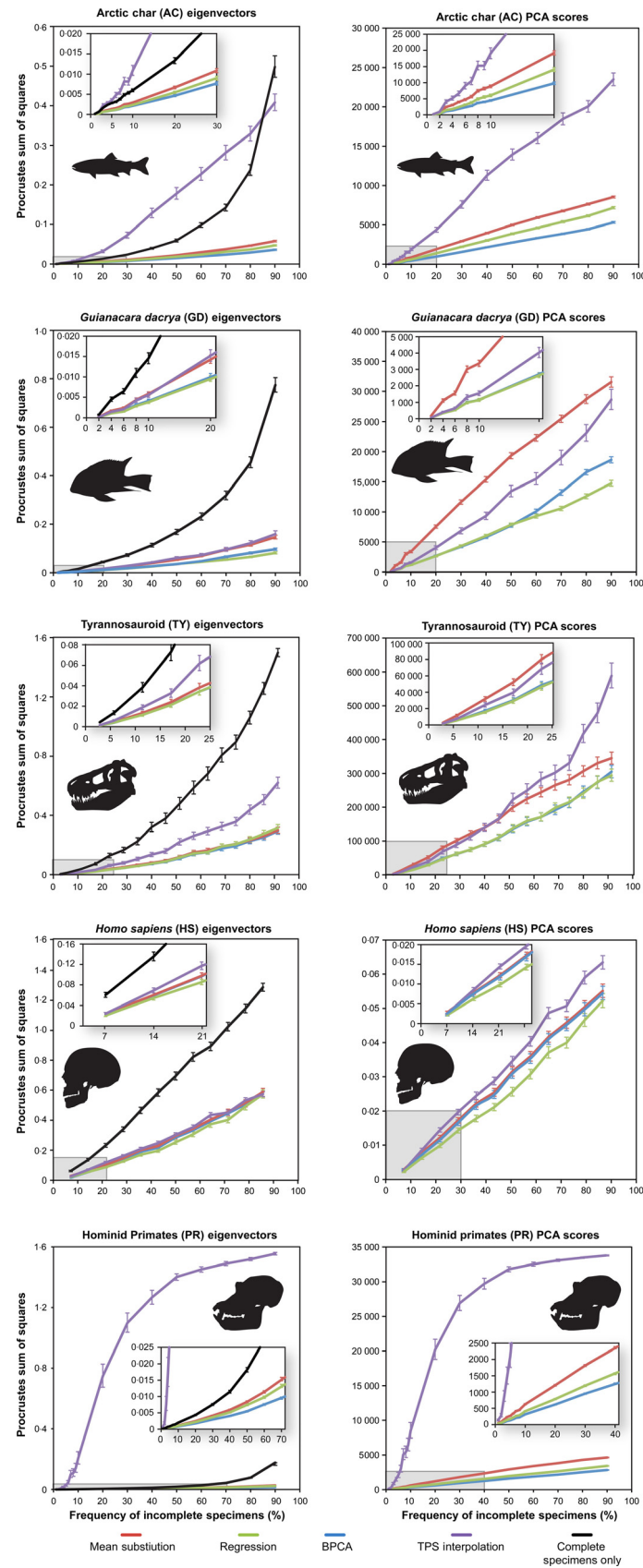


Fig. 3. A comparison of incomplete specimen reconstruction versus exclusion, and the effectiveness of estimators. Procrustes sum of squares from a comparison of the eigenvectors (left) and PCA scores (right), between the original and estimated (see legend) landmark data, when the frequency of incomplete specimens varied from 1 to 90%. Error bars indicate 95% confidence intervals around mean values.

