Comparing the Accuracy and Precision of Three Techniques Used for Estimating Missing Landmarks when Reconstructing Fossil Hominin Crania

Rudolph Neeser,[1]* Rebecca Rogers Ackermann,[2] and James Gain[1]

[1]*Department of Computer Science, University of Cape Town, Rondebosch 7701, South Africa*
[2]*Department of Archaeology, University of Cape Town, Rondebosch 7701, South Africa*

KEY WORDS hominoids; estimation; mean substitution; thin plate splines; regression

ABSTRACT Various methodological approaches have been used for reconstructing fossil hominin remains in order to increase sample sizes and to better understand morphological variation. Among these, morphometric quantitative techniques for reconstruction are increasingly common. Here we compare the accuracy of three approaches—mean substitution, thin plate splines, and multiple linear regression—for estimating missing landmarks of damaged fossil specimens. Comparisons are made varying the number of missing landmarks, sample sizes, and the reference species of the population used to perform the estimation. The testing is performed on landmark data from individuals of *Homo sapiens*, *Pan troglodytes* and *Gorilla gorilla*, and nine hominin fossil specimens. Results suggest that when a small, same-species fossil reference sample is available to guide reconstructions, thin plate spline approaches perform best. However, if no such sample is available (or if the species of the damaged individual is uncertain), estimates of missing morphology based on a single individual (or even a small sample) of close taxonomic affinity are less accurate than those based on a large sample of individuals drawn from more distantly related extant populations using a technique (such as a regression method) able to leverage the information (e.g., variation/covariation patterning) contained in this large sample. Thin plate splines also show an unexpectedly large amount of error in estimating landmarks, especially over large areas. Recommendations are made for estimating missing landmarks under various scenarios. Am J Phys Anthropol 140:1–18, 2009.

Taphonomic damage or distortion of hominin fossil material is widespread, and many techniques have been used to reconstruct such specimens. These reconstructions have often involved traditional approaches—such as the physical manipulation of remains by a trained anatomist—although with the advent of computer-based techniques reconstruction methods have increasingly used virtual, statistical, and morphometric tools (for some examples, ranging from digital mirroring to thin plate splines, see: Conroy et al., 1998, 2000; Ponce de León and Zollikofer, 1999; Zollikofer et al., 2002, 2005; Neubauer et al., 2004). Analytic procedures are advantageous in that they generally increase a reconstruction's repeatability and come with associated error metrics. Both of these allow a reconstruction to be more readily reasoned over and falsified (Zollikofer and Ponce de León, 1998; Weber, 2001).

However, though these approaches offer great promise and may be less subjective than more traditional approaches, little has been done to compare the real world efficacy of the various morphometric techniques, or to understand how the amount of fossil or extant comparative material available to the researcher might affect the accuracy of these reconstruction methods. The study reported here explores the relative accuracy and precision of morphometric-based techniques in fossil reconstruction by comparing the results obtained from three analytic techniques used for estimating missing landmarks: mean substitution, thin plate spline warping, and multiple linear regression. Although the estimation of missing landmarks is only a small part of the reconstruction process, it plays a pivotal role: landmark estimates typically guide the reconstruction of the remaining interlandmark morphology, and are the data that can most easily be examined and tested.

We have focused on these three techniques—rather than, say, shape space analysis—for the following reasons. First, thin plate spline techniques are widely used in the literature (see examples below). Second, regression-based methods are also generally mentioned, though are rarely used, being considered accurate but too reliant on large data sets. Further, regression-based methods are biologically relevant and intuitive; the implications of this given our results will be explored further in the Discussion. Third, composite reconstruction methods and mirroring (both special cases of mean substitution) have been widely used by researchers, though mean substitution is probably the approach with the weakest support biologically. As such, we consider it to be a baseline technique which the others should always outperform.

We compare the techniques by estimating landmarks for samples of individuals drawn from extant gorilla,

Grant sponsors: The National Research Foundation of South Africa, The Paleontological Scientific Trust.

*Correspondence to: Rudolph Neeser, Department of Computer Science, University of Cape Town, Private Bag Rondebosch 7701, South Africa. E-mail: rudy@cs.uct.ac.za

Received 14 July 2008; accepted 17 December 2008

DOI 10.1002/ajpa.21023
Published online 10 February 2009 in Wiley InterScience (www.interscience.wiley.com).

human, and chimpanzee populations as well as a sample of fossil hominin specimens. In particular, our analyses are structured to manipulate the effects of landmark loss (i.e., the amount of damage), reference-sample sizes (i.e., the amount of data used to estimate the missing landmarks), and the species of the population from which the reference-sample was drawn (which may differ from the species of the individual being reconstructed) to assess the accuracy of the different reconstructions under different circumstances (i.e., how close to the true landmark an estimate is, as measured by mean residuals). We also examine some precision-related issues, here measured by error spread.

MATERIALS AND METHODS

The techniques

Mean substitution (MS). MS replaces a missing landmark with the landmark's average position, as calculated over a sample of undamaged reference specimens. It proceeds by aligning an undamaged reference form to the damaged individual's form, and substitutes the reference form's landmarks for those missing on the damaged form. It is wise to use more than a single reference individual: an average Procrustes form can be determined from a sample of undamaged reference individuals, thus limiting the effect of any individual with unexpected or atypical morphology. An extra scaling step may also be used: the reference form is scaled so as to match the damaged individual in centroid size. We use such a scaling step.

Mirroring, a special case of MS, is performed by substituting a damaged individual's contralateral landmarks (if present) for any missing landmarks.

Thin plate spline substitution (TPS). TPS is an extension of MS that also substitutes landmark values, but first uses thin plate splines to fit the average form to what remains of the damaged individual in an attempt to better match the damaged individual's known morphology. TPS has been used in paleoanthropology both for estimating landmark positions and for replacing missing anatomy with the known morphology of an undamaged specimen (e.g., Ponce de León and Zollikofer, 1999; Neubauer et al., 2004). The splines are defined by specifying homologous points between two individuals. These points are mapped exactly to one another, whereas the remaining landmarks are mapped via an interpolation that attempts to minimize bending energy, an integral of curvature. See Bookstein (1991) for more details.

Multiple linear regression (RM). RM models the relationships between landmarks over a set of undamaged reference specimens to obtain a collection of regression coefficients. These coefficients, along with the nonmissing landmarks of the damaged individual, are used to predict the position of missing landmarks. Of the three methods used, RM makes the most use of biological information—specifically that captured by the pattern of variation/covariation (V/CV) between the landmark positions. Although it is adapted from the literature, it is not widely used, and what follows is therefore a more detailed methodological description than that given for the other two approaches.

We use a method similar to that of Richtsmeier et al. (1992). All individuals are represented as form matrices, which is a matrix whose components represent distances between landmarks. The missing distances are estimated using a regression method—Richtsmeier proposed the use of projection pursuit regression, whereas we use standard multiple linear regression (the lm function of the R statistical language, R Development Core Team, 2005).

For each of the missing distances in the damaged individual, regression coefficients are calculated from a reference-sample, but only using the distances not "homologous" to any of the missing distances. These coefficients are used—along with the damaged individual's nonmissing distances—to estimate those that are missing. Once all distances have been estimated, multidimensional scaling (Cox and Cox, 1994) is used to convert the distances into Euclidean coordinates. These coordinates need to be aligned to those of the damaged individual, and the missing landmarks are estimated by substituting their homologous counterparts (we use this substitution due to both numerical imprecision and application of least mean squares in the multidimensional scaling causing the Euclidean coordinates of the nonmissing landmarks to be slightly shifted—the original coordinates should be used instead).

Of course, the general problem with regression methods is that the reference-sample must be at least as large as the number of predictor variables. Because available data are at a premium, we reduce the number of variables used in the estimation: the form matrix diagonal contains distances of landmarks to themselves—this is always zero and so is removed; form matrices are symmetric, hence either the lower or upper triangles may be removed.

We also remove variables by considering those that contribute to multicollinearity; these exist due to the extreme interdependence between interlandmark distances (Richtsmeier et al.,1992). This is done using the principal component analysis (PCA) method of Jolliffe (1986): in essence, PCA is used to determine which distance variables explain the least variance in the data set. Those that explain zero variance (i.e., have a zero eigenvalue) are not independent of the other predictors, and are removed.

The method is as follows: after removing both the diagonal and either the lower or upper triangles, each distance matrix is converted to a vector of distances (by, say, removing a matrix component subscript) and is supplied as input to PCA. PCA returns eigenvectors and their associated eigenvalues (variances). The eigenvectors with eigenvalues below a cutoff point (here set at 0.5×10^{-6}—because of numerical imprecision we cannot expect eigenvalues to ever be exactly zero) are considered to explain little additional variance. A distance variable is associated with each such eigenvector and is removed from the set of predictor variables. This association is done by noticing that the nth component of the input variables to the PCA method contributes to the nth component of any eigenvector. We locate the largest component in the eigenvector and its associated position in the input vectors. But this position merely encodes a distance variable, and so this variable is no longer used as a predictor variable.

If on looking for the largest eigenvector component, we locate a distance variable that has already been removed, we instead search the eigenvector for its next largest component. Once all the eigenvectors with eigenvalues below the cutoff are associated with a distance variable, we use the remaining distance variables as pre-

dictors. We found that this method reduces the number of variables to slightly below those of coordinate-based approaches. If larger data reduction is required, larger cutoff values can be used. Jolliffe (1986) suggests the value 0.7 (far larger than the value of 0.5×10^{-6}, because we were only interested in removing multicollinearities).

More details on these three estimation techniques can be found in Neeser (2007).

The tests

This article uses the following terminology: within-species estimation uses individuals of the same species for both the test-sample (those individuals being reconstructed) and the reference-sample (individuals used to perform the reconstruction). Across-species estimation draws the reference-sample from a species other than that of the test-sample individuals.

Because of the scarcity of fossil material, the bulk of the testing is performed using landmark data from individuals of three extant primate species: *Homo sapiens*, *Pan troglodytes,* and *Gorilla gorilla*. Landmark data were obtained from previous studies using Microscribe and Polhemus contact digitizers (see Ackermann, 1998). Each analysis uses two samples: the actual test-sample and a reference-sample. These samples never overlap. A maximum of 67 chimpanzee, 107 gorilla, and 628 human individuals were available to the authors. All individuals are adult, and all samples have roughly balanced the numbers of male and female individuals. Each individual is represented by up to 29 landmark measurements, shown in Figure 1. Of these, most of the analyses are carried out on thirteen landmarks—NA, NSL, IS, FMN (Left and Right), ZI (L/R), FM (L/R), ZTS (L/R), MT (L/R)— to reduce the number of variables involved in the reconstructions while maximizing data overlap with the fossil specimens.

Means are compared using Welch's approximate *t*-test for heteroscedastic data (Sokal and Rohlf, 1995), because *F*-tests show that the variances of the residuals are unequal.

