Forensic Anthropology Vol. 2, No. 3: 178–188
DOI: 10.5744/fa.2019.1015

TECHNICAL NOTE

Missing Data Imputation Using Morphoscopic Traits and Their Performance in the Estimation of Ancestry

Michael Kenyhercz[a,b*] • Nicholas Vere Passalacqua[c] • Joseph T. Hefner[d]

ABSTRACT: Missing data are an inherent problem in biological anthropology for both reference data sets and individual cases. The goal of data imputation for forensic anthropological applications is to accurately estimate missing values by using other, observed values. To quantify the accuracy of macromorphoscopic data in conditions with slight (10%), moderate (25%), and severe (50%, 75%, and 90%) amounts of missing data, we selected four data-imputation techniques: Hot Deck, iterative robust model-based imputation (IRMI), *k*-nearest neighbor (*k*-NN), and the variable medians. Hefner's Macromorphoscopic Databank was used (Hefner 2018); the full sample consisted of 688 individuals from 3 U.S. populations (Blacks, Hispanics, and Whites). Six cranial macromorphoscopic variants were scored in accordance with Hefner (2009). The five data sets with missing data were randomly simulated over multiple iterations ($N = 500$ each) from the original data. These data sets were compared for agreement using weighted Cohen's kappa and correct classification accuracies over multiple iterations ($N = 500$) calculated for the original data set. The latter comparisons were also used to examine the effects of imputed data on classification accuracies. Results suggest that IRMI is the most accurate method for imputing missing data, followed by *k*-NN, in each of the comparisons for nearly all of the variables imputed.

KEYWORDS: forensic anthropology, missing data imputation, macromorphoscopic traits, nonmetric data, classification

Missing data are an inherent problem in biological anthropology due to the fragile nature of osseous material; these issues are compounded in forensic anthropology, as remains in forensic contexts are often subjected to perimortem trauma and/or taphonomic alterations that damage or destroy bony morphology. Missing data are often an analytical issue in which certain statistics cannot accommodate missing data values. A common way to circumvent the issue of missing data is to omit any case in which at least one variable is missing, also known as listwise deletion. In other circumstances, variables that have higher amounts of missing data may be removed from an analysis entirely. Thus, missing data can reduce sample sizes or limit the number of variables used within a given analysis, which has the potential to significantly decrease reliable estimates of some aspects of the biological profile (Passalacqua et al. 2013).

[a]Department of Defense POW/MIA Accounting Agency, Central Identification Laboratory, Joint Base Pearl Harbor–Hickam, HI 96853, USA
[b]University of Pretoria, Anatomy, Pretoria, South Africa
[c]Anthropology and Sociology, Western Carolina University, Cullowhee, NC 28723, USA
[d]Department of Anthropology, Michigan State University, East Lansing, MI 48824, USA
*Correspondence to: Michael Kenyhercz, Department of Defense POW/MIA Accounting Agency, Central Identification Laboratory, 590 Moffet St., Bldg. 4077, Joint Base Pearl Harbor–Hickam, HI 96853, USA
E-mail: michael.kenyhercz@gmail.com

Received 25 February 2017; Revised 14 September 2018; Accepted 15 November 2018

Missing data are an issue for comparative, reference data sets and also individual forensic cases. Given the typical handling of missing data mentioned above, missing data within reference data sets can limit sample sizes, limit the number of variables to compare, or both, which can affect the strength of the analysis and any conclusions based upon the results. Conversely, when faced with a fragmentary cranium, a full suite of macromorphoscopic traits may not be present for analysis, limiting method applicability. As Hefner (2009) has pointed out, individual macromorphoscopic traits are not particular to one ancestry group, but the expression of several traits in tandem can be used with an appropriate classification statistic to accurately classify individuals. Many classification statistics and data-reduction techniques (e.g., principal components analysis, principal coordinate analysis) require complete data sets, which makes missing data an issue for forensic casework and also research at large.

Cranial macromorphoscopic traits have a long history in forensic anthropology for the estimation of ancestry (Hefner et al. 2012; Rhine 1990). Unfortunately, the manner in which these traits have been used to estimate ancestry is fraught with analytical bias, subjectivity, and a lack of statistical rigor. Early on, Hooton (1946) recognized the subjectivity inherent in trait scoring, noting that "even veteran anthropologists have difficulty in maintaining consistency in these subjective ratings and still greater difficulty in equating their standards with those of equally experienced observers" (1946:715). His "Harvard Blanks" were designed to reduce

some of that subjectivity, and in many ways they formalized the collection of cranial nonmetric data. For most of its existence, forensic anthropology, regrettably, did not move very far beyond the list of traits proffered by Hooton. In fact, many forensic anthropologists maintained and championed the typological approach to race by incorporating Hooton's traits into lists using race-specific character states and extreme values based on small sample sizes that added very little to any understanding of human variation. In direct contrast to the typological, trait-list approach, Hefner (2009) provided detailed definitions and illustrations of the individual character states of each trait without any consideration of their distribution within and between populations. Using these definitions, Hefner (2009) provided frequency data within multiple populations to demonstrate their distribution across the globe. Later, Hefner and Ousley (2014) utilized that data in appropriate statistical models to demonstrate the effectiveness of those traits in the estimation of ancestry. Today, the typological approach to human variation using macromorphoscopic traits is considered inferior to these best-practice statistical approaches. Other researchers are building on Hefner's initial research, adding additional traits, testing and validating new statistical methods, and strengthening the macromorphoscopic approach to ancestry estimations (Hefner 2018). Unfortunately, missing data remain a problem in all these approaches.

Recently, increases in computational power have allowed for the creation and simple implementation of complex methods for imputing missing data, beyond using variable medians or modes. Instead of omitting missing cases or variables, practitioners can now use their own full data sets with missing data by employing data imputation. In practical terms, this means the missing values are modeled from values observed within the original data set.

Previously, Kenyhercz and Passalacqua (2016) demonstrated the utility of several imputation methods for craniometric data. However, to date there has been no treatment on the use of imputation methods for cranial macromorphoscopic data. The goals of this article are twofold: (1) to evaluate which data-imputation technique most accurately estimates macromorphoscopic trait scores; and (2) to evaluate the impact of each imputation method on ancestry classification accuracies.