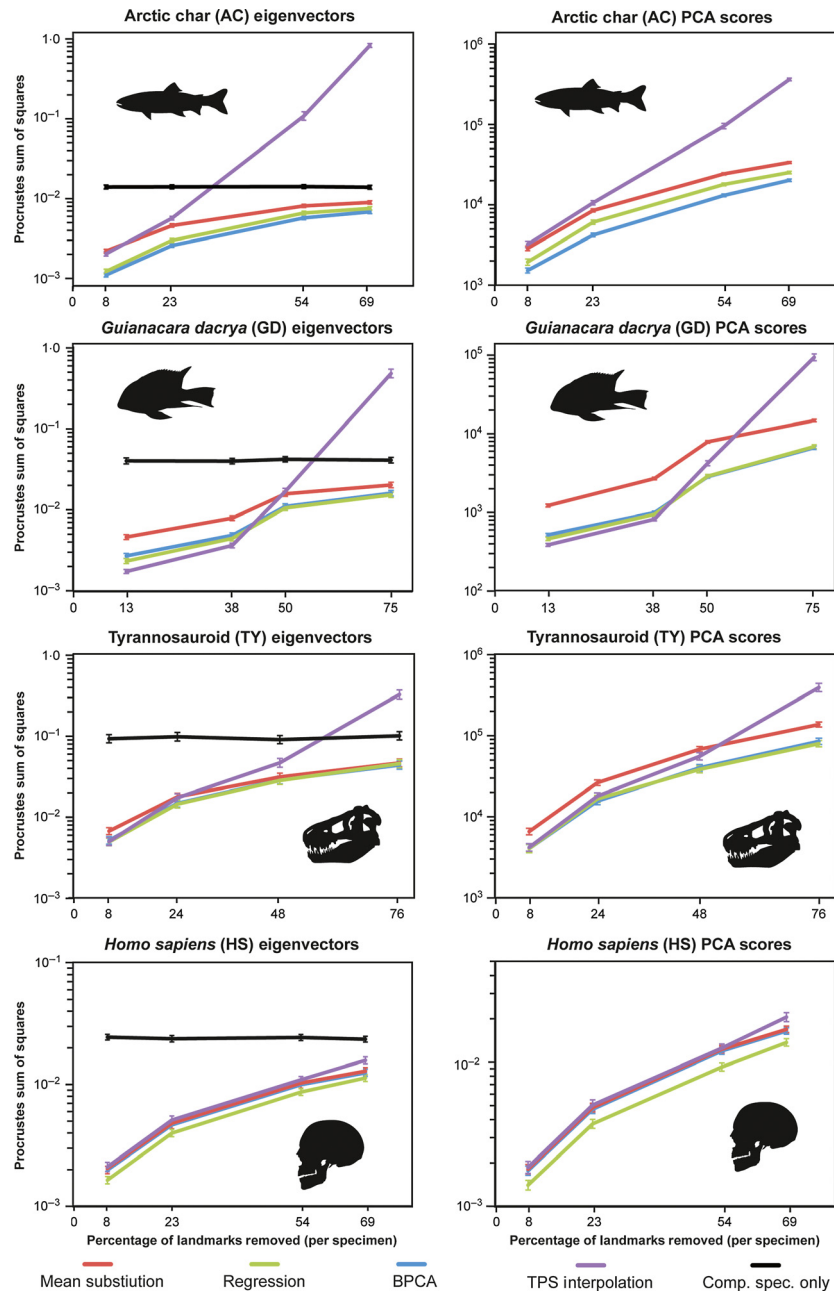


Fig. 4. The effect of varying the number of missing landmarks per incomplete specimen. Procrustes sum of squares from a comparison of the eigenvectors (left) and PCA scores (right), between the original and estimated (see legend) landmark data, when the frequency of incomplete specimens was 20% and the percentage of landmarks missing from incomplete specimens varied up to ~75% (x-axis). Error bars indicate 95% confidence intervals around mean values. The Y-axis (PSS) is presented as a log-scale to accommodate the proportionately high impact of TPS.

between the estimated and original data eigenvectors showed an overall increase with decreasing sample size, indicating a poorer fit at smaller sample sizes (Fig. 5). The GD data set showed greater stability in estimation error across sample size, with a substantial increase in PSS only occurring below 50% of the original sample size. Comparatively, both the PR and AC data sets show progressively higher PSS with each decrease in sample size. The relative performance of MS, BPCA and REG (as compared to excluding incomplete specimens) was relatively consistent across sample sizes, with no clear pattern: for example, REG performed better at small

sample sizes in the AC data set (33% of COM PSS at $n = 18$, and 41% of COM PSS at $n = 73$) but better at larger sample sizes in the GD data set (30% of COM PSS at $n = 18$, and 22% of COM PSS at $n = 73$). The size of the data set did not have an impact on whether any given estimator produced a better fit to the original eigenvectors than using only complete specimens (i.e. BPCA showed a better fit and TPS did not show a better fit than excluding incomplete specimens across all samples sizes in the AC data set; Fig. 5). Across the three data sets that were examined, the PSS for the PCA scores of MS, BPCA and REG did not show an