Analysis I: estimating landmarks using MS, TPS, and RM. The analysis is repeated for all three methods. Three test-samples (of $n = 10$) are drawn, one each of *H. sapiens*, *P. troglodytes,* and *G. gorilla*. A reference-sample of 57 individuals is drawn from each species. For each reference-sample, a consensus form is created for both MS and TPS, whereas regression-coefficients are calculated for RM (as previously described). The analysis is carried out for each consensus form/test-sample pair, and is repeated for each estimation technique. The analysis itself proceeds as follows: for each test individual, a landmark is removed and treated as missing. The landmark is then estimated using an estimation technique, as previously outlined. This is repeated for each landmark.

The residuals between the landmarks' true and estimated positions are calculated, and an average residual is calculated for each individual. The mean of the individual averages gives a mean residual for each test-sample, reference-sample, and estimation technique combination.

Analysis II: estimating the point at which RM outperforms MS and TPS. Unlike TPS and MS, regression techniques are highly dependent on reference-sample sizes, and we wish to know how large a sample RM requires to be competitive with MS and TPS. This test repeatedly corrects the same test-sample while increasing the reference-sample sizes in increments of 10 individuals, from 10 to 600. Only a human reference-sample is used for this analysis, as it is the only data set available to the authors of large enough size. There are, however, also chimpanzee and gorilla test-samples. All these samples are of $n = 28$; as there is no need to construct reference-samples for the chimpanzee and gorilla data sets, some of the individuals can be used in the test-samples. Landmark estimations are performed as outlined in Analysis I.

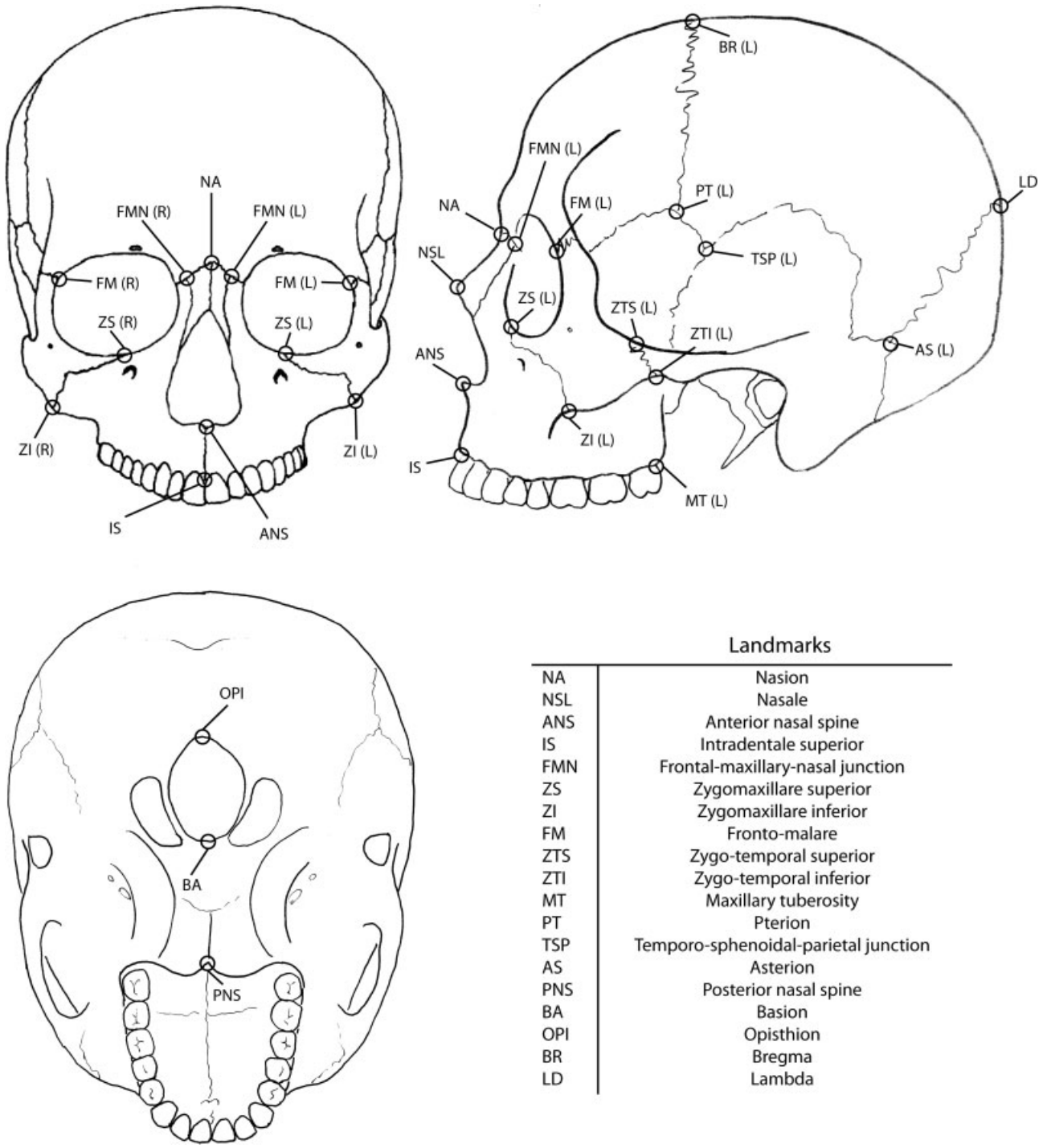
Analysis III: examining the increase in estimation errors with increasing number of missing landmarks. The previous analyses calculate residuals as if only a single landmark were missing. This is unrealistic, as taphonomic distortion typically affects multiple landmarks. Analysis III examines how an increasing number of missing landmarks affects landmark estimates, with the number of missing landmarks being a proxy for the amount of specimen damage. All the test individuals (three test-samples, each with $n = 28$, constructed as previously) are corrected using a human reference-sample of 600 individuals, allowing RM to perform without the hindrance of small sample sizes. Each individual is corrected 10 times, each iteration removing, in succession, 1–10 landmarks. The missing landmarks are estimated as previously (only now with less information about the damaged individual), and a residual is calculated. A mean is calculated over these residuals.

A loss of 10 landmarks represents a 77% landmark loss, and leaves three landmarks for use by the estimation techniques, some of which require a minimum of three landmarks to function. The test is also repeated using a single, random reference individual and the following changes: RM is not tested; small reference-samples free up chimpanzee and gorilla data for use as reference individuals.

Analysis IV: testing which estimated landmarks produce the largest residuals. This analysis uses all 29 landmarks to examine the distribution of error over the whole cranium. A human reference-sample ($n = 178$) is used to correct a human test-sample ($n = 33$). The analysis uses all three methods and proceeds as in Analysis I to obtain residuals and means.

We also use this full data set to test the efficacy of a special case of MS—mirroring. We implement this as follows: the sagittal plane is calculated by taking the third principal component of all mid-sagittal landmarks, and the constant in the parametric plane equation is calculated as the dot product between Nasion and the normal. A missing lateral landmark is estimated by taking its corresponding contralateral landmark (called CL) and calculating its orthogonal projection (OP) onto the sagittal plane. The line from CL to OP is extended by its distance through OP—the end of this line is the landmark's estimated position. Only landmark positions for the left lateral landmarks are estimated (the residuals for the right lateral landmarks are identical). Residuals are examined and compared with the other methods, as above.

Analysis V: estimating landmarks for fossil specimens. This analysis examines the behavior of the techniques when "correcting" fossil specimens. This analysis



Landmarks	
NA	Nasion
NSL	Nasale
ANS	Anterior nasal spine
IS	Intradentale superior
FMN	Frontal-maxillary-nasal junction
ZS	Zygomaxillare superior
ZI	Zygomaxillare inferior
FM	Fronto-malare
ZTS	Zygo-temporal superior
ZTI	Zygo-temporal inferior
MT	Maxillary tuberosity
PT	Pterion
TSP	Temporo-sphenoidal-parietal junction
AS	Asterion
PNS	Posterior nasal spine
BA	Basion
OPI	Opisthion
BR	Bregma
LD	Lambda

Fig. 1. Landmarks used in this study.

does not attempt to estimate landmarks that are truly missing from these fossils. Rather, we remove and estimate known landmarks, allowing us to determine the amount of reconstruction error associated with each technique. Landmark estimations are performed as in Analysis I.

Various species are used in this analysis: the australopiths are represented by *Australopithecus africanus* (STS 5, Taung), *Paranthropus boisei* (KNM-ER 406, KNM-WT 17400), and *Paranthropus aethiopicus* (KNM-WT 17000). Members of our genus are represented by *Homo habilis* (KNM-ER 1470, KNM-ER 1813) and *Homo erectus* (KNM-ER 3733, KNM-WT 15000). Each specimen displays a different state of preservation which has affected which landmarks have been collected, as listed in Table 1. This specimen list also includes juveniles (Taung, KNM-WT 15000) and a young adult (KNM-WT 17400).

Each specimen is corrected once with each method/reference-sample combination (human: $n = 628$; chimpanzee:

TABLE 1. Results from estimating landmarks on fossil specimens

Specimen	Species	Recorded landmarks	MS			TPS			RM		
			C	G	H	C	G	H	C	G	H
STS 5	Australopithecus africanus	NA, NSL, IS, FMN, ZI, FM left, FM right, ZTS right, MT left, MT right	8.42	9.31	10.88	8.24	9.20	11.92	13.57	6.91	7.68
KNM-ER 1470	Homo habilis	NA, NSL, IS, FMN left, FMN right, FM left, FM right	14.66	15.47	14.00	20.76	25.25	17.64	8.10	10.49	8.58
KNM-ER 1813	Homo habilis	NA, NSL, IS, FMN left, FMN right, FM right, MT left, MT right	9.16	11.67	9.45	11.19	14.09	10.32	10.80	8.89	5.29
KNM-ER 3733	Homo erectus	NA, NSL, IS, FMN left, FMN right, ZI left, ZI right, FM left, ZTS left, ZTS right, MT left, MT right	8.84	10.34	8.64	10.11	9.39	4.44	18.98	8.24	5.32
KNM-ER 406	Paranthropus boisei	NA, NSL, IS, FMN left, FMN right, ZI left, ZI right, FM left, FM right, ZTS left, ZTS right, MT left, MT right	16.16	16.95	16.38	19.60	16.01	15.70	34.46	29.92	14.62
KNM-WT 17400	Paranthropus boisei	NA, NSL, IS, FMN left, FMN right, MT left, MT right	9.80	11.40	11.62	15.63	19.11	15.90	8.91	9.65	7.83
Taung	Australopithecus africanus	NA, NSL, IS, FMN left, FMN right, ZI left, ZI right, FM left, FM right, ZTS right, MT left	4.50	5.69	4.68	4.53	5.67	5.30	4.55	3.41	3.30
KNM-WT 15000	Homo erectus	NA, IS, FMN left, FMN right, ZI left, ZI right, FM left, FM right, ZTS right, MT left	9.90	10.90	9.35	16.14	14.71	6.22	9.36	10.70	8.03
KNM-WT 17000	Paranthropus aethiopicus	NA, NSL, IS, FMN left, FMN right, ZI left, ZI right, FM left, MT left, MT right	16.52	18.13	19.30	19.09	18.60	8.29	32.82	12.44	10.86
Total means			10.83	12.12	11.56	13.57	13.89	11.34	16.84	11.68	8.09
Method means				11.50			12.94			12.21	

Mean residuals are listed for reference samples (chimpanzee, gorilla, human) and estimation technique. The total means are weighted means. For landmarks which occur in pairs, such as FMN for example, listing only FMN indicates that both FMN left and FMN right were measured. Measurements are in millimeters.