Material and Methods

Sample

A total of 688 individual crania were scored by one of the authors (JTH) as part of the Macromorphoscopic Databank (MaMD) (Hefner 2018) (Table 1). The primary purpose of the MaMD is to address a substantial gap in forensic anthropology by providing the reference data needed to more accurately and objectively assess ancestry using macromorphoscopic trait data from numerous populations. As of November of 2018, the MaMD contained data on 17 variables for over 8,000 individuals from around the world. The end product will be a free-to-use data-collection and data-analysis program available to researchers.

TABLE 1—*Sample demographics by sex and ancestry.*

	Black	Hispanic	White	Total
Female	116	31	127	274
Male	176	86	83	345
Indeterminate	0	69	0	69
Total	292	186	210	688

Traits and Scoring

Six commonly employed cranial macromorphoscopic traits were included in this study (Hefner 2009; Hefner & Ousley 2014) (Table 2). The six traits used in the current study are those required for use with the Optimized Summed Scored Attributes (OSSA) method (Hefner 2009; Hefner & Ousley 2014). These six traits were selected because the OSSA method is currently employed in forensic casework (Kenyhercz et al. 2017). Previous work by Hefner (2009) found that the expression of cranial macromorphoscopic traits were not sex biased, so all individuals were pooled into their respective ancestry for all analyses.

Data Preparation

Prior to analysis, each individual in the study sample had a complete macromorphoscopic data set. A custom R function was used to randomly delete predefined proportions of the data (0.10, 0.25, 0.50, 0.75, and 0.90) and is available from the authors upon request. The function iteratively replaces the values designated NA (the commonly used abbreviation for "not available" to let R know that the cell has no information) using four different imputation methods. Each imputed data set is then fit to a canonical analysis of principal coordinates (CAP; described below) and the model accuracy stored within R. Once all four imputation methods have been simulated 500 times and all 2,000 new data tables have been

TABLE 2—*Traits used in the current study, their abbreviations, and potential character states.*

Trait	Abbreviation	Score
Anterior nasal spine	ANS	1–3
Inferior nasal aperture	INA	1–5
Interorbital breadth	IOB	1–3
Nasal aperture width	NAW	1–3
Nasal bone structure	NBS	0–4
Post-bregmatic depression	PBD	0–1

subjected to the CAP, the function returns a classification matrix for each simulation and each imputation method. A total of 2,500 different missing data sets were generated: 1,000 with slight to moderate amounts of missing data (10% and 25% missing for each trait) and 1,500 with severe amounts of missing data (50%, 75%, and 90% missing for each trait).

Data-Imputation Techniques

The four imputation methods used in the current study include (1) Hot Deck; (2) iterative robust model-based imputation (IRMI); (3) *k*-nearest neighbor (*k*-NN); and (4) variable medians. All analyses were conducted in R (R Core Team 2016), and all observations were set as ordered factors so the program did not treat trait scores as continuous variables. All but one of the data-imputation techniques was completed using the package *'VIM'* (Schopfhauser et al. 2014); median imputation was completed using the *'imputeMissings'* package (Meire et al. forthcoming). Previously, Little and Rubin (2002) explained that the goal of data imputation is to generate plausible values over more accurate values. However, Kenyhercz and Passalacqua (2016) argued that, for forensic purposes, the most accurate imputation method is the most favorable. Accurate imputation methods allow for less bias in the results and thus stronger analytical conclusions. To assess stability within each imputation method across the 500 simulations, descriptive statistics and diagnostic plots were generated and evaluated.

Hot Deck. Hot Deck imputation is an inductive data-ordering imputation technique similar to *k*-NN. However, unlike *k*-NN, Hot Deck employs a randomly selected donor from a pool of similar individuals to impute the missing values (Andridge & Little 2010). Here, similarity between the donor pool and the recipient is based on implicit assumptions regarding the choice of a distance metric to match potential donors to potential recipients. For this application, random Hot Deck imputations using a generalized distance function to measure similarity were used, first, to match donors and recipients, and second, to randomly select viable donors for imputation. The advantage of random Hot Deck imputation is that it does not rely on model building and is not as sensitive to misspecification, an issue that may occur using parametric methods such as regression imputation (Andridge & Little 2010). Conversely, if missing variables are imputed one at a time, relationships among the data are ignored, which may obscure trait interactions necessary for ancestry estimation.

IRMI. In IRMI, each missing data value is treated as a response variable while the other variables are used as the regressors (Templ et al. 2011). Hefner (2009) and Hefner and Ousley (2014) previously noted the non-normal distribution of cranial macromorphoscopic traits which may not be appropriately modeled by some imputation methods. However, IRMI is an iterative, model-based approach to missing data robustly handling continuous, categorical, binary, and mixed-response variables (Kowarik & Templ 2017; Templ et al. 2012) through fully conditional specification (FCS) and imputation via chained equations. That is, each variable is imputed using a regression model, conditioned on all the other variables, tailored to the specific nature of the imputed variable: continuous variables are imputed via linear regression, whereas binary variables are imputed by logistic regression. In this way, however, IRMI can produce values not represented in the scoring structure of the macromorphoscopic data. If IRMI is employed, we suggest replacing any imputed values greater than possible character state scores with the greatest expression for that trait.

k-*NN.* The *k*-NN algorithm identifies a number (*k*) of similar neighbors (individuals without missing values) having a similar distribution of available (non-missing) trait scores using a similarity distance measure (default measure is Euclidean). The median value of this missing trait score is calculated from these *k*-nearest neighbors and used for each of the missing imputations (Batista & Monard 2002). The current study used the nearest five neighbors, which was previously an effective and unbiased representation of the actual values (Hastie et al. 1999). However, in the event of severe missing data, the number of neighbors may be adjusted to reach a more representative number of nearest neighbors.

Variable Medians. The last imputation technique was median replacement. The median values were calculated from the entire data set, as it is unrealistic to assume a specific ancestral group prior to analysis, which is not an issue with the other methods described above since they are based on inter-individual measures of similarity. Variable medians act primarily as placeholders, because no new information is actually gleaned from their involvement.