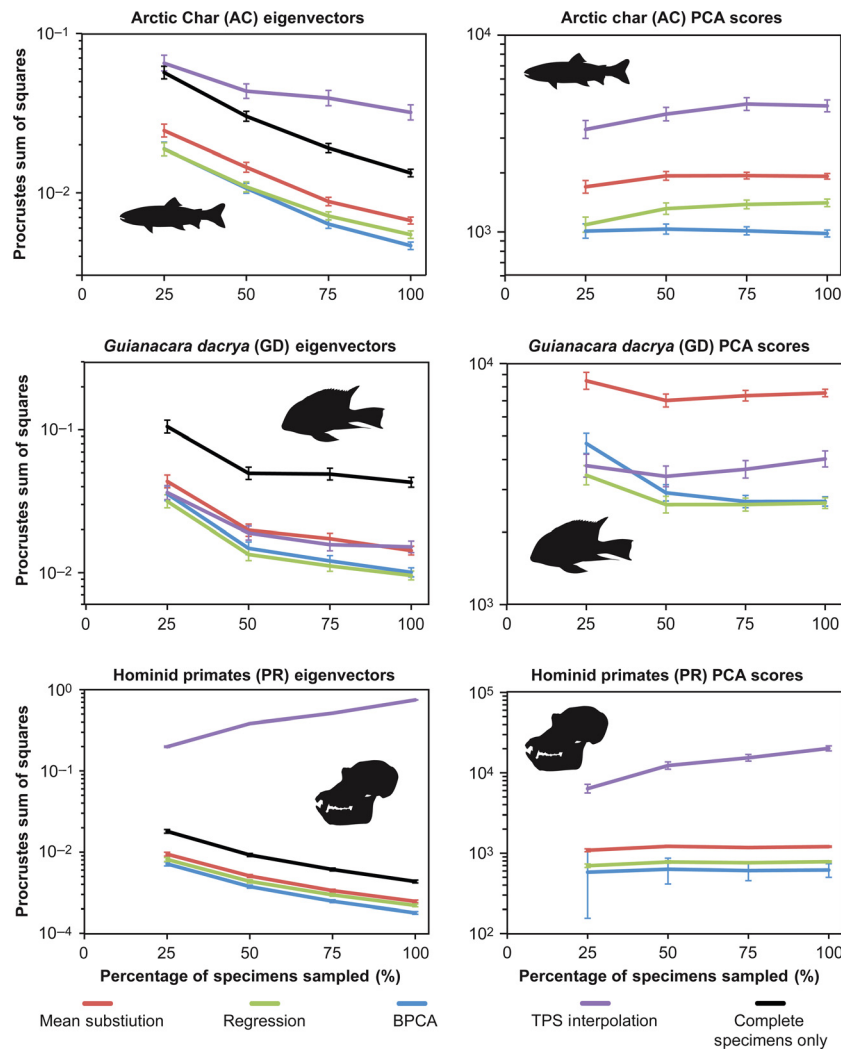


Fig. 5. The effect of varying sample size on the effectiveness of missing data estimation. Procrustes sum of squares from a comparison of the eigenvectors (left) and PCA scores (right), between a subsample (between 25 and 100%, X-axis) of the specimens in the original and estimated (see legend) data sets corresponding to these subsamples, when the frequency of incomplete specimens was 20%. Error bars indicate 95% confidence intervals around mean values. The y-axis (PSS) is presented as a log-scale to accommodate the proportionately high impact of TPS.

overall pattern of increase or decrease with sample size (Fig. 5, right). The estimation method that produced the lowest PSS values for each data set in the initial simulations also produced the lowest PSS values in each subsampling regime. TPS showed a unique pattern of marked increases in PSS with sample size for PCA in the PR data set, suggesting a poorer fit between the original and estimated data at larger sample sizes (Fig. 5, PR).

Discussion

The study of morphological disparity in an evolutionary or ecological context can be limited by the current inability of morphometric analyses to accommodate missing data (Adams, Rohlf & Slice 2004; Strauss 2010). We examined several taxonomically varied data sets, a combination of standard and geometric-based estimation methods, variation in specimen incompleteness, variation in sample size, as well as

compared missing data estimation to the more common practice of excluding incomplete specimens. Data sets reacted very differently to missing data estimation, and landmarks representing different structures within the same data set may vary in how well they can be estimated by different techniques. Across the data sets examined, missing data estimation was preferable to excluding incomplete specimens, and BPCA and REG were both reliable missing data estimators, with low sensitivity to sample size and specimen incompleteness.