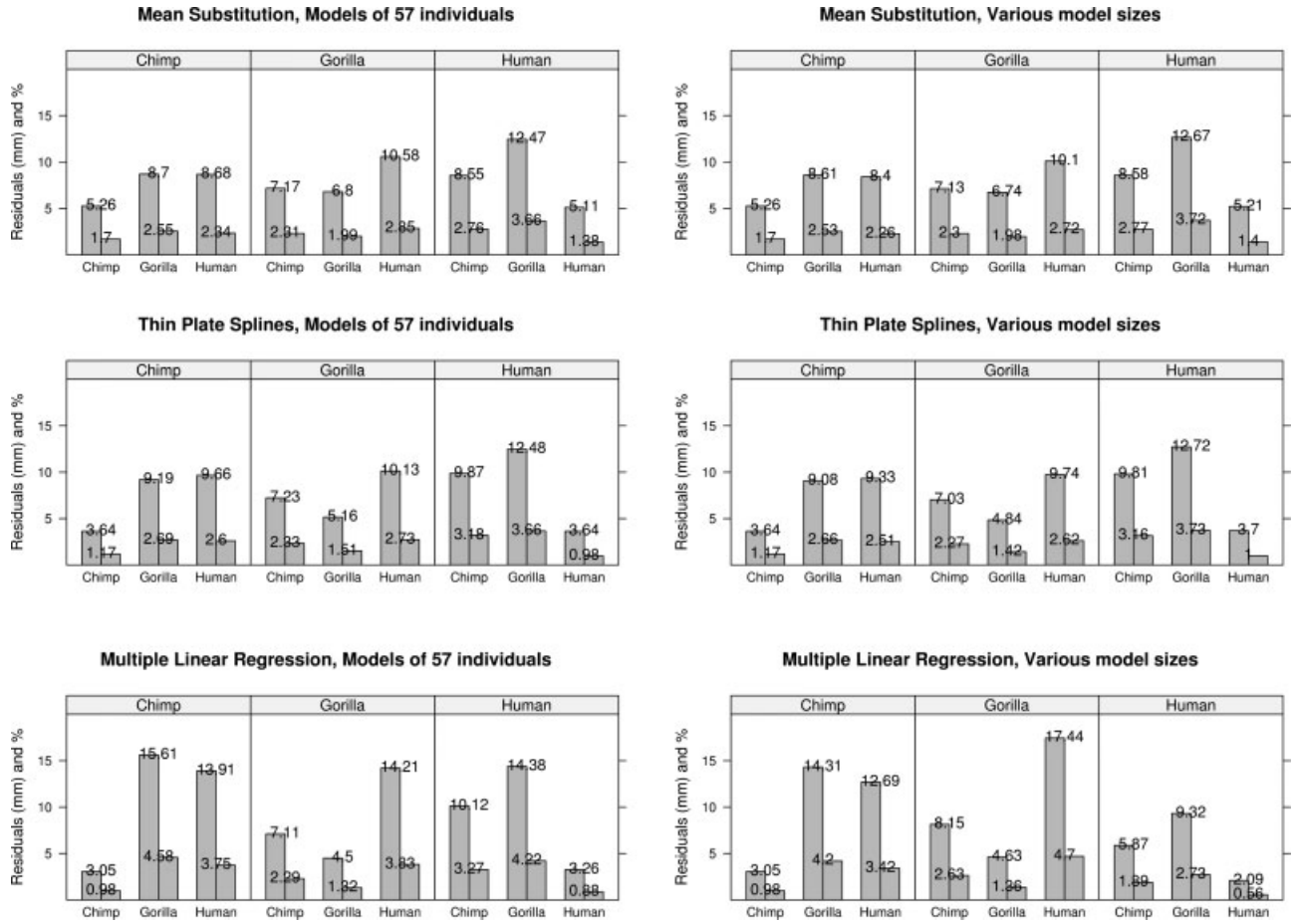


Fig. 2. The residuals obtained from MS, TPS and RM. Corrections are performed with within- and across-species reference-samples. The *x*-axis shows the species of the test-sample, the groupings show the species of the reference-sample. Each result is given twice: the right bar in mm, the left bar as the percentage of the test-sample's mean centroid size. The left column gives results for the 57 individual reference-samples, and the right column the larger reference-samples.

$n = 67$; gorilla: $n = 117$). Each landmark is in turn removed then estimated, as in Analysis I.

Because a mixed reference-sample might better represent the shared, ancestral population's pattern of covariation in early hominins (this corresponds to the strategy of taxon-independent inference proposed in Zollikofer and Ponce de León (2005) for reconstruction work), we also created pooled, heterogeneous reference-samples that consist of individuals from each of the three extant species for use by the RM method. We created two such groups: the first, consisting of an equal number of chimpanzee, gorilla, and human individuals (67 for each species, for a total of 201 individuals), and a second consisting of all of our data pooled, to create a sample of 812 individuals. Data were analyzed as above.

It is important to keep in mind that this analysis may be methodologically problematic for a number of reasons. First, the varying number and selection of landmarks may have unwanted effects (varying the number of landmarks certainly affects how much information is available to each technique). Second, we cannot be completely sure how much plastic deformation may be affecting the specimens (more than likely there is some), and hence the obtained residuals. Finally, the fossils are drawn from various species, and it is possible that different techniques may be more or less appropriate for reconstructing certain species (e.g., certain shapes).

RESULTS

All the results use a $P < 0.05$ significance level unless stated otherwise. Reported values are in millimeters (indicated by mm). Where multiple test-samples drawn from different species are used, these values may be followed in parenthesis with the value reported as a percentage of the sample's mean centroid size. This percentage is calculated as follows: first a percentage is calculated for each residual, then the percentages are averaged across landmarks, then across individuals; centroid sizes are always calculated using the full 29 landmarks, rather than only the 13 landmarks typically used in the analyses. Although metrics standardized by size are interesting and illuminating, absolute error sizes remain the critical metric to be considered.

Analysis I: estimating landmarks using MS, TPS, and RM

Figure 2 shows the obtained sample means for Analysis 1. The combined-means calculated across all test and reference-sample combinations are as follows: MS $\overline{\overline{X}} = 8.15\,\text{mm}$ (2.392); TPS $\overline{\overline{X}} = 7.89\,\text{mm}$ (2.318); RM $\overline{\overline{X}} = 9.57\,\text{cm}$ (2.790). For comparison with later tests, means obtained from larger reference-samples are also supplied in Figure 2 [chimpanzee reference-sample

TABLE 2. Comparison of results from analyses I, using 57 reference individuals

	TPS				RM			
		Reference species				Reference species		
	Test species	C	G	H	Test species	C	G	H
MS	C	8.794×10^{-4} TPS	0.860 –	0.024 MS	C	1.674×10^{-4} RM	0.885 –	0.109 –
	G	0.497 –	2.097×10^{-4} TPS	0.990 –	G	7.384×10^{-5} MS	9.576×10^{-8} RM	0.034 MS
	H	0.046 MS	0.290 –	1.523×10^{-4} TPS	H	0.030 MS	4.578×10^{-5} MS	9.961×10^{-6} RM
		Reference species						
	Test species	C	G	H				
RM	C	0.265 –	0.813 –	0.798 –				
	G	1.251×10^{-4} TPS	0.051 –	0.050 –				
	H	0.067 –	1.845×10^{-5} TPS	0.288 –				

Columns and rows represent the estimation methods being compared. Each cell shows the observed P value, and the method with the lower mean residual is indicated for means which are significantly different ($P < 0.05$).

TABLE 3. Comparison of the results obtained in analysis I, using larger, unequal reference-samples

	TPS				RM			
		Reference species				Reference species		
	Test species	C	G	H	Test species	C	G	H
MS	C	8.794×10^{-4} TPS	0.747 –	0.034 MS	C	1.674×10^{-4} RM	0.200 –	2.151×10^{-6} RM
	G	0.379 –	6.463×10^{-8} TPS	0.935 –	G	1.208×10^{-6} MS	3.965×10^{-6} RM	2.998×10^{-8} RM
	H	0.152 –	0.049 TPS	2.2×10^{-16} TPS	H	9.228×10^{-11} MS	2.2×10^{-16} MS	2.2×10^{-16} RM
		Reference species						
	Test species	C	G	H				
RM	C	0.265 –	0.179 –	1.485×10^{-6} RM				
	G	5.837×10^{-6} TPS	0.573 –	9.275×10^{-7} RM				
	H	1.072×10^{-7} TPS	2.2×10^{-16} TPS	2.2×10^{-16} RM				

Compares the means obtained for Analyses I using the larger, unequally sized reference-samples. Columns and rows represent the estimation methods being compared. Each cell shows the observed P value, and the method with the lower mean residual is indicated for means which are significantly different ($P < 0.05$).

size = 57, gorilla = 97, human = 280; MS $\overline{\overline{X}} = 8.12$ mm (2.278); TPS $\overline{\overline{X}} = 7.79$ mm (2.186); RM $\overline{\overline{X}} = 9.88$ mm (2.745)]. The difference between the combined-means for smaller and larger sample sizes is not significant for the first two methods, and increasing the sample sizes also shows no significant change in the means of the individual tests. For RM, the observed P-value between these means is 0.109, which approaches significance at the 0.1 level, but not at the 0.05 level used here. The difference between the test-sample means obtained for RM as reference-sample sizes vary are all significantly different, as is the human test-sample corrected with the gorilla sample. All the other differences are not significant. Tables 2 and 3 contains the means for each method, test-sample and reference-sample. Table 2 compares the tests using 57 reference individuals, and Table 3 compares the tests which use the unequal reference-sample sizes.