Data Agreement

All of the 500 simulated NA data sets for each imputation method were compared to the original values with a weighted Cohen's kappa statistic (κ). The weighted Cohen's kappa was calculated using the 'irr' package (Gamer et al. 2012) using equal weights so that each disagreement was treated equally. Essentially, the original data set and each imputed data set were considered different observers. A κ value of 1.0 would indicate perfect agreement between the original data set and the respective imputation method. Given that κ is designed

to examine the reproducibility of measurements, it is appropriate to treat each imputation method as a separate observer, because the aim of the current study is to identify which method, or methods, most accurately reproduces the original data sets. Following Cicchetti (1994), agreement is considered poor (<0.40), fair (0.40–0.59), good (0.60–0.74), or excellent (0.75–1.00).

The κ values and their standard deviations were averaged from the 500 simulations for each of the imputation techniques at each of the different levels of missing data in order to generate an unbiased assessment of agreement. From these, the mean κ values of each individual method were averaged as a measure of overall method performance (method mean), and the mean κ for each individual trait was tabulated as a measure of trait imputability (trait agreement mean). Further, all of the imputed values for a particular imputation method were averaged to show which method produced the highest overall κ values, regardless of level of missing data. Next, the mean imputed κ values for each variable were averaged across all methods and levels of missing data to examine the overall imputability of each trait. Furthermore, the mean κ values for all traits were averaged at each level of missing data to explore the impact of missing data on agreement. A one-way ANOVA was used to test for the significance between the different imputation techniques estimates of κ for each variable at each of the different levels of missing data. Lastly, pairwise comparisons of the each imputation method's estimate of κ was performed with Tukey's HSD (honest significance difference) test for each variable and at each level of missing data.

Classification

To assess the potential of each imputation method, we measured the classificatory power of each derived data set using a canonical analysis of principal coordinates (CAP). Legendre and Legendre (1998) proposed a canonical discriminant analysis performed on the transformed values of the principal coordinates. In short, a CAP applies a principal coordinate analysis using any one of several distance measures (Anderson & Willis 2003), essentially transforming categorical variables into continuous, normally distributed variables. In that way, the CAP method is highly effective for dealing with macromorphoscopic data and enables classification and visualization of the groups in a manner approximating craniometric analyses. To that end, the chi-square distance metric (Anderson 2005) was used to calculate inter-individual similarities (and a distance matrix) within each imputed data set. Permutations were applied on the resulting similarity matrices. For the interpretation of significance we considered permutation *p*-values, since the permutation number (4,999) was considered high (Anderson 2005; Hefner 2016). Following 500 simulations, an evaluation of the cross-validated (leave-one-out) classification accuracies within and between the imputation methods provides a robust comparison of the four imputation methods. The CAP analysis was conducted using the *'CAPdiscrim'* function in the *BiodiversityR* package (Kindt & Coe 2005).

Results

Agreement

10% Missing Data. The mean κ and standard deviations from the 500 simulations for each imputed data set with 10% missing data and the original data set are presented in Table 3. Each imputation method has excellent agreement with the original data set. Across the board, IRMI has the strongest correlations with the actual data set for each trait (method mean κ = 0.938), followed by *k*-NN (method mean κ = 0.928) and, lastly, median replacement (method mean κ = 0.923). Hot Deck, while still having excellent agreement, was outperformed by the other imputation techniques and has a slightly

TABLE 3—*Mean κ and standard deviations of 500 simulations for each imputation technique and trait with the complete data set and each of the NA data sets with 10% missing data. The greatest correlation between an imputed trait and actual score is shown in bold for the first missing data set and in bold underline for the second missing data set.*

	k-NN		IRMI		Hot Deck		Variable Median		Trait Agreement
Trait	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean
ANS	**<u>0.933</u>**	0.009	**0.938**	0.009	0.900	0.011	0.925	0.008	0.924
INA	**<u>0.935</u>**	0.007	**0.938**	0.007	0.900	0.009	0.924	0.009	0.924
IOB	**<u>0.934</u>**	0.009	**0.943**	0.008	0.900	0.011	0.930	0.008	0.927
NAW	**<u>0.939</u>**	0.009	**0.949**	0.008	0.900	0.011	0.928	0.008	0.929
NBS	**<u>0.918</u>**	0.009	**0.937**	0.008	0.900	0.010	0.915	0.009	0.917
PBD	0.911	0.013	**0.923**	0.013	0.900	0.013	**<u>0.918</u>**	0.013	0.913
Method Mean	0.928	0.009	0.938	0.009	0.900	0.011	0.923	0.009	—

higher standard deviation (0.011). Lastly, using trait agreement mean κ values as a proxy for trait imputability, NAW (0.929) has the highest trait agreement mean κ value and PBD has the lowest (0.913). Imputation techniques are significantly different for each variable (all $p < 0.001$), and all of the pairwise comparisons from the Tukey's HSD tests are also significantly different at $p < 0.001$.

25% Missing Data. The mean κ and standard deviations from the 500 simulations for each imputed data set with 25% missing data and the original data set are presented in Table 4. Again, each imputation method has excellent overall agreement with the original data set. IRMI has the highest level of agreement for each variable (method mean $\kappa = 0.849$), followed by *k*-NN (method mean $\kappa = 0.829$), median replacement (method mean $\kappa = 0.812$), and Hot Deck (method mean $\kappa = 0.812$). Hot Deck and median replacement have marginally larger standard deviations compared to the other imputation techniques. In terms of trait imputability, NAW again has the highest trait agreement mean (0.828) and PBD the lowest (0.792). Imputation techniques are significantly different for each variable (all $p < 0.001$), and all of the pairwise comparisons from the Tukey's HSD tests are also significantly different at $p < 0.001$.

50% Missing Data. The mean κ and standard deviations from the 500 simulations for each imputed data set with 50% missing data and the original data set are presented in Table 5. Each imputation method has good agreement with the original data set. Following previous trends, IRMI has the highest mean κ for each variable (method mean $\kappa = 0.676$), followed by *k*-NN (method mean $\kappa = 0.649$), median replacement (method mean $\kappa = 0.595$), and Hot Deck (method mean $\kappa = 0.528$). Again, median replacement and Hot Deck have slightly larger standard deviations than the other two imputation techniques. Trait imputability is greatest in NAW (0.636) and lowest in PBD (0.571). Imputation techniques are significantly different for each variable (all $p < 0.001$), and all of the pairwise comparisons from the Tukey's HSD tests are also significantly different at $p < 0.001$ except in PBD between IRMI and *k*-NN ($p = 0.424$), and also between IRMI and median replacement ($p = 0.316$).