Missing data estimation was most accurate in the AC and GD data sets, which both feature landmark data from full lateral profiles (all others include only a structure from the organism) and are both composed of Teleost fishes. Qualitatively, missing landmark estimation within these data sets appears to fall within the range of measurement error in landmark placement (J. H. Arbour, Pers. Obs.). This is despite the considerable ecological- and ontogenetic-based variation in shape within each data set (Arbour, Hardie & Hutchings 2011; Arbour &

López-fernández 2011). These data sets did not appear to have the most inherent shape structure: PC1 and PC2 summarized a cumulative total of 61.1% and 58.9% of shape variation in the AC and GD data sets, compared to 46.8% in TY, 38.8% in HS and 65.6% in PR. Nor did either data set have the largest sample size (PR), the largest or smallest number of landmarks (TY and PR respectively). Both the AC and GD data sets included only one species; however, the HS data set also included one species and showed much higher estimation error.

The overall estimation accuracy (relative to scale) was poorest for the TY data set, which also had the highest taxonomic diversity of the five data sets (Fig. 2). As this data set consisted entirely of fossil specimens, it is possible that taphonomic distortion or erosion decreased the amount of meaningful shape variation present in the data set and reduced the effectiveness of missing data estimation. However, at least for some missing data estimators (BPCA and REG), the accuracy of missing data estimation was comparable with a similarly sized but taxonomically less variable, extant taxa data set (HS). Furthermore, the analysis of eigenvectors suggests that even for this small, fossil-based data set, the error introduced by estimation altered the results of geometric morphometric analyses less than the removal of incomplete specimens would.

Especially for TPS, the accuracy of missing data estimation appeared to be related to the variation in shape. For example, the AC data set includes two morphotypes of *Salvelinus alpinus* that differ primarily in head shape and body depth (Arbour, Hardie & Hutchings 2011), and the highest variation in estimation accuracy was observed in landmarks corresponding to the head and dorsal fin region (left and upper landmarks in Fig. 2). In the original analysis of the PR data set, higher variation (especially associated with sexual dimorphism) was observed among facial landmarks compared with cranial landmarks (O'Higgins & Dryden 1993). TPS was found to very poorly estimate PR landmarks in general, but this was particularly the case for some facial landmarks (Fig. 2, PR purple, left and upper left landmarks). Interestingly, Neeser, Ackermann & Gain (2009) found the opposite pattern from the PR data set in landmark estimation accuracy in *Homo sapiens* skull characteristics: nonfacial landmarks had significantly higher estimation error than facial landmarks, and TPS performed better than mean substitution (MS). This may result from the higher number of landmarks (29 vs. 8 in the PR data) or the lower taxonomic diversity (one species with a substantial reference set of 178 complete skulls). Overall, maximizing the effectiveness of missing data estimation requires an understanding of which anatomical regions show the greatest shape variation. The use of TPS should be evaluated on a case-by-case basis; however, the effectiveness of this method is likely comparable with other estimators when taxonomic diversity is low and the number of landmarks is high, or if there is a well chosen reference specimen (Gunz *et al.* 2009; Neeser, Ackermann & Gain 2009).

For practical reasons, we used only single imputation in examining the performance of missing data estimation (i.e. for each simulated data set, only one set of coordinates was estimated for each missing landmark). Comparatively, multiple imputation creates sets estimated values for each missing data

point, after which the mean and total standard error including estimation uncertainty can be determined (Rubin 1987). For example, under a regression-based multiple imputation method, sets of imputed values could be drawn from a distribution around the value estimated by a regression function with the standard deviation taken from the complete values for that variable (Schafer 1997), and this could easily be applied to the regression function examined here. For practical applications of the methods described in this paper, where the true values of missing landmarks are unknown, it is advised that multiple imputation is incorporated to more accurately reflect the uncertainty associated with missing data estimation.

The decision to estimate missing data or exclude incomplete specimens is rooted in the trade-off between the additional information represented by incomplete specimens versus the error introduced during estimation. These analyses suggest that several missing data estimators are more effective at reconstructing patterns of shape variation (as summarized by its eigenvectors) than using only fully complete specimens. This would support that for some missing data estimators, the additional shape information captured by incomplete specimens has a greater impact on geometric morphometric analyses than the error introduced during estimation. While some previous studies have concluded only small amounts of missing data can be estimated reliably, these studies did not consider the loss of information resulting from the exclusion of incomplete specimens (Strauss, Atanassov & Alves de Oliveira 2003; Couette & White 2010).