Analysis II: estimating the point at which RM outperforms MS and TPS

Figure 3 displays the results obtained from Analysis II, overlaying the MS, TPS, and RM residuals. Table 4 shows t-test comparisons between the RM, MS, and TPS methods. From this it appears that, given a large enough reference-sample, RM outperforms both LM and TPS, and this is discussed further later. Figure 4 shows the test repeated for reference-samples incrementing from 1 to 10 individuals, allowing us to examine how MS and TPS react to small reference-sample sizes, this, unfortunately, being the norm in paleoanthropological work. RM is not included in this analysis due to the extremely small reference-samples. With the smaller reference-samples, enough data are available to us to construct reference-samples from each species. Using these small samples, we see that with the human reference samples,

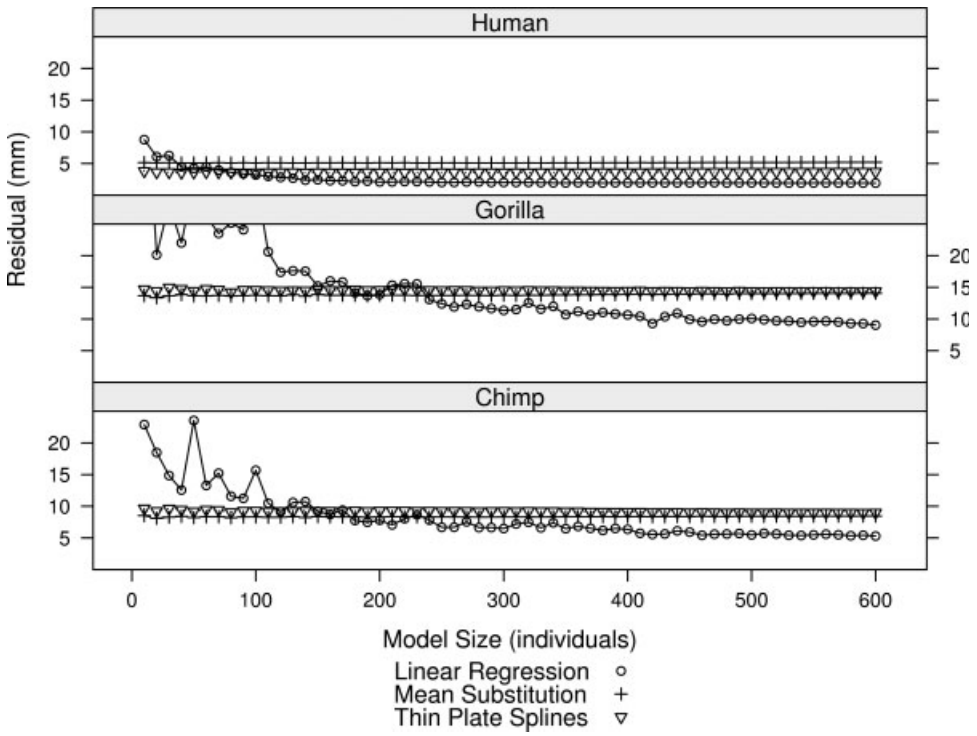


Fig. 3. A comparison between MS, TPS and RM as reference-sample sizes are increased. Groups are given for the human, chimpanzee, and gorilla test-samples. Each plot point is the sample average of the individual mean residuals for each sample. Only a human reference-sample is used, whose size increases in increments of 10 individuals.

TABLE 4. Comparing the effects of varying reference-sample sizes

	TPS				RM			
		Test species				Test species		
	Sample size	H	C	G	Sample size	H	C	G
MS	100	2.421×10^{-10} TPS	1.337×10^{-4} MS	0.126 –	100	2.307×10^{-10} RM	5.684×10^{-12} MS	3.463×10^{-13} MS
	150	3.884×10^{-10} TPS	3.154×10^{-4} MS	0.159 –	150	2.2×10^{-16} RM	0.031 MS	0.110 –
	200	2.496×10^{-10} TPS	9.500×10^{-5} MS	0.123 –	200	2.2×10^{-16} RM	0.076 –	0.760 –
	250	2.058×10^{-10} TPS	2.002×10^{-4} MS	0.139 –	250	2.2×10^{-16} RM	7.021×10^{-8} RM	0.005 –
	300	3.486×10^{-10} TPS	2.013×10^{-4} MS	0.151 –	300	2.2×10^{-16} RM	1.287×10^{-10} RM	7.733×10^{-06} RM
		Test species						
	Sample size	H	C	G				
RM	100	0.119 –	1.155×10^{-10} TPS	7.557×10^{-13} TPS				
	150	1.227×10^{-6} RM	0.879 –	0.528 –				
	200	7.581×10^{-10} RM	1.078×10^{-05} RM	0.231 –				
	250	9.783×10^{-11} RM	3.465×10^{-12} RM	1.877×10^{-4} RM				

The rows and columns give the methods being compared to each other, while each cell gives the observed P value, and the method with the lower mean if there are is a significant difference ($P < 0.05$). Tests for sample sizes below those reported here are identical to the lowest reported size, and sample sizes above are identical to the highest reported sample size.

TPS consistently produces the lower mean residuals ($P < 0.05$) when estimating human landmarks, and MS for chimpanzee landmarks ($P < 0.05$), while the converse is true using the chimpanzee reference sample. There is no significant difference between the mean residuals obtained for estimating gorilla landmarks for either method. When using the gorilla reference-sample, TPS consistently produces the lower residuals when estimating gorilla and chimpanzee landmarks, whereas for the human landmarks TPS produces the lower mean or a

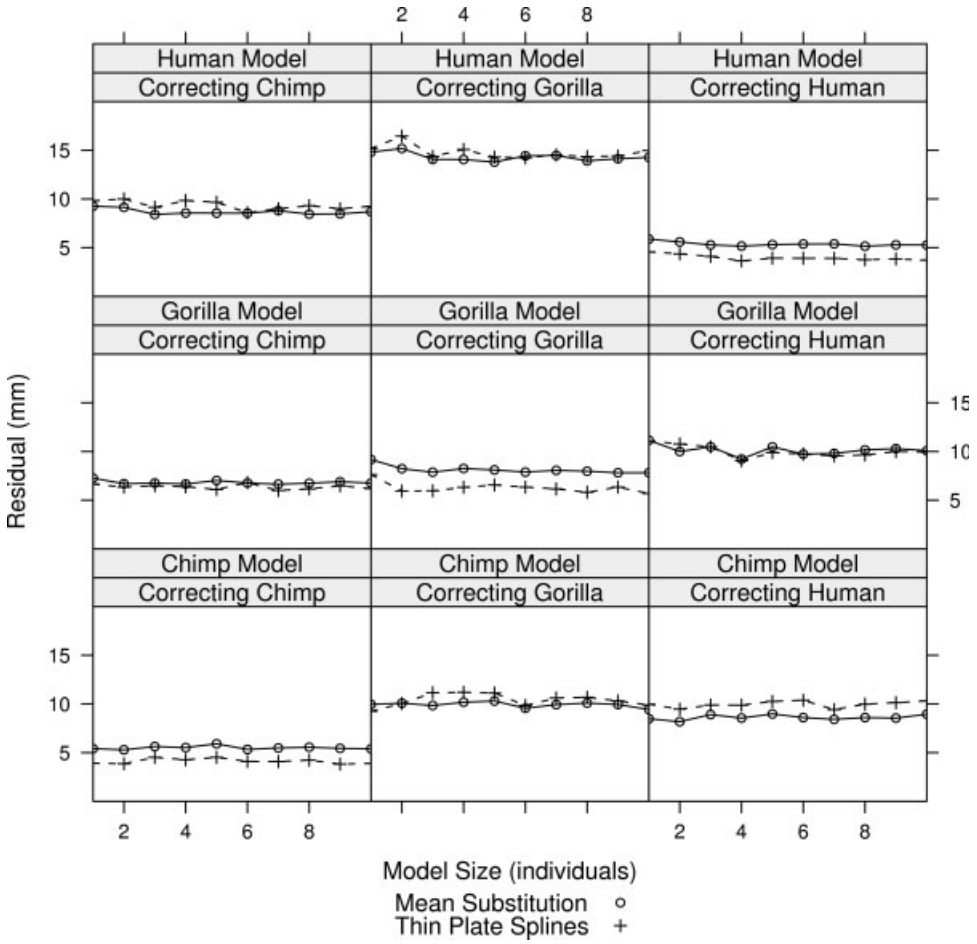


Fig. 4. Comparison of the MS and TPS methods as reference-sample sizes increase from one to 10 individuals, as in Figure 2.

TABLE 5. Means (mm) obtained for testing across-species estimation methods

	Reference species		
Method	C	G	H
Chimp. test sample			
MS	5.29	6.67	8.45
TPS	3.71	6.21	8.98
RM	–	–	5.30
Gorilla test sample			
MS	9.70	7.92	14.08
TPS	10.01	5.98	14.38
RM	–	–	9.05
Human test sample			
MS	8.44	9.89	5.19
TPS	9.69	9.85	3.69
RM	–	–	1.90

The human reference column uses the large, 600 member reference-sample. Estimations for chimpanzee and gorilla were not made using RM.

TABLE 6. Comparison of the means obtained in testing across-species reconstruction

	Reference species	
Method	C	G
Chimp. test sample		
vs. MS	0.937	$1.998 \times 10-10$
	–	RM
vs. TPS	$5.056 \times 10-11$	$1.446 \times 10-4$
	TPS	RM
Gorilla test sample		
vs. MS	0.097	0.002
	–	MS
vs. TPS	0.036	4.741×10^{-10}
	RM	TPS
Human test sample		
vs. MS	$< 2.2 \times 10-16$	$< 2.2 \times 10-16$
	RM	RM
vs. TPS	$< 2.2 \times 10-16$	$< 2.2 \times 10-16$
	RM	RM

Observed P values are reported, as is the method with the lower residual if there is a significant ($P < 0.05$) difference in the means.

mean not significantly different to that of MS. It does seem that of the two techniques, TPS produces the lower mean when performing within-species correction. To see if the large reference-sample can make viable across-species estimation using RM, Table 5 supplies the means obtained from correcting these test-samples using RM, TPS, and MS and small chimpanzee ($n = 29$), small gorilla ($n = 64$), and large human ($n = 600$) reference samples. The means are compared in Table 6.

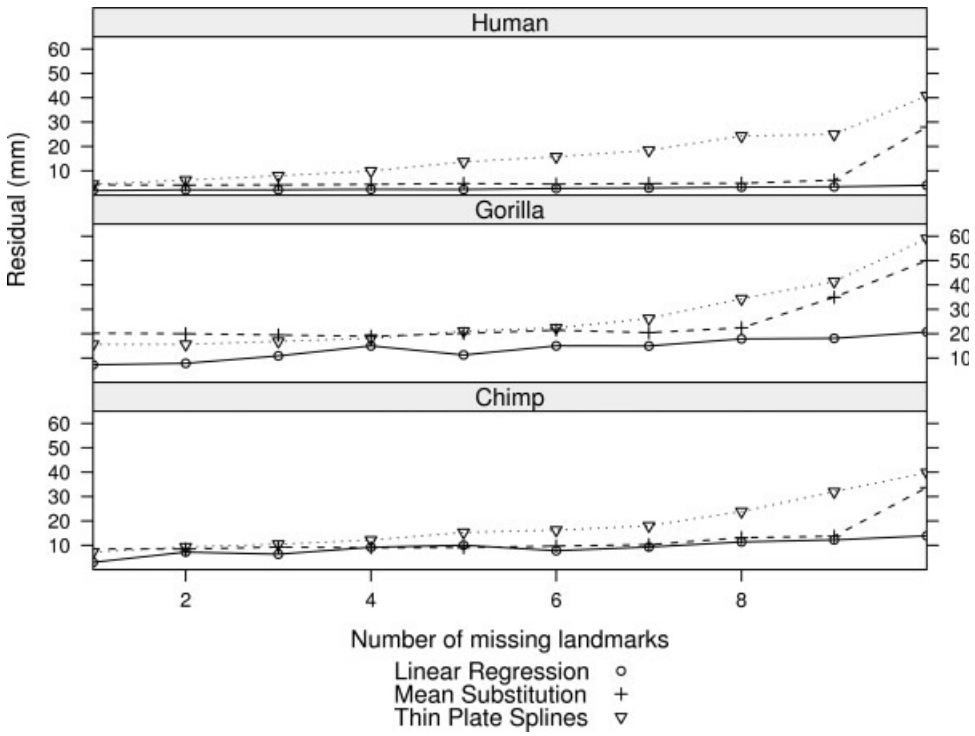


Fig. 5. Comparison of the three techniques as landmarks are cumulatively removed from the test individuals. The *x*-axis displays the number of missing landmarks, and the *y*-axis shows the obtained mean residuals.