75% Missing Data. The mean κ and standard deviations from the 500 simulations for each imputed data set with 75% missing data and the original data set are presented in Table 6. IRMI, *k*-NN, and most of the median replacement imputed traits have fair agreement with the original data set, while Hot Deck and two median replacement imputed traits have

TABLE 4—*Mean κ and standard deviations of 500 simulations for each imputation technique and trait with the complete data set and each of the NA data sets with 25% missing data. The greatest correlation between an imputed trait and actual score is shown in bold for the first missing data set and in bold underline for the second missing data set.*

	k-NN		IRMI		Hot Deck		Median		Trait Agreement
Trait	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean
ANS	**<u>0.838</u>**	0.013	**0.850**	0.012	0.764	0.016	0.816	0.014	0.817
INA	**<u>0.843</u>**	0.010	**0.853**	0.012	0.764	0.015	0.810	0.022	0.818
IOB	**<u>0.839</u>**	0.014	**0.860**	0.013	0.763	0.017	0.827	0.013	0.822
NAW	**<u>0.855</u>**	0.014	**0.871**	0.014	0.764	0.016	0.823	0.013	0.662
NBS	**<u>0.811</u>**	0.013	**0.849**	0.012	0.764	0.015	0.794	0.013	0.804
PBD	0.790	0.019	**0.813**	0.020	0.764	0.020	**<u>0.801</u>**	0.019	0.792
Method Mean	0.829	0.014	0.849	0.014	0.764	0.017	0.812	0.016	—

TABLE 5—*Mean κ and standard deviations of 500 simulations for each imputation technique and trait with the complete data set and each of the NA data sets with 50% missing data. The greatest correlation between an imputed trait and actual score is shown in bold for the first missing data set and in bold underline for the second missing data set.*

	k-NN		IRMI		Hot Deck		Median		Trait Agreement
Trait	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean
ANS	**<u>0.659</u>**	0.019	**0.681**	0.017	0.528	0.022	0.597	0.029	0.616
INA	**<u>0.675</u>**	0.015	**0.694**	0.017	0.528	0.020	0.583	0.041	0.620
IOB	**<u>0.662</u>**	0.020	**0.692**	0.020	0.527	0.022	0.623	0.018	0.626
NAW	**<u>0.690</u>**	0.019	**0.709**	0.021	0.528	0.023	0.616	0.018	0.636
NBS	**<u>0.628</u>**	0.017	**0.682**	0.019	0.528	0.020	0.571	0.018	0.602
PBD	0.578	0.028	**0.595**	0.031	0.530	0.028	**<u>0.582</u>**	0.026	0.571
Method Mean	0.649	0.020	0.676	0.021	0.528	0.023	0.595	0.025	—

TABLE 6—*Mean κ and standard deviations of 500 simulations for each imputation technique and trait with the complete data set and each of the NA data sets with 75% missing data. The greatest correlation between an imputed trait and actual score is shown in bold for the first missing data set and in bold underline for the second missing data set.*

	k-NN		IRMI		Hot Deck		Median		Trait Agreement
Trait	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean
ANS	**<u>0.515</u>**	0.024	**0.542**	0.023	0.362	0.025	0.414	0.031	0.458
INA	**<u>0.540</u>**	0.019	**0.563**	0.019	0.358	0.020	0.399	0.041	0.465
IOB	**<u>0.515</u>**	0.028	**0.547**	0.037	0.351	0.025	0.449	0.020	0.466
NAW	**<u>0.551</u>**	0.025	**0.563**	0.027	0.352	0.028	0.437	0.019	0.476
NBS	**<u>0.498</u>**	0.024	**0.542**	0.021	0.353	0.028	0.395	0.013	0.447
PBD	**<u>0.413</u>**	0.041	**0.416**	0.047	0.361	0.031	0.408	0.032	0.399
Method Mean	0.505	0.027	0.529	0.029	0.356	0.026	0.417	0.026	—

poor agreement. IRMI has the highest levels of agreement with each variable, as well as overall agreement (method mean $\kappa = 0.529$), followed by *k*-NN (method mean $\kappa = 0.505$). Median replacement (method mean $\kappa = 0.417$) and Hot Deck (method mean $\kappa = 0.356$) have much lower overall agreement. As before, NAW is the most imputable trait (0.476) and PBD is the least (0.399). Each imputation technique is significantly different for each variable; however, pairwise comparisons of the Tukey's HSD test show that the IRMI and *k*-NN values are not significantly different ($p = 0.131$) and that for PBD the following comparisons are not significantly different: median replacement and *k*-NN ($p = 0.961$), IRMI and *k*-NN ($p = 0.991$), and IRMI and median replacement ($p = 0.997$).

90% Missing Data. The mean κ and standard deviations from the 500 simulations for each imputed data set with 90% missing data and the original data set are presented in Table 7. Fair agreement is only observed in the IRMI imputed INA, while the rest of the imputations all show poor agreement. IRMI and *k*-NN have the highest level of agreement for three traits, though IRMI has an overall greater method mean κ (0.338). Hot Deck and median replacement both have substantially lower agreements, both in individual traits and overall. INA has the highest overall imputability (0.277), and PBD, again, has the lowest (0.178). There are no significant differences among imputation techniques, nor pairwise comparisons with each displaying values over $p > 0.05$.

All Data Sets. A summary of the mean κ values for each imputation method at each level of missing data is presented in Table 8 and shown graphically in Figure 1. The average of the mean κ values for IRMI is the greatest at 0.666, followed by *k*-NN at 0.649, median replacement at 0.586, and Hot Deck at 0.536. In terms of imputability, NAW has the greatest overall average among all of the imputation methods and levels (0.627; see Fig. 1D) followed by INA (0.621; see Fig. 1B), IOB (0.620; see Fig. 1C), ANS (0.614; see Fig. 1A), NBS (0.605; see Fig. 1E), and PBD (0.571; see Fig. 1F). At 10% missing data, the average of each of the traits' mean κ across all imputation methods is 0.922, at 25% it is 0.814, at 50% it is 0.612, at 75% it is 0.451, and at 90% it is 0.247.