The results herein are somewhat paralleled by those found in the investigations of the effect of missing data on phylogenetic resolution in cladistic analyses, using both simulated and empirical data sets. There it was found that it is not the amount of missing data within a taxon, that was important to its effect of the resolution, rather it was the quality and the characters that are present – 'too few characters' rather than 'too many missing characters' (Kearney & Clark 2003; Wiens 2003a,b; Santini & Tyler 2004). Based on this, the inclusion of incomplete specimens, particularly those preserving character combinations that resolve polytomies and long-branch attraction, often improves the resolution of the phylogeny relative to that based on only complete taxa, and incomplete taxa should not be excluded *ad hoc*.

In the present study, the patterns observed in the fit between the reconstructed and original landmark data when the three data sets were subsampled reflected the patterns observed when the whole data sets were analysed. This indicates that if a sufficient portion of a data set was comprised of complete specimens (in our analyses this would be at least 25% of the total sample) the simulations described in this study, could be used to evaluate the effectiveness of missing data estimators prior to their application. This would allow for the optimal methods specific to each data set to be determined. Alternatively, if the use of missing data estimators cannot be evaluated on a subsample of complete specimens, it is advisable to use BPCA, regression-based or otherwise similar method to estimate missing data, as these were the most consistent across all data sets examined.

It is interesting to note that the fit between the morphological relationships (eigenvectors) and individual specimen values (PCA scores) between the original and reconstructed data was more strongly related to the maximum rather than the average number of landmarks removed. This is not necessarily surprising considering the substantial reduction in fit to the original data set when the number of landmarks removed was increased (Fig. 3). This suggests that while the use of effective missing data estimators can be extremely beneficial in approximating the holomorphospace, or absolute morphospace (Willis 2001), of a group of organisms, the usefulness of missing data estimation could be maximized in some cases by excluding a small number of the most damaged/fragmented specimens especially those where missing landmarks are associated with highly variable anatomical structures.

Theoretically, a perfect missing data estimator would eventually produce results in geometric morphometric analyses that converged with that of a data set containing only the complete specimens when the sample size became sufficiently large to have sampled the entire holomorphospace of a group of organisms (Willis 2001). Under this scenario, 'perfectly' reconstructed incomplete specimens would not represent novel, un-sampled regions of the morphospace and would provide no additional information on the patterns of morphological diversity within the group. Until a sufficient sample size for a particular group is available to reach this limit, the inclusion of incomplete specimens by way of effective missing data estimation (for example, by using BPCA or regression-based methods) provides a better approximation of the morphological relationships present within a group of organisms than complete specimens alone.

Acknowledgements

We are grateful to H. Lopez-Fernandez (Royal Ontario Museum), J. Hutchings (Dalhousie University) and L. Shychoski (University of Alberta) for permission to utilize the GD, AC and TY data sets (respectively) in this study. We are also grateful to D. Staples (York University) for providing constructive comments to an early draft of the manuscript and to V. Arbour (University of Alberta) for assistance in data acquisition. This study was partly funded by an Alexander Graham Bell Canada Graduate Scholarship from NSERC to JHA.

Data accessibility

Landmark coordinate data is accessible for the AC and GD datasets through Dryad (Arbour & Brown 2014) and the HS and PR data sets are accessible as data files through the R package 'shapes' (Dryden 2009). See Table 1 for references associated with the TY data set. Functions for the alignment of incomplete landmark data sets and the application of missing data estimators are available through the R package 'LOST' (Arbour & Brown 2012).

References

Adams, D. & Rohlf, F.J. (2000) Ecological character displacement in *Plethodon*: biomechanical differences found from a geometric morphometric study. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 4106–4111.

Arbour, J.H. & Brown, C. (2012) LOST: Missing morphometric data simulation and estimation. Retrieved from: <http://cran.r-project.org/web/packages/LOST/index.html>

Arbour, J.H. & Brown, C.M. (2014) Data from: Incomplete specimens in geometric morphometric analyses. *Dryad Digital Repository*, doi:10.5061/dryad.mp713

Arbour, J.H., Hardie, D.C. & Hutchings, J.A. (2011) Morphometric and genetic analyses of two sympatric morphs of Arctic char (*Salvelinus alpinus*) in the Canadian High Arctic. *Canadian Journal of Zoology*, **89**, 19–30.