TABLE 7. Mean residuals (mm) obtained by the estimation techniques as an increasing number of landmarks are iteratively removed

	Number of missing landmarks									
Method	1	2	3	4	5	6	7	8	9	10
Chimp. test sample										
MS	8.49	8.43	9.27	8.99	9.04	9.68	10.24	13.11	13.71	33.67
TPS	7.18	9.31	10.43	12.13	15.27	16.21	17.98	23.84	31.94	39.78
RM	2.99	7.19	6.33	9.17	10.01	7.77	9.27	11.37	12.20	13.87
Gorilla test sample										
MS	20.19	19.98	19.45	18.91	20.08	21.42	20.43	22.40	34.94	49.94
TPS	15.58	15.60	16.85	18.08	21.02	22.35	26.23	34.30	41.43	59.02
RM	7.21	7.83	10.87	14.99	11.26	15.00	14.95	17.86	18.08	20.69
Human test sample										
MS	4.30	4.15	4.25	4.43	4.78	4.54	4.76	4.90	6.12	27.95
TPS	4.43	6.20	7.96	9.92	13.68	15.75	18.44	24.27	24.97	40.80
RM	1.88	2.19	2.19	2.40	2.32	2.76	2.94	3.29	3.43	4.05

Analysis III: examining the increase in estimation errors with increasing number of missing landmarks

Figure 5 shows the results of Analysis III. Table 7 give the means obtained for each technique at each iteration of the analysis. MS produces means smaller than TPS for the chimpanzee and human test-samples ($P < 0.05$) from the third and second iterations, respectively, with no significant difference before that. For the gorilla reference sample, TPS produces smaller residuals at the second and third iterations, MS at the seventh and eight; there are no significant differences at the other iterations. RM produces significantly lower means than TPS at all iterations for all test-samples, except the fifth iteration of the gorilla test-sample, which shows no significant difference. RM produces significantly lower means than MS for all iterations of all test-samples except iterations two, four, seven and nine of chimpanzee, and eight and nine of gorilla, with these iterations showing no significant difference. MS produces the significantly lower mean for the fifth iteration of the chimpanzee test-sample.

It is interesting to consider when the mean obtained for correcting only a single landmark becomes significantly different from the later means (i.e., when does an increase in the number of damaged landmarks degrade the technique's performance). This occurs at the following points. MS: Chimpanzee test data: nine landmarks or more, although correcting seven landmarks is also significantly different; Gorilla test data: nine landmarks or more; Human test data: 10 landmarks. TPS: Chimpan-

zee test data: three landmarks or more; Gorilla test data: five landmarks or more; Human test data: two landmarks or more. RM: Chimpanzee test data: two landmarks or more; Gorilla test data: three landmarks or more; Human test data: six landmarks or more.

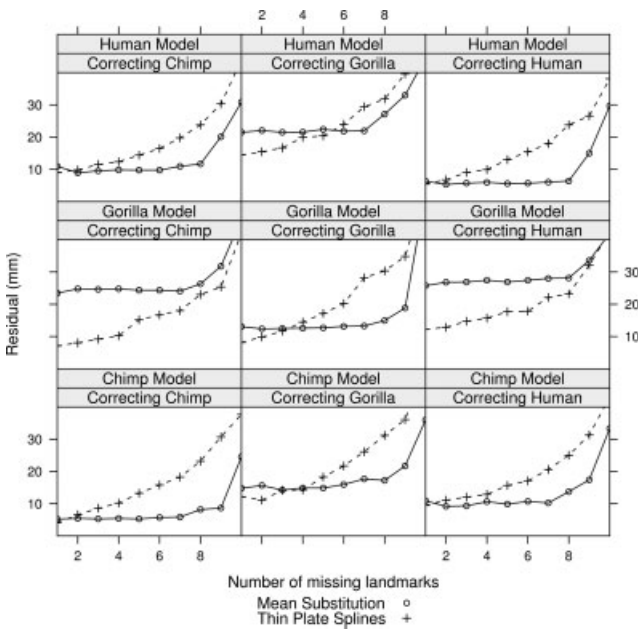


Fig. 6. Comparison of MS and TPS while landmarks are cumulatively removed. This is as seen in Figure 5, only the reference-sample consists of one randomly chosen individual.

The results for using the single reference individual are given in Figure 6 and Table 8. For the chimpanzee reference, MS produces significantly lower means from the third, fifth, and second iterations of the chimpanzee, gorilla and human test-samples, respectively. There are no significant differences before this except for the second iteration on the gorilla test, where TPS produces the lower mean. For the human reference-sample, MS produces the significantly lower mean from the third and second iterations on the chimpanzee and human test-samples, respectively. There are no significant differences before this. TPS produces significantly lower means up to and including the third iteration of gorilla, after which there is no significant difference. Using the gorilla reference-sample, TPS produces significantly lower means up to and including the seventh, second and eight iterations of chimpanzee, gorilla, and human, respectively. For chimpanzee and human, there are no significant differences after these iterations, whereas MS produces significantly lower means for gorilla from the fifth iteration onwards.

Again, we look for the number of missing landmarks at which the mean residual is significantly different from the mean obtained correcting only one landmark. MS: Chimpanzee reference-sample–Chimpanzee test data: nine landmarks or more; Gorilla test data: nine landmarks or more; Human test data: nine landmarks or more; Gorilla reference-sample–Chimpanzee test data: nine landmarks or more; Gorilla test data: 10 landmarks; Human test data: nine landmarks or more; Human reference-sample–Chimpanzee test data: nine

TABLE 8. Means residuals (mm) obtained by the estimation techniques as an increasing number of landmarks are iteratively removed, using only a single reference individual

	Number of missing landmarks									
Method	1	2	3	4	5	6	7	8	9	10
					Chimp. reference sample					
Chimp. test sample										
MS	5.23	5.45	5.28	5.43	5.28	5.69	5.84	8.18	8.64	24.54
TPS	4.08	6.55	8.59	10.19	13.23	15.79	18.23	23.18	30.71	37.69
Gorilla test sample										
MS	14.78	15.64	14.23	14.85	14.87	15.94	17.68	17.22	21.71	35.97
TPS	12.16	11.11	14.14	14.27	18.26	21.68	26.10	31.19	36.07	53.61
Human test sample										
MS	10.87	9.17	9.33	10.62	9.87	10.68	10.22	13.71	17.40	33.32
TPS	9.45	11.09	12.04	12.88	15.65	17.09	20.58	25.02	31.30	43.07
					Gorilla reference sample					
Chimp. test sample										
MS	22.36	24.80	24.64	24.78	24.36	24.35	24.08	26.31	31.80	46.51
TPS	7.08	8.03	9.18	10.24	15.17	16.69	18.03	23.02	25.37	42.50
Gorilla test sample										
MS	13.06	12.37	12.50	12.59	12.69	13.14	13.27	14.96	18.85	54.28
TPS	8.14	9.84	11.67	14.46	17.16	20.21	28.20	30.35	34.83	55.41
Human test sample										
MS	25.78	26.82	26.90	27.45	26.91	27.43	28.01	28.14	33.60	41.98
TPS	12.21	12.86	14.74	15.70	17.70	17.85	22.10	23.21	32.16	42.56
					Human reference sample					
Chimp. test sample										
MS	10.93	8.91	9.54	9.83	9.75	9.75	10.94	11.66	20.10	30.85
TPS	8.97	9.82	11.51	12.35	14.40	16.48	19.75	23.79	30.31	43.95
Gorilla test sample										
MS	21.49	22.04	21.49	21.54	22.41	21.83	21.95	27.12	32.98	44.57
TPS	14.44	15.45	16.66	19.91	20.37	23.94	29.39	31.83	39.60	54.64
Human test sample										
MS	6.31	5.32	5.69	5.96	5.58	5.68	6.04	6.38	14.99	29.67
TPS	5.42	6.76	9.03	9.93	13.01	15.43	17.99	23.79	26.56	38.10

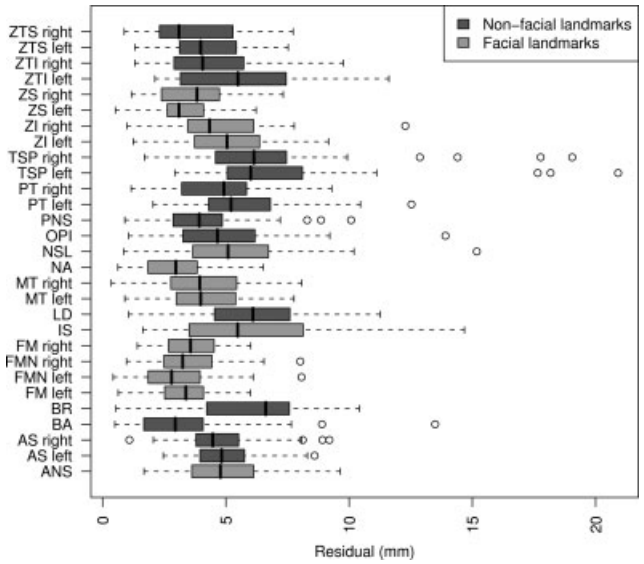


Fig. 7. A plot of the residuals obtained by each landmark for the MS method in Analysis IV.

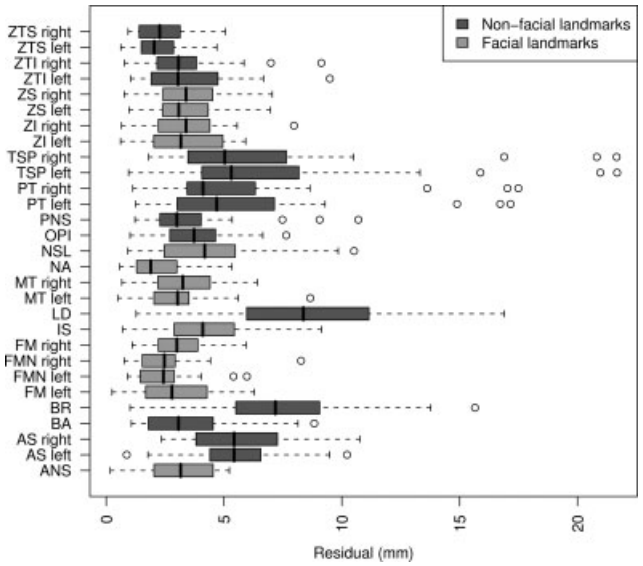


Fig. 8. A plot of the residuals obtained by each landmark for the TPS method in Analysis IV.

landmarks or more; Gorilla test data: nine landmarks or more; Human test data: nine landmarks or more. TPS: Chimpanzee reference-sample–Chimpanzee test data: two landmarks or more; Gorilla test data: five landmarks or more; Human test data: three landmarks or more; Gorilla reference-sample–Chimpanzee test data: three landmarks or more; Gorilla test data: three landmarks or more; Human test data: three landmarks or more; Human reference-sample–Chimpanzee test data: three landmarks or more; Gorilla test data: four landmarks or more; Human test data: two landmarks or more.