Classification

Table 9 provides the global classification accuracies for all five data sets, averaged across the four imputation methods and the original data set. All values are significant at $p < 0.05$ (based on 4,999 permutations for each of the 500 simulations). Figure 2 illustrates the accuracy for each simulation ($N = 500$) by imputation method, and Figure 3 provides kernel density

TABLE 7—*Mean κ and standard deviations of 500 simulations for each imputation technique and trait with the complete data set and each of the NA data sets with 90% missing data. The greatest correlation between an imputed trait and actual score is shown in bold for the first missing data set and in bold underline for the second missing data set.*

	k-NN		IRMI		Hot Deck		Median		Trait Agreement
Trait	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean
ANS	**<u>0.336</u>**	0.000	**0.352**	0.047	0.154	0.012	0.169	0.026	0.253
INA	**<u>0.362</u>**	0.006	**0.409**	0.007	0.142	0.005	0.193	0.003	0.277
IOB	**0.358**	0.012	**<u>0.347</u>**	0.012	0.122	0.051	0.210	0.003	0.259
NAW	**0.391**	0.074	**<u>0.361</u>**	0.085	0.113	0.053	0.190	0.014	0.264
NBS	**<u>0.325</u>**	0.063	**0.369**	0.088	0.140	0.003	0.180	0.006	0.254
PBD	**0.239**	0.035	**<u>0.193</u>**	0.032	0.120	0.040	0.162	0.004	0.178
Method Mean	0.335	0.032	0.338	0.045	0.132	0.028	0.184	0.009	—

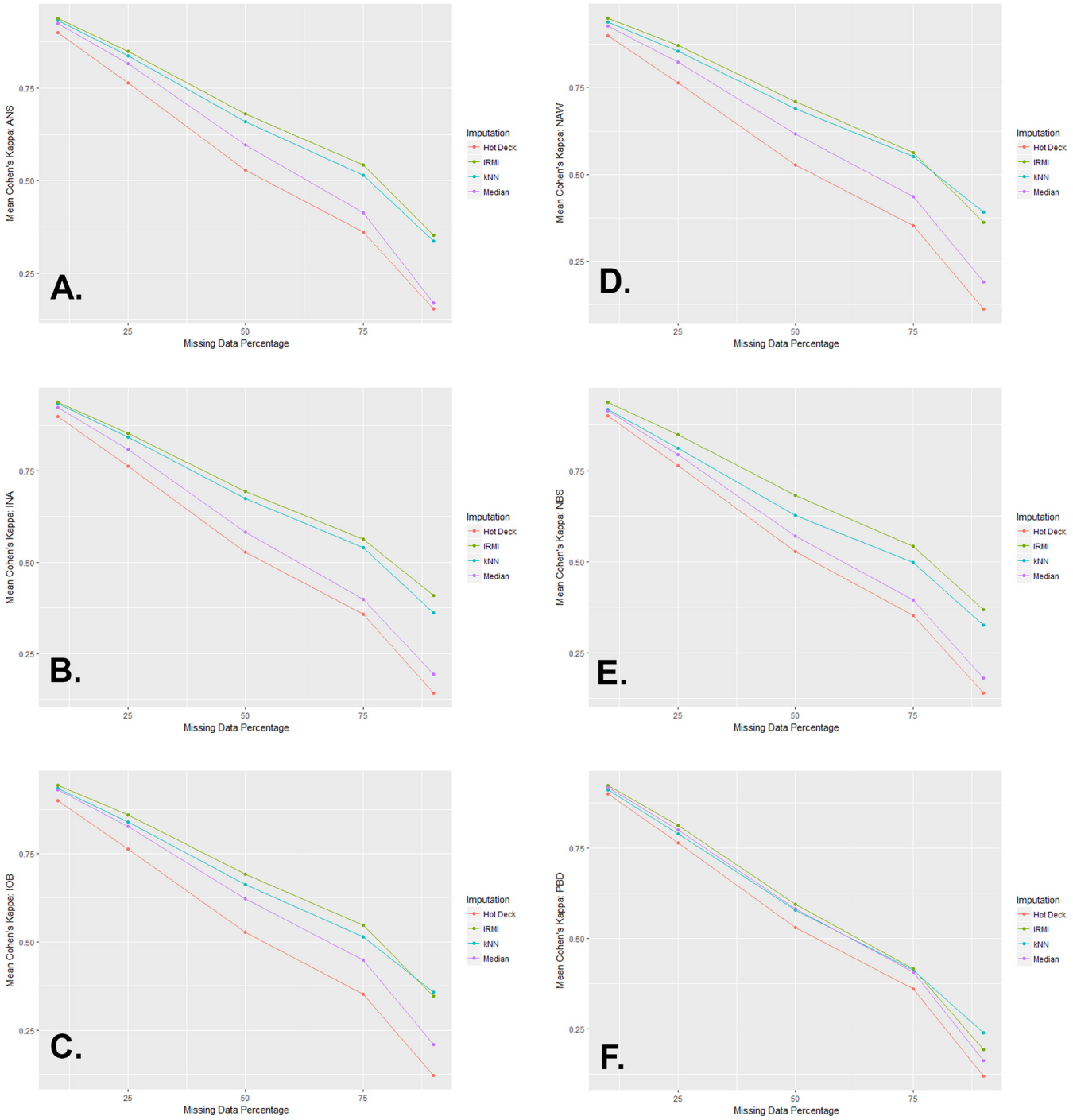

FIG. 1—*Line plot of mean K (*N *simulations = 500) over percentage of missing data by each imputation method. A = ANS; B = INA; C = IOB; D = NAW; E = NBS; F = PBD.*

estimates of the correct classification accuracy of each of the four imputation methods by the percent of missing data values introduced to the original data set. All four methods performed similarly within the various levels of missingness and the imputation accuracies within each did not vary significantly. For 10% missing data, Hot Deck shows the highest total correct classification (68.70%); in fact, it is higher than the original data set (65.55%). However, IRMI shows the lowest bias (–0.70%). At 25% missing data, again, Hot Deck provides the greatest total correct classification (69.68%), which is again higher than the original data set, but IRMI still shows the lowest bias (–2.51%). With 50% of the data missing, IRMI has the highest correct classification (58.01%) and the lowest bias (–7.64%). Following suit, at