Arbour, J.H. & López-fernández, H. (2011) *Guianacara dacrya*, a new species from the rio Branco and Essequibo River drainages of the Guiana Shield (Perciformes: Cichlidae). *Neotropical Ichthyology*, **9**, 87–96.

Bland, J.M. & Altman, D.G. (1996) The use of transformation when comparing two means. *British Medical Journal*, **312**, 1153.

Bookstein, F.L. (1991) Landmarks. *Morphometric Tools for Landmark Data: Geometry and Biology* (ed. F.L. Bookstein), pp. 55–87. Cambridge University Press, Cambridge.

Bookstein, F.L. (1996) Biometrics, biomathematics and the morphometric synthesis. *Bulletin of Mathematical Biology*, **58**, 313–365.

Bookstein, F.L. (1997) Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. *Medical image analysis*, **1**, 225–243.

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. *Systematic biology*, **61**, 941–954.

Cardini, A., Higgins, P.O. & Road, C. (2004). Patterns of morphological evolution in Marmota (Rodentia, Sciuridae): geometric morphometrics of the cranium in the context of marmot phylogeny, ecology and conservation. 385–407.

Claude, J. (2008). *Thin-Plate Splines. Morphometrics with R*. pp. 181–188. Springer, New York.

Claude, J., Pritchard, P., Tong, H., Paradis, E. & Auffray, J.-C. (2004) Ecological correlates and evolutionary divergence in the skull of turtles: a geometric morphometric assessment. *Systematic Biology*, **53**, 933–948.

Couette, S. & White, J. (2010) 3D geometric morphometrics and missing data. Can extant taxa give clues for the analysis of fossil primates? *Comptes Rendus Palevol*, **9**, 423–433.

Dryden, I. (2009). shapes: Statistical shape analysis. Retrieved from <http://cran.r-project.org/package=shapes>

Foote, M. (1997) The evolution of morphological diversity. *Annual Review of Ecology and Systematics*, **28**, 129–152.

Gauch, H.G. Jr (1982) Noise reduction by eigenvector ordinations. *Ecology*, **63** (1643–1649), 5.

Gavrilets, S. & Losos, J.B. (2009) Adaptive radiation: contrasting theory with data. *Science*, **323**, 732–737.

Glor, R.E. (2010) Phylogenetic Insights on Adaptive Radiation. *Annual Review of Ecology, Evolution, and Systematics*, **41**, 251–270.

Gunz, P., Mitteroecker, P., Bookstein, F.L. & Weber, G.W. (2004). Computer-aided reconstructions of incomplete human crania. *Enter the Past: Computer Applications and Quantitative Methods in Archaeology, BAR International Series 1227*, pp. 92–94. Archaeopress, Oxford.

Gunz, P., Mitteroecker, P., Neubauer, S., Weber, G.W. & Bookstein, F.L. (2009) Principles for the virtual reconstruction of hominin crania. *Journal of human evolution*, **57**, 48–62.

Jackson, D. (1993) Stopping rules in principal component analysis: a comparison of heuristical and statistical approaches. *Ecology*, **74**, 2204–2214.

Kearney, M. (2002) Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. *Systematic Biology*, **51**, 369–381.

Kearney, M. & Clark, J.M. (2003) Problems due to missing data in phylogenetic analyses including fossils: a critical review. *Journal of Vertebrate Paleontology*, **23**, 263–274.

Little, R.J.A. & Rubin, D.B. (1987) *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

López-Fernández, H., Arbour, J.H., Winemiller, K.O. & Honeycutt, R.L. (2013) Testing for ancient adaptive radiations in Neotropical cichlid fishes. *Evolution*, **67**, 1321–1337.

Mahler, D.L., Revell, L.J., Glor, R.E. & Losos, J.B. (2010) Ecological opportunity and the rate of morphological evolution in the diversification of Greater Antillean anoles. *Evolution*, **64**, 2731–2745.

Mardia, K.V., Kent, J.T. & Bibby, J.M. (1979). *Multivariate Analysis*. Academic Press, London, UK.

Mitteroecker, P. & Gunz, P. (2009) Advances in Geometric Morphometrics. *Evolutionary Biology*, **36**, 235–247.

Moen, D.S. (2008) Missing data and the accuracy of Bayesian phylogenetics. *Evolution*, **46**, 307–314.