It is also interesting to ask if there is a significant difference obtained when using the large vs. the small reference-samples. TPS shows no significant difference between the means at any iteration of the tests; MS also shows no significant difference for the chimpanzee and gorilla tests samples, but produces significantly lower means using the large-reference sample for the human test sample on all iterations except the fifth and tenth.

Analysis IV: testing which estimated landmarks produce the largest residuals

Figures 7–9, show results of Analysis IV for MS, TPS, and RM, respectively. The mean residuals are as follows: MS $\overline{\overline{X}} = 4.7\,\text{mm}$; RM $\overline{\overline{X}} = 3.5\,\text{mm}$; and TPS $\overline{\overline{X}} = 4.1\,\text{mm}$. The landmarks can be partitioned into facial and nonfacial landmarks. MS obtains means of $\overline{\overline{X}} = 4.2\,\text{mm}$ for the facial landmarks, and $\overline{\overline{X}} = 5.2\,\text{mm}$ for nonfacial. A one sided *t*-test shows a significant difference between these means. RM obtains a mean of $\overline{\overline{X}} = 3.0\,\text{mm}$ for the facial landmarks, and $\overline{\overline{X}} = 3.9\,\text{mm}$ for the nonfacial. These means are again significantly different. The pattern is similar for TPS: the facial mean: ($\overline{\overline{X}} = 3.2\,\text{mm}$), is less than the nonfacial mean: ($\overline{\overline{X}} = 4.9\,\text{mm}$), with a significant difference between the means.

We have assumed mirroring to be a special case of MS. But, of course, in terms of biology it is worth considering whether contralateral landmarks can be used (via mirroring) in place of missing landmarks, and indeed

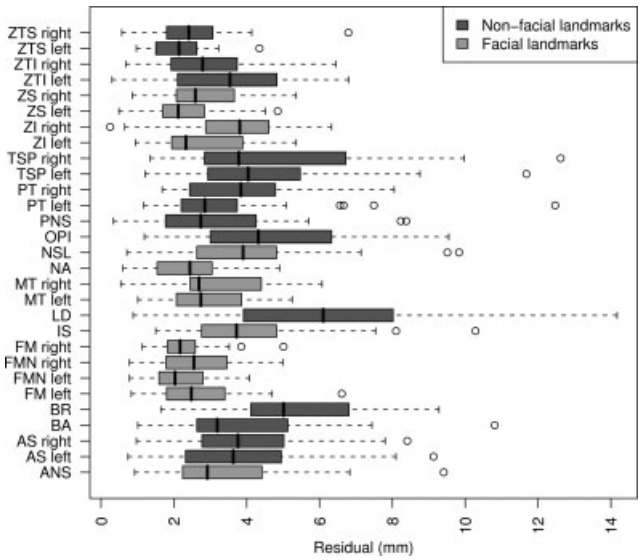


Fig. 9. A plot of the residuals obtained by each landmark for the RM method in Analysis IV.

whether this approach is superior to using TPS, MS, or RM methods of landmark estimation. To test this, we also corrected the human data set using mirroring as previously described. Over these 10 landmarks, mirroring obtained a mean residual of 4.25 mm. MS obtains a mean residual of 4.68 mm and 4.44 mm for the right side; neither of these means differ significantly from mirroring. TPS results in 3.89 mm (left) and 3.90 mm (right); again, neither differ significantly from that obtained via mirroring. RM results in 3.06 mm (left) and 3.24 (right); here, both values do differ significantly from that obtained by mirroring.

Analysis V: estimating landmarks for fossil specimens

The results of Analysis V are presented in Table 1. Averaged over all three reference-samples, MS produces

TABLE 9. Weighted means (mm) obtained for correcting fossil specimens

	MS			TPS			RM		
Genus	C	G	H	C	G	H	C	G	H
Australopith	11.21	12.37	12.63	13.28	13.22	13.21	19.84	13.23	9.09
	11.79		12.63	13.25		13.21	16.54		9.09
Homo	10.30	11.74	10.02	13.99	14.85	8.68	12.55	9.47	6.66
	11.02		10.02	14.42		8.68	11.01		6.66

The columns represent the species (chimpanzee, gorilla, human) and estimation method; the rows group the individuals by genus, the australopith group being made up of *Paranthropus* and *Australopithecus*.

a total weighted mean of 11.50 mm, TPS a mean of 12.94 mm, and RM of 12.21 mm. There are no significant differences between these means. However, RM, when using the human reference-sample of $n = 628$ individuals, produces a mean residual of only 8.09 mm, far below any of the means of the other techniques ($P < 0.05$).

Test specimens may be grouped together with reference-samples based on morphological similarity: australopiths with chimpanzees and gorillas, *Homo* with humans. It might be expected that landmark estimations produced from these morphologically similar reference-samples would produce smaller residuals, and so Table 9 supplies mean residuals for australopith and *Homo* reconstructions calculated using each extant genus, and also presents the average of the chimpanzee + gorilla mean residuals. The australopith group shows no difference between the residuals obtained for MS and TPS as the reference group varies between chimpanzee/gorilla and human ($P > 0.05$). RM produces a significantly smaller mean when using the human reference-sample (which has the larger sample). For the *Homo* group, MS shows no significant difference between the chimpanzee/gorilla and human reference-samples. Both TPS and RM produce smaller residuals with the human reference-sample. The mean residual for correcting *Homo* individuals using RM and the human reference-sample is only 6.66 mm.

Landmark estimation is also performed using a single fossil specimen as the reference individual. Only MS and TPS are used. Table 10 presents the results obtained using STS 5, KNM-ER 406, and KNM-ER 3733 as the reference individuals. TPS and MS do not obtain means significantly different from one another. It is notable that RM, when using the large human reference-sample, produces a smaller mean residual than either MS or TPS when these methods use a single fossil specimen as a reference-sample ($P < 0.05$). Further, there is no significant difference in using MS with either a living or extinct reference-sample, as with TPS.

The results of an RM reconstruction of the fossils based on pooled extant references species are presented in Table 11. Both heterogeneous reference-samples obtain total means (weighted by landmark count) of over 10 mm: 10.57 mm for the equal sized sample, and 10.23 mm for the complete group. These two means are not significantly different from each other. Of interest, the mean obtained for the complete heterogeneous group is significantly larger than that obtained using just the 628 human individual reference-sample. Neither the equal nor the complete heterogeneous groups are significantly different to the TPS or MS total means, although the complete group is close, with an observed *P*-value of 0.054.

TABLE 10. Means (mm) obtained for estimating landmarks on fossil specimens using a single reference individual

Specimen	No. landmarks	MS	TPS
STS 5 Reference individual			
STS 5	–	–	–
KNM-ER 1470	7	15.24	23.25
KNM-ER 1813	8	9.77	12.51
KNM-ER 3733	11	12.21	12.58
KNM-ER 406	12	11.43	14.92
KNM-WT 17400	7	11.05	32.51
Taung	**11**	**5.44**	**4.99**
KNM-WT 15000	10	11.35	16.16
KNM-WT 17000	10	14.10	15.24
Total Means		11.16	15.48
KNM-ER 406 Reference individual			
STS 5	12	9.60	11.03
KNM-ER	7	16.01	20.76
KNM-ER 1813	8	8.06	10.52
KNM-ER 3733	12	13.64	14.08
KNM-ER 406	–	–	–
KNM-WT 17400	**7**	**10.74**	**16.85**
Taung	11	7.98	9.51
KNM-WT 15000	10	10.85	13.00
KNM-WT 17000	**10**	**13.67**	**16.28**
Total Means		11.21	13.58
KNM-ER 3733 Reference individual			
STS 5	11	11.25	12.07
KNM-ER 1470	**6**	**14.75**	**13.72**
KNM-ER 1813	**7**	**9.69**	**13.62**
KNM-ER 3733	–	–	–
KNM-ER 406	12	15.15	15.70
KNM-WT 17400	7	9.62	12.81
Taung	10	5.42	5.78
KNM-WT 15000	**9**	**9.42**	**7.76**
KNM-WT 17000	10	17.70	17.53
Total Means		12.04	12.85

The total means are weighted means, whereas the bolded rows are individuals of the same species as the reference individual.

DISCUSSION

The analyses

Analysis I: estimating landmarks using MS, TPS, and RM. Analysis I shows a standard and expected pattern: within-species estimation outperforms across-species estimation. Of note, the smallest residual is obtained for within-species estimation of humans using RM, which gives a residual just larger than 2 mm—a magnitude of error comparable to accepted interuser error. In fact, RM performs the best for all within-species estimates, although TPS appears no different to RM for all but the largest reference-sample. However, in across-species comparison, RM only outperforms the other techniques when using the 280-individual, human reference-sample. TPS and MS estimations are statistically

TABLE 11. Results from estimating landmarks on fossil specimens using the regression method and heterogeneous reference-samples

Specimen	Equal sized heterogeneous RM	Complete heterogeneous RM
STS 5	9.27	9.13
KNM-ER 1470	10.08	9.61
KNM-ER 1813	7.89	7.15
KNM-ER 3733	8.83	9.28
KNM-ER 406	18.02	14.15
KNM-WT 17400	7.91	10.76
Taung	5.20	5.32
KNM-WT 15000	8.44	9.04
KNM-WT 17000	16.90	11.72
Total Means	10.57	10.23

Mean residuals (mm) are listed for using heterogeneous reference-samples composed of 67 chimpanzee, 67 gorilla, and 67 human individuals for the equal sized group, and 67 chimpanzee, 117 gorilla, and 628 human for the complete group. Total means are weighted by the number of landmarks in the individual.

equivalent across-species, while TPS outperforms MS within-species. Both methods appear invariant to increases in reference-sample sizes (no significant difference between means as the reference-sample sizes increase), which is supported by later analyses. Conversely, the results suggest that small reference-samples play a large role in RM's poor performance.

Analysis II: estimating the point at which RM outperforms MS and TPS. To interrogate the degree to which sample-size affects RM's performance, Analysis II determines the point, relative to reference-sample size, at which RM outperforms the other techniques. This analysis highlights RM's need for large reference-samples, and indeed with large samples RM outperforms both MS and TPS. It also demonstrates the existence of an asymptote: there is a point after which increased reference-sample sizes give negligible returns in residual reduction. As with the first analysis, both MS and TPS are relatively invariant to changes in reference-sample sizes. RM outperforms MS from approximately 50 reference individuals onwards; from between 50 and 100 individuals it also outperform TPS, with significant differences between means at these points, and onwards. RM requires reference-samples of some few hundred individuals to effectively estimate landmark positions from across-species reference-samples. From between 250 and 300 individuals, RM outperforms MS on both the chimpanzee and gorilla test-samples. Fewer individuals are required to outperform TPS: from between 200 and 250 individuals onwards. The results obtained for TPS and MS using small reference-samples are similar to those obtained using the larger samples. As previously seen, TPS appears to outperform MS for within-species estimation.