TABLE 8—*Summary of mean κ values for each imputation method at each level of missing data, and averages of variable and method means.*

Method	ANS	INA	IOB	NAW	NBS	PBD	Method Mean
k-NN							
10%	0.933	0.935	0.934	0.939	0.918	0.911	0.928
25%	0.838	0.843	0.839	0.855	0.811	0.790	0.829
50%	0.659	0.675	0.662	0.690	0.628	0.578	0.649
75%	0.515	0.540	0.515	0.551	0.498	0.413	0.505
90%	0.336	0.362	0.358	0.391	0.325	0.239	0.335
IRMI							
10%	0.938	0.938	0.943	0.949	0.937	0.923	0.938
25%	0.850	0.853	0.860	0.871	0.849	0.813	0.849
50%	0.681	0.694	0.692	0.709	0.682	0.595	0.676
75%	0.542	0.563	0.547	0.563	0.542	0.416	0.529
90%	0.352	0.409	0.347	0.361	0.369	0.193	0.339
Hot Deck							
10%	0.900	0.900	0.900	0.900	0.900	0.900	0.900
25%	0.764	0.764	0.763	0.764	0.764	0.764	0.764
50%	0.528	0.528	0.527	0.528	0.528	0.530	0.528
75%	0.362	0.358	0.351	0.352	0.353	0.361	0.356
90%	0.154	0.142	0.122	0.113	0.140	0.120	0.132
Median							
10%	0.925	0.924	0.930	0.928	0.915	0.918	0.923
25%	0.816	0.810	0.827	0.823	0.794	0.801	0.812
50%	0.597	0.583	0.623	0.616	0.571	0.582	0.595
75%	0.414	0.399	0.449	0.437	0.395	0.408	0.417
90%	0.169	0.193	0.210	0.190	0.180	0.162	0.184
Trait Agreement Mean	0.614	0.621	0.620	0.627	0.605	0.571	—

TABLE 9—*Global classification accuracies and bias for all five data sets, averaged across the imputation methods. The greatest total correct classifications are shown in bold. The lowest bias is italicized. Note, the original, non-imputed data set has a mean correct classification of 65.55% with a standard deviation of 1.42%.*

	k-NN		IRMI		Hot Deck		Median	
Percent Missing	Total Correct	Bias	Total Correct	Bias	Total Correct	Bias	Total Correct	Bias
10%								
Mean	64.75%	–0.80%	64.85%	*–0.70%*	**68.70%**	3.15%	64.58%	–0.97%
SD	0.91%	—	0.81%	—	1.07%	—	0.74%	—
25%								
Mean	62.64%	–2.91%	63.04%	*–2.51%*	**69.68%**	4.13%	62.30%	–3.25%
SD	1.24%	—	1.33%	—	1.53%	—	1.18%	—
50%								
Mean	56.67%	–8.88%	**58.01%**	*–7.54%*	52.55%	–13.00%	56.41%	–9.14%
SD	1.70%	—	1.67%	—	1.58%	—	1.55%	—
75%								
Mean	49.86%	–15.69%	**51.05%**	*–14.50%*	46.12%	–19.43%	49.68%	–15.87%
SD	1.74%	—	1.29%	—	1.47%	—	1.48%	—
90%								
Mean	45.76%	–19.79%	**46.50%**	*–19.05%*	43.41%	–22.14%	45.57%	–19.98%
SD	1.50%	—	**0.92%**	—	1.12%	—	1.12%	—

75% missing data, IRMI has the highest correct classification (51.05%) and the lowest bias (–14.50%). Lastly, with 90% missing data, IRMI again has the greatest classification (46.50%) and the lowest bias (–19.05%). The single greatest bias is observed with Hot Deck imputation at 90% missing data (22.14%).

Overall, IRMI has the highest mean total correct classification averaged across all levels of missing data (56.69%), followed by Hot Deck (56.09%), *k*-NN (55.94%), and median replacement (55.71%). The averaged biases reflect the same trend, with IRMI showing the least averaged bias across each level of missing data (–8.86%), followed by Hot Deck (–9.46%),