Montaña, C.G. & Winemiller, K.O. (2010) Local-scale habitat influences morphological diversity of species assemblages of cichlid fishes in a tropical flood-plain river. *Ecology of Freshwater Fish*, **19**, 216–227.

Nakagawa, S. & Freckleton, R.P. (2008) Missing inaction: the dangers of ignoring missing data. *Trends in Ecology Evolution*, **23**, 592–596.

- Neeser, R., Ackermann, R.R. & Gain, J. (2009) Comparing the accuracy and precision of three techniques used for estimating missing landmarks when reconstructing fossil hominin crania. *American Journal of Physical Anthropology*, **140**, 1–18.
- Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K. & Ishii, S. (2003) A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, **19**, 2088–2096.
- O'Higgins, P. & Dryden, I.L. (1993) Sexual dimorphism in hominoids: further studies of craniofacial shape differences in *Pan*, *Gorilla* and *Pongo*. *Journal of Human Evolution*, **24**, 183–205.
- Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H. & Wagner, H. (2010). *vegan*: Community Ecology Package. Retrieved from <http://cran.r-project.org/package=vegan>
- Peres-Neto, P. & Jackson, D. (2001) How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia*, **129**, 169–178.
- Price, S.A., Wainwright, P.C., Bellwood, D.R., Kazancioglu, E., Collar, D.C. & Near, T.J. (2010) Functional innovations and morphological diversification in parrotfish. *Evolution*, **64**, 3057–3068.
- Rijssel, J.C. & Witte, F. (2012) Adaptive responses in resurgent Lake Victoria cichlids over the past 30 years. *Evolutionary Ecology*, **27**, 253–267.
- Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.
- Sallan, L.C. & Friedman, M. (2011) Heads or tails: staged diversification in vertebrate evolutionary radiations. *Proceedings of the Royal Society B*, **279**, 2025–2032.
- Santini, F. & Tyler, J.C. (2004) The importance of even highly incomplete fossil taxa in reconstructing the phylogenetic relationships of the tetraodontiformes (Acanthomorpha: Pisces). *Integrative and Comparative Biology*, **44**, 349–357.
- Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.
- Shychoski, L., Rayfield, E.J. & Sakamoto, M. (2006). *A geometric morphometric and finite element analysis investigating Tyrannosauroid phylogeny and ontogeny emphasizing the biomechanical implications of scale*. PhD thesis, University of Britsol, Britsol.
- 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. (2010) Prospectus: The Future of Morphometrics *Morphometrics for Nonmorphometricians* (ed. A.M.T. Elewa), pp. 345–352. Springer-Verlag, Berlin.
- 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. *Biological Journal of the Linnean Society*, **88**, 309–328.
- Strauss, R.E., Atanassov, M.N. & Alves de Oliveira, J. (2003) Evaluation of the principal-component and expectation maximization methods for estimating missing morphometric studies. *Journal of Vertebrate Paleontology*, **23**, 284–296.
- Wainwright, P.C. (2007) Functional versus morphological diversity in macroevolution. *Annual Review of Ecology, Evolution, and Systematics*, **38**, 381–401.
- Wiens, J.J. (1998) Does adding characters with missing data increase or decrease phylogenetic accuracy? *Systematic Biology*, **47**, 625–640.
- Wiens, J.J. (2003a) Incomplete taxa, incomplete characters, and phylogenetic accuracy: is there a missing data problem? *Journal of Vertebrate Paleontology*, **23**, 297–310.
- Wiens, J.J. (2003b) Missing data, incomplete taxa, and phylogenetic accuracy. *Systematic Biology*, **52**, 528–538.
- Wiens, J.J. (2005) Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? *Systematic Biology*, **54**, 731–742.
- Wiens, J.J. (2006) Missing data and the design of phylogenetic analyses. *Journal of Biomedical Informatics*, **39**, 34–42.
- Willis, M.A. (2001). Morphological disparity: a primer. *Fossils, Phylogeny and Form* (eds M. Adrian, G.D. Edgecombe & B.S. Lieberman), pp. 55–144. Kluwer Academic, New York.
- Zelditch, M.L., Swiderski, D.L., Sheets, H.D. & Fink, W.L. (2004). *Geometric Morphometrics for Biologists: A Primer*. Elsevier Academic Press, London.

Received 20 August 2013; accepted 9 October 2013

Handling Editor: Shinichi Nakagawa