To test the across-species estimation power of RM, we compared the means obtained by correcting the chimpanzee and gorilla test-samples using RM (and the 600 individual, human reference-sample), and those obtained correcting test-samples with MS and TPS (driven by within-species reference-samples). An across-species application of RM is unable to outperform within-species MS and TPS, although it is of note that RM shows no significant difference to MS when correcting the chimpanzee test-sample. When all three techniques use an across-species reference-sample, however, RM produces the smaller residuals in all but one case: against MS correcting the gorilla. However, in this case the obtained means are not significantly different. This suggests that if even a small, within-species reference-sample is available, TPS should be the method of choice. However, if no such sample is available (or perhaps even if the species of the damaged individual is uncertain), RM using a large, across-species reference-sample should be used.

Analysis III: examining the increase in estimation errors with increasing number of missing landmarks. Analysis III considers the amount of damage requiring reconstruction and, indirectly, the amount of information from the damaged individual available to the estimation technique. Both MS and RM show a low rate of residual increase relative to TPS. Significance tests between the means show that TPS produces the largest residuals as the number of missing landmarks increases. From four landmarks upwards, even MS produces means significantly lower, or of no significant difference, to those of TPS. RM appears identical to TPS when estimating a single landmark, and produces the lower mean from two upwards. When compared with MS, RM produces a lower mean from the first missing landmark onwards. Notably, MS appears fairly invariant to the number of missing landmarks (or lack of morphological information), producing a significantly different mean only after a large number of landmarks are missing (nine or 10, roughly 70–75% of the landmarks). Both TPS and RM do not react as well to the number of missing landmarks, probably due to both techniques requiring undamaged morphology to guide the estimation, either as predictor variables or for defining the splines. Both RM and TPS produce significantly larger residuals from two missing landmarks onwards.

Notably, TPS residuals grow far more rapidly than those of the other techniques. This may be due to a property of TPS themselves: once defined, the deformation applied by such a spline does not gradually approach zero the further one moves from the subset of the spline's domain used in its construction: instead the spline becomes increasingly "flat" (constant), but not necessarily zero, as one moves away from the defining subset (Bookstein, 1991). Hence, the further one moves from any of the landmarks used in its construction, the less the spline reflects the required warp in the damaged area. As more landmarks require estimation, the distance between missing and non-missing landmarks increases, exacerbating this effect. Clumping of correct landmarks (such as is seen in fossil hominins where only a portion of the cranium exists), or a large area with few landmarks, are both situations in which TPS is not ideal. The TPS pattern of a rapid residual increase remains true for small reference-samples. Notice that, as with previous analyses, TPS produces smaller residuals with the gorilla reference-sample, but this appears to be the only case in which TPS would be chosen over, say, MS.

Analysis IV: testing which estimated landmarks produce the largest residuals. Analysis IV demonstrates that for all three techniques the nonfacial landmarks suffer from the most estimation error, with facial landmarks giving residuals smaller on average by between 1 mm and 2 mm, or 20% for MS, 35% for TPS, and 24% for RM. Although the reason for this is not clear, it does suggest that the various techniques

perform better reconstructing closely spaced landmarks, such as the mostly facial landmark set used in the other analyses. We have already seen a suggestion of this relating to TPS. However, this correlation between sparse landmarks and increased residual size is not well tested, and further work is required to demonstrate that such a relationship indeed exists.

As part of Analysis IV, we also examined mirroring's relative performance. From an anatomical standpoint, mirroring might be expected to result in reduced residuals for bilateral landmarks, but our results suggest that this is not the case. Indeed, landmark estimation through mirroring provides errors comparable to estimates from both TPS and MS. RM, however, appears to estimate these landmarks with more precision than mirroring.

Analysis V: estimating landmarks for fossil specimens. Analysis V estimates landmarks for fossil specimens rather than extant species. Although there are no significant differences between the means calculated across all reference-samples for each method, RM using the large, human reference-sample outperforms the other methods ($P < 0.05$), producing a residual smaller by between 2 mm and 8 mm. This is in line with the results of Analysis II concerning across-species estimates. Because of the overall morphological similarity, one might expect the chimpanzee or perhaps the gorilla reference-samples to produce smaller residuals than the human reference-sample when correcting the australopith specimens, but we see instead that RM, using the large human reference-sample, produces smaller residuals than the other reconstruction method/reference-sample combinations. This implies that reference-sample size, and presumably the associated robust model of variance/covariance structure (and the ability to exploit such a model with an estimation technique such as RM), may be more important than morphological similarity, as will be discussed further below.

The fossil specimens were also corrected using a case typically seen in the literature: using a single reference individual of an extinct species (such as in Ponce de León and Zollikofer, 1999; Neubauer et al., 2004). Although the Taung child generally reacts well to the reconstruction using these samples (it is not clear as to why—although Taung is the smallest and youngest of the specimens), in general, the across-species estimation via RM and the large, human reference-sample outperforms these techniques, producing a mean smaller by between 2 mm and 7 mm. Across-species RM is also the only technique/reference-sample combination to produce mean residuals below 10 mm. Using an across-species RM with a reference-sample drawn from a living species appears significantly (and greatly) better than using a single reference individual (or even a sample) from an extinct species when the reference individual's species differs from that of the damaged individual. Again, this implies that sample size is more important than morphological closeness.

Finally, we also analyzed the performance of RM using pooled, heterogeneous reference samples, under the assumption that such a sample might better represent the shared, ancestral population of covariation in early hominins (e.g., Zollikofer and Ponce de Leon, 2005). These analyses indicate that mixed models produce worse results than those obtained using a homogenous reference-sample. With no significant difference between either of the heterogeneous groups and MS/TPS, it appears that heterogeneous groups should not be used over homogenous groups. It is not clear why this is so, but we may speculate: a more homogenous sample should have a tighter pattern of covariation, and therefore be a more accurate data set from which to predict, thereby providing smaller residuals. A similar effect might explain the poorer results typically obtained when estimating landmarks using gorilla data sets, as greater sexual dimorphism and overall variation in this extant taxon could result in poorer landmark estimation power.

The importance of biological relevance

All the previous observations can be summed up as follows: when performing within-species estimation of a single landmark using reference-samples of large enough size, RM produces the smallest residuals, TPS slightly larger residuals, and MS larger yet. However, as the number of missing landmarks increases, TPS quickly produces errors larger than MS (in the results presented here, from 30% and upwards of missing landmarks). When performing across-species estimation, TPS and MS appear to produce similar sized residuals, whereas RM outperforms both.

From a biological standpoint, there are many assumptions behind each of the reconstruction techniques that may offer some insight into the relative performance of each, and especially the good performance of RM. Mean substitution, which replaces missing morphology with the morphology of another individual, assumes that individuals are similar enough that their anatomy is interchangeable, either with that of a single individual, or of an average over a sample of individuals. Composite physical reconstructions and mirroring techniques are special cases of mean substitution. Apart from mirroring (which cannot always be applied), mean substitution methods are not based on the known anatomy of a damaged individual, and hence what this anatomy could imply about the missing portions. The thin plate spline method is an extension to the mean substitution technique, but recognizes that *ad hoc* substitution of mean morphology may not be sufficient, and instead aims to fit the expected morphology to the damaged specimen's known morphology in an attempt to reduce estimation error. Although the thin plate spline interpolation method is widely used to perform this fitting, many other such methods exist, some of which may prove better for this purpose than TPS, although this has not been well-evaluated in the literature. Importantly, thin plate splines mimic how thin sheets of metal deform under pressure, and as such also have limited biological relevance. In contrast, regression-based methods have inherently biological underpinnings, as they assume that there is a pattern to how morphology varies in relation to other anatomical areas. They use known morphology on the damaged specimen to drive the estimation of the missing regions, along with variation/covariation information from undamaged specimens. This method has stronger ties to the damaged specimen's (and the model specimens') biology than either of the other methods, both of which have little biological meaning.

Viewed in this light, it is perhaps not surprising that RM performs so well. What is more surprising is that MS performs as well as it does: MS is capable of estimating multiple landmarks with little increase in error. This is biologically unintuitive, and the only comparable—

albeit unpublished—study suggests that MS is a uniformly poor performer (Gunz, 2005). The rapid increase in TPS residual sizes is also both surprising and worrying, it appears that the method may not be appropriate when reconstructing large areas of damaged morphology, or when used with few known landmarks in the undamaged individual.

The unimportance of morphological distance

Zollikofer and Ponce de León (2005, pg 179) give the advice that fossil reconstruction should use reference-samples drawn from the same species as the specimen being reconstructed, and that the reference-sample should "represent the shared ancestral pattern of variation rather than patterns of variation characteristic of the derived taxa". This is in order to avoid biasing the reconstructions towards "preconceived morphologies." This approach is ideal, but difficult to achieve because of extremely small samples of fossil material available for the various hominin species. Indeed, samples are often so small that robust statistical inference (and, indeed, robust inferences in general) and reconstruction become difficult (Smith, 2005). This means that we may not be able to a) draw a reference-sample from the same population to which the damaged individual belongs, or b) draw a reference-sample from an ancestral population, or even c) create a large enough sample of related individuals, whether they share the derived form, the ancestral form, or have autapomorphies of their own. Of particular note is the result in Analysis V showing that our heterogeneous samples fail to capture these shared ancestral patterns—in other words, pooling extant species data does not bring us closer to an estimate of ancestral populations.

Viewed within the context of previous studies that demonstrate shared patterns of variation/covariation in taxa within the primate order (Ackermann and Cheverud, 2000, 2002; Ackermann, 2002, 2005), our results suggest a way forward: use extant species as the reference-sample, as estimates derived from these shared patterns of covariation allow for a more accurate reconstruction of the relative relationships among landmarks. This approach has its own methodological flaws: there are differences in variance and covariance patterns among species, and the consequences of such differences are demonstrated by the results presented here. However, reservations of biasing the reconstruction aside, using large samples drawn from a living species in an across-species estimation of landmarks via RM appears to be a much better choice than an across-species estimation using small samples of extinct species (typically using TPS). This is likely due to the shared, and presumably evolutionarily conserved, pattern of covariation seen across the primate order (e.g., Cheverud, 1996; Ackermann and Cheverud, 2000, 2004; Marroig and Cheverud, 2001; Ackermann, 2002, 2003, 2005; Gonzalez-José et al., 2004) and is also shared by our hominin ancestors. In other words, morphological similarity appears to be of less importance than large reference-samples, and the ability to leverage that knowledge.

Stated differently, our work strongly suggests that reconstruction methods using a single specimen of a species other than that of the damaged specimen—but perhaps closely allied to it—to approximate missing morphology, such as by modeling missing cranial portions of *Homo habilis* based on the morphology of *Homo erectus* using TPS, is less accurate than using a large sample, possibly drawn from an extant (and hence not as closely related) population, such as *H. sapiens*, and a method (such as a regression-based method) able to take advantage of the information contained in this sample. In other words, using a small sample (or even a single individual) of a closely related species to guide fossil reconstruction or landmark estimation (as is frequently done) is a poor choice over using a larger, more morphologically robust model and reconstruction method, even if the model is that of a more distantly related species.