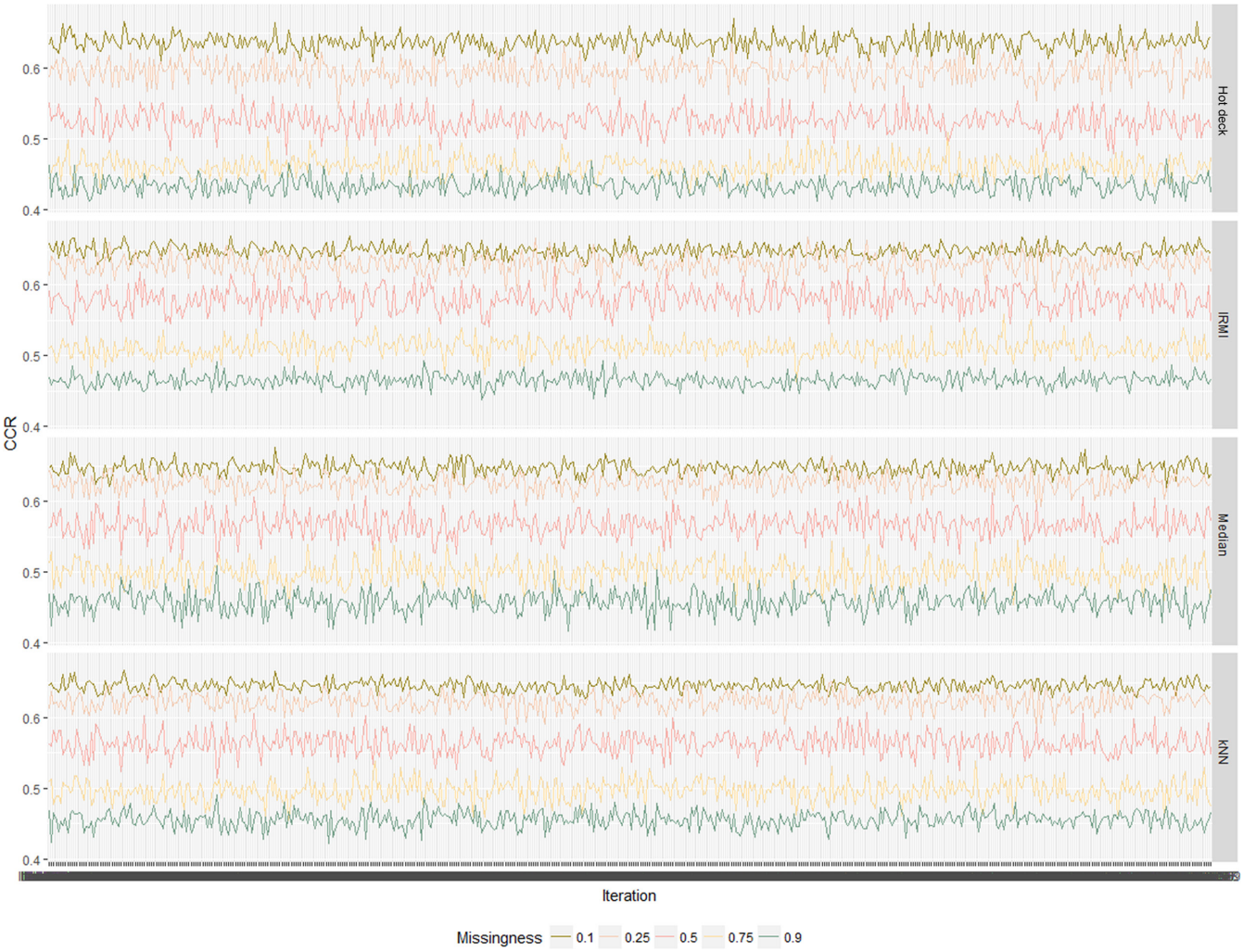

FIG. 2—*Accuracy for each simulation (*N = *500), by imputation method.*

k-NN (–9.61%), and median replacement (–9.84%). On average, there is essentially no meaningful difference in correct classifications of any of the imputation methods with 10% and 25% missing data and the original data set. At 50% missing data, model performance is diminished by –9.64% on average; at 75%, by –16.37% on average, and at 90%, by –20.24%.

Discussion

The various macromorphoscopic data-imputation methods all performed relatively well, even in data sets with severe amounts of missing data. In all instances, IRMI was the most successful imputation technique in terms of overall agreement and in overall classifications (with the least overall bias). The success of IRMI on macromorphoscopic traits is very likely due to the independent modeling of each variable, which is particularly important given that these traits have different levels of expression from two through five. To compound this issue, macromorphoscopic traits are not relegated solely to one ancestral group—some groups may have higher frequencies of certain trait scores, but all trait variants are observed in all groups—though suites of these traits do tend to vary among groups (Hefner 2009; Hefner & Ousley 2014). The genetic basis for the expression of macromorphoscopic traits is likely captured by and informing the individual regressions of the variables more than the other imputation methods do so. Consistently, *k*-NN shows the second-highest levels of agreement. Kenyhercz and Passalacqua (2016) found *k*-NN to outperform other imputation techniques when working with craniometrics data. However, for these macromorphoscopic traits, *k*-NN was not as effective as IRMI.

Hot Deck imputation consistently underperformed the other techniques in each of the different levels of missing data. Hot Deck relies on randomly selecting donors that have similar expressions of available traits to impute the missing values. As the amount of missing data increases, the number of suitable donors decreases. While median replacement