Methodological problems

There are undoubtedly a number of problems and biases associated with the approaches used here. For example, Analysis III randomly chooses missing landmarks, implicitly assuming that the probability of a specimen missing a given landmark is independent of it missing any other. However, taphonomic distortion or damage tends to affect whole, connected areas: it is more likely that missing landmarks will lie closer together than be randomly scattered over the specimen. Because a random scatter should be more easily corrected, it is likely that the calculated residuals underestimate those that would actually be obtained. However, as the number of damaged landmarks increases, randomized landmark loss begins to approximate true taphonomic distortion. Additionally, studying a technique's obtained errors in relation to the number of missing landmarks is perhaps deceptive. Most of the techniques (except MS) have a reliance on the distance to the closest nonmissing landmark, and this is arguably a stronger determinant of accuracy than the number of missing landmarks, and more than likely a confounding factor in this study. Also, although one would like to hypothesize that these distance-effects resulted in the greater error associated with the neurocranial landmarks, MS also displays increased errors with reconstructed neurocranial landmarks even though it should not. It is not clear why this is so. Finally, this analysis was constructed based on a fairly restricted set of landmarks, extracted from a previously-collected data set because they were shared among fossil specimens. We did not test whether different combinations of landmarks would give different results, nor can we know if very different data sets would result in the same conclusions. Further work needs to be done to better understand the universality of these results for different landmark data and other organisms, especially as regression-based methods are promising but largely untested. Yet despite these caveats, the results do provide some tentative guidelines for how to choose methods and appropriate samples when attempting to reconstruct the missing morphology of fossil hominins, as outlined below.

CONCLUDING RECOMMENDATIONS

On choosing a reference-sample

When possible, reference-samples should be drawn from the same species as that of the individual being reconstructed, which results in landmark estimates with lower mean residuals. Small, within-species reference-samples often prove adequate to drive MS and TPS, especially if the amount of damage to be corrected is small. However, we know that small samples (especially "samples" of one individual) can unduly affect the recon-

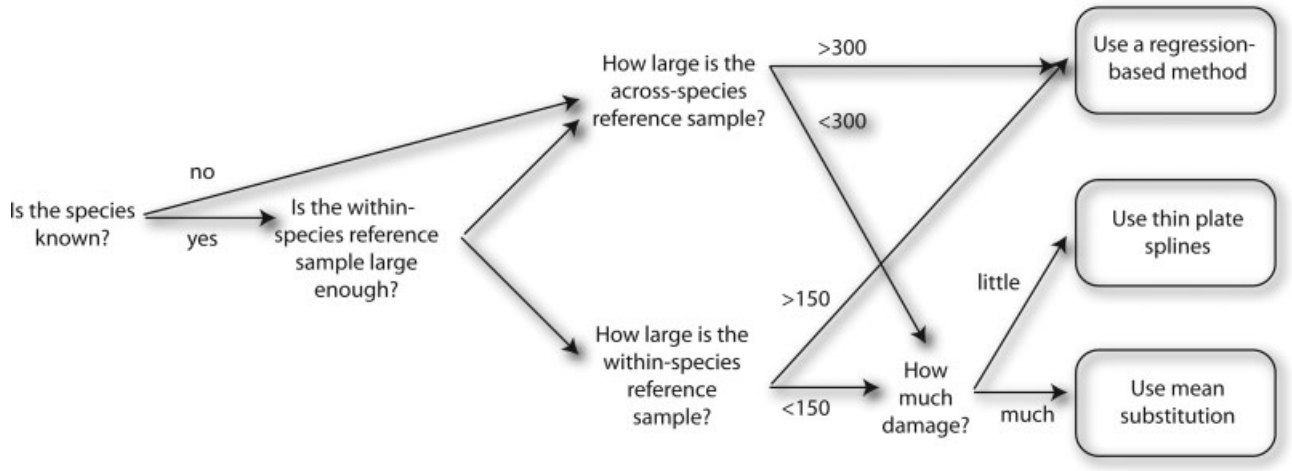


Fig. 10. A flowchart summarizing our concluding recommendations on selecting a landmark estimation technique and reference-samples.

struction, and while they may prove adequate, they should generally be avoided if possible. Specifically, if the reference-sample is drawn from well known hypodigms such as *Homo sapiens*, *Homo neanderthalensis*, or *Homo erectus*, there is no reason to use only one reference individual.

If the damaged individual's species is unknown or uncertain, we recommend using a large across-species reference-sample driven by a technique such as RM. In this case, reference-samples of a few hundred individuals drawn from an extant species prove satisfactory.

Reference-samples drawn from a species showing great intraspecific variability (e.g., *Gorilla gorilla*) should be avoided: they appear to provide larger residuals than those obtained otherwise. If such a species must be used, TPS appears to be a good estimation technique to use.

On choosing the technique

Figure 10 is a flow chart illustrating the following recommendations. The most clear cut recommendation involves correcting individuals of a species for which a large reference-sample can be drawn. Regression methods are, in this case, superior to other methods. However, the following guidelines seem appropriate when using smaller-reference samples.

Researchers should first ask themselves this question concerning the reconstruction: Is the species of the damaged individual known or unknown?

- For a known species, use a within-species reference sample.
  a. If it is not possible to create a within-species reference-sample, proceed as if the individual's species were unknown.
  b. If the reference-sample has 150 individuals or more, use RM.
  c. For small reference-samples, proceed as follows: Is there much damage (or few landmarks on the damaged individual)? If yes, use MS unless the reference-sample is drawn from a species with large, intraspecific variability, in which case TPS should be used. If there is little damage, use TPS.
- For an unknown species, use across-species estimation.
  a. Where large reference samples can be compiled ($N > 300$) from an extant species that does not show great intraspecific variability (such as that shown by gorillas) is available, use an estimation method such as RM. Because of the ease with which landmark data for extant species may be obtained, there is little reason not to use such an estimation regime.
  b. If the reference-sample is smaller than roughly 250–300 individuals, proceed as follows: Is there much damage (or few known landmarks on the damaged individual)? If yes, use MS unless the reference-sample is drawn from a species with large intraspecific variability, in which case use TPS. With little damage, use TPS.

ACKNOWLEDGMENTS

We would like to thank the curators of the various museums and institutions from which the original data were collected for generously proving access to materials. Ongoing work on data sets such as these could not occur without their generous support. We would like to thank the reviewers and editors for their helpful comments.

LITERATURE CITED

Ackermann RR. 1998. A quantitative assessment of variability in the australopithecine, human, chimpanzee, and gorilla face [dissertation]. St. Louis: Washington University.

Ackermann RR. 2002. Patterns of covariation in the hominoid craniofacial skeleton: implications for paleoanthropological models. J Hum Evol 43:167–187.

Ackermann RR. 2003. Morphological integration in hominoids: a tool for understanding human evolution. Am J Phys Anthropol Supp 36:55.

Ackermann RR. 2005. Ontogenetic integration in the honinoid face. J Hum Evol 48:175–197.

Ackermann RR, Cheverud JM. 2000. Phenotypic covariance structure in tamarins (genus *Saguinus*): a comparison of vari-

ation patterns using matrix correlation and common principal component analysis. Am J Phys Anthropol 111:489–501.
Ackermann RR, Cheverud JM. 2002. Discerning evolutionary processes in patterns of tamarin (genus *Saguinus*) craniofacial variation. Am J Phys Anthropol 117:260–271.
Ackermann RR, Cheverud JM. 2004. Morphological integration in primate evolution. In: Pigliucci M, Preston K, editors. Phynotypic integration: studying the ecology and evolution of complex phenotypes. Oxford: Oxford University Press. p 302–319.
Bookstein FL. 1991. Morphometric tools for landmark data: geometry and biology. New York: Cambridge University Press.
Cheverud JM. 1996. Quantitative genetic analysis of cranial morphology in the cotton-top (*Saguinus oedipus*) and saddleback (*S. fuscicollis*) tamarins. J Evol Biol 88:5–42.
Conroy GC, Falk D, Guyer J, Weber GW, Seidler H, Recheis W. 2000. Endocranial capacity in Sts 71 (*Australopithecus africanus*) by three-dimensional computed tomography. Anat Rec 258:391–396.
Conroy GC, Weber GW, Seidler H, Tobias PV, Kane A, Brunsden B. 1998. Endocranial capacity in an early hominid cranium from Sterkfontein, South Africa. Science 280:1730–1731.
Cox TF, Cox MAA. 1994. Multidimensional scaling. London: Chapman & Hall.
González-José R, Van Der Molen S, González-Pérez S, Hernéndez M. 2004. Patterns of phenotypic covariation and correlation in modern humans as viewed from morphological integration. Am J Phys Anthropol 123:69–77.
Gunz P. 2005. Statistical & geometric reconstruction of hominid crania: reconstructing australopithecine ontogeny [dissertation]. Wien: Universitat Wien.
Jolliffe IT. 1986. Principal component analysis. New york: Springer-Verlag.
Marroig G, Cheverud JM. 2001. A comparison of phenotypic variation and covariation patterns and the role of phylogeny, ecology, and ontogeny during cranial evolution of new world monkeys. Evolution 55:2576–6000.
Neeser R. 2007. A comparison of statistical and geometric reconstruction techniques: guidelines for correcting fossil hominin crania [dissertation]. Cape Town: University of Cape Town. Available at: http://pubs.cs.uct.ac.za/archive/00000413/.
Neubauer S, Gunz P, Mitteroecker P, Weber GW. 2004. Three-dimensional digital imaging of the partial *Australopithecus africanus* endocranium MLD 37/38. Can Assoc Radiol J 55:271–278.
Ponce de León MS, Zollikofer CPE. 1999. New evidence from le moustier 1: computer-assisted reconstruction and morphometry of the skull. Anat Rec 254:474–489.
R Development Core Team. 2005. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.
Richtsmeier JT, Cheverud JM, Subhash L. 1992. Advances in anthropological morphometrics. Annu Rev Anthropol 21:283–305.
Smith RJ. 2005. Species recognition in paleoanthropology: implications of small sample sizes. In: Lieberman D, Smith R, Kelley J, editors. Interpreting the past: essays on human, primate and mammal evolution in honor of David Pilbeam. Boston: Brill Academic Publishers. p 207–219.
Sokal RR, Rohlf FJ. 1995. Biometry, 2nd ed. New York: Freeman.
Weber GW. 2001. Virtual anthropology (VA): a call for glasnost in paleoanthropology. Anat Rec 265:193–201.
Zollikofer CPE, Ponce de León MS. 1998. Computer-assisted paleoanthropology. Evol Anthropol 6:41–54.
Zollikofer CPE, Ponce de León MS. 2005. Virtual reconstruction: A primer in computer-assisted paleontology and biomedicine. Hoboken, New Jersey: Wiley.
Zollikofer CPE, Ponce de León MS, Esteves F, Silva T, Pacheco D. 2002. The computer-assisted reconstruction of the skull. In: Zilhão J, Trinkaus E, editors. Portrait of the artist as a child. The Gravettian Human Skeleton from the Abrigo do Lagar Velho and its Archeological context. Lisbon: Instituto Portugeuês de Arqueologia. p 326–341
Zollikofer CPE, Ponce de León MS, Lieberman DE, Guy F, Pilbeam D, Likius A, Mackaye HT, Vignaud P, Brunet M. 2005. Virtual cranial reconstruction of *Sahelanthropus tchadensis*. Nature 434:755–759.