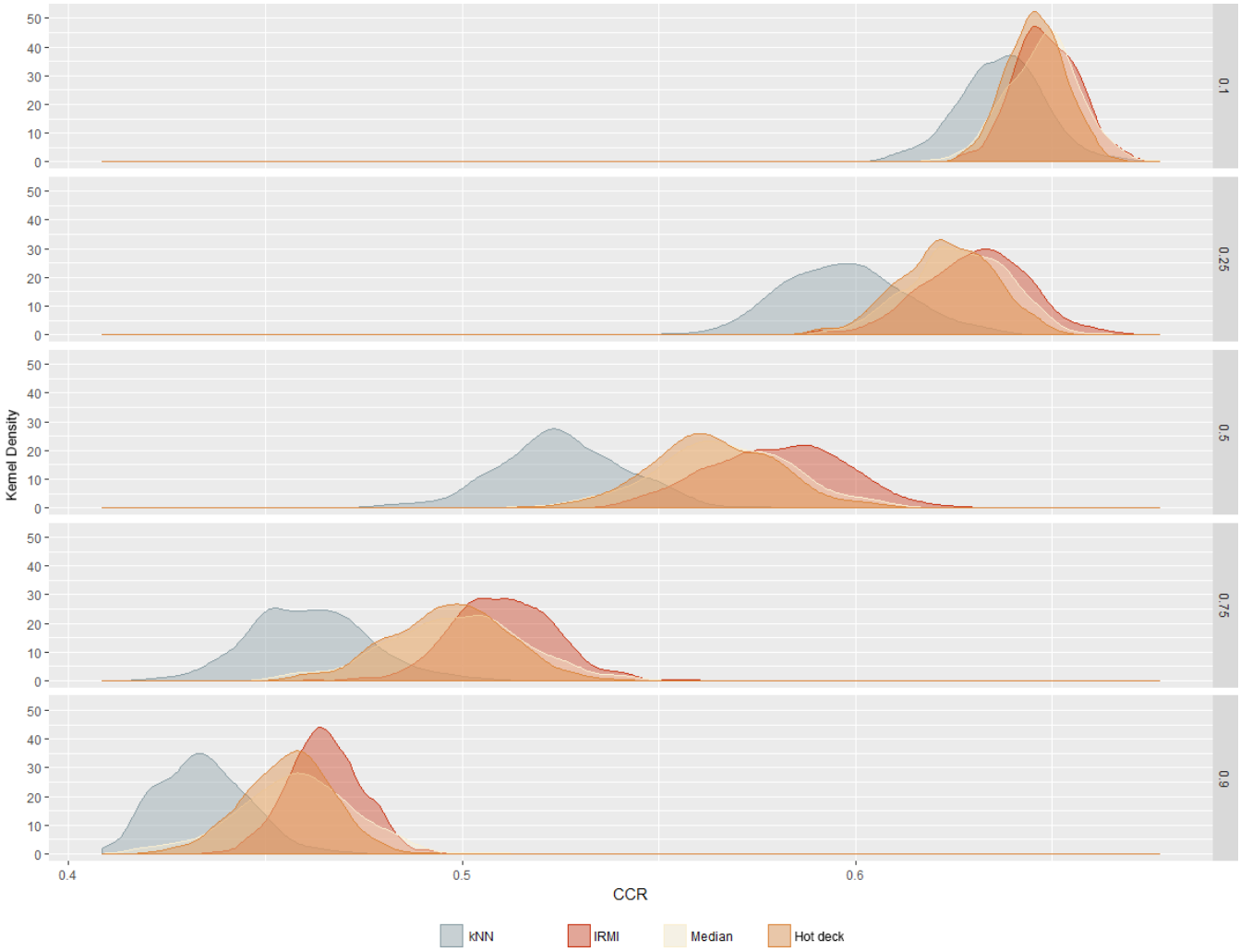

FIG. 3—*Kernel density estimates of the correct classification accuracy of each of the four imputation methods, by the percent of missing data values introduced to the original data set.*

performed well for a simple way to impute missing values, better-informed models should be used when imputing data, especially given the ease involved in applying the more robust imputation methods. Further, in instances of severe missing data (75% and 90%), the median replacement agreement dropped relative to the IRMI and *k*-NN imputations. Overall, for macromorphoscopic traits with missing data, IRMI imputations provided the best results.

In terms of total correct classifications, Hot Deck provided greater correct classifications than the original data set for the 10% and 25% NA data sets. However, while there was better classification with the Hot Deck imputed values, they were not as representative of the actual data set and should not be used. As Kenyhercz and Passalacqua (2016) argued, the goal of data imputation should be producing the most accurate values, particularly for forensic contexts. To this end, IRMI was again the most successful imputation model, as it consistently demonstrated the least amount of absolute classification bias across each level of missing data.

Missing data, particularly missing data in skeletal populations, may not always be missing completely at random (MCAR) or missing at random (MAR). For example, a cranium missing some midfacial structure like the nasal bones will have multiple missing data values from the region (e.g., NBC and NAW). Mealli and Rubin (2016) clarify issues of random and nonrandom missing data and the exchangeability of data values under the assumption of randomly missing values. Following Mealli and Rubin (2016), we have to assume any mechanism leading to missing data in a given skeletal sample will depend only on the fully observed variables. In other words, the observable values (and thus the values used for imputation) will always have conditional independence, regardless of the mechanism leading to the missing data. Further research into values missing not at random (MNAR) in skeletal samples will shed light on the effects of these assumptions in missing data imputation.

Data can be imputed for reference data sets and also individual cases with relative ease. For example, using the

VIMGUI package (Schopfhauser et al. 2014) in R, many of the methods discussed above can be used to quickly impute missing data values. In the event of a single case with missing data, the practitioner can include that case with a unique identifier into a reference data set and, using VIMGUI, impute the missing values for further analysis. However, we recommend that at least 50% of the observable variables be present to generate an accurate, or meaningful, imputation; as Hefner (2009) pointed out, it is the suite of nonmetric traits that vary among populations, not individual traits alone. Without at least half of the traits, it is difficult to model macromorphoscopic trait expression meaningfully.

Conclusion

The statistical estimation of ancestry from crania with missing data using macromorphoscopic traits is actually a relatively straightforward endeavor. Different levels of missing data affect the effectiveness of various imputation techniques on macromorphoscopic cranial data and the accuracy of ancestry estimation from incomplete crania. When we are faced with missing macromorphoscopic data, IRMI should be used for imputation, as it consistently generated the highest levels of agreement and overall classification with the least amount of overall absolute bias. If more than half of observable data are missing in an individual case, imputation may not be appropriate. Nonetheless, abandoning statistical methods or analyses altogether in cases where data are missing is not an appropriate or effective analytical option.

Acknowledgments

We wish to thank all of the collection's managers for making this research possible. Special thanks to Dhruv Sharma for assistance in the R code for the simulations. Our thanks also to the anonymous reviewers for their comments and critiques, which strengthened the article.

References

Anderson MJ. CAP: A computer program. Auckland, New Zealand; 2005.

Anderson MJ, Willis TJ. Canonical analysis of principal coordinates: A useful method of constrained ordination for ecology. *Ecology* 2003;84:511–525.

Andridge RR, Little RJ. A review of Hot Deck imputation for survey non-response. *International Statistical Review* 2010;78(1); 40–64.

Batista GE, Monard MC. A study of *k*-nearest neighbor as an imputation method. *HIS* 2002:87;48.

Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* 1994;6(4):284–290.

Gamer M, Lemon J, Fellows I, Singh P. Various coefficients of interrater reliability and agreement. R package version 0.84 version. 2012. Accessed April 20, 2016.

Hastie T, Tibshirani R, Sherlock G, Eisen M, Brown P, Botstein D. Imputing missing data for gene expression arrays. Stanford, CA: Stanford University Statistics Department; 1999.

Hefner JT. Biological distance analysis, cranial morphoscopic traits, and ancestry assessments in forensic anthropology. In: Pilloud MA, Hefner JT, eds. *Biological Distance Analysis: Forensic and Bioarchaeological Perspectives*. San Diego: Academic Press; 2016:301–316.

Hefner JT. Cranial nonmetric variation and estimating ancestry. *Journal of Forensic Sciences* 2009;54(5):985–995.

Hefner JT. The Macromorphoscopic Databank. *American Journal of Physical Anthropology* 2018;166(4):994–1004.

Hefner JT, Ousley SD. Statistical classification methods for estimating ancestry using morphoscopic traits. *Journal of Forensic Sciences* 2014;59(4):883–890.

Hefner JT, Ousley SD, Dirkmaat DC. Morphoscopic traits and the assessment of ancestry. In: Dirkmaat DC, ed. *A Companion to Forensic Anthropology*. New York: John Wiley & Sons; 2012: 287–310.

Hooton EA. *Up from the Ape*. 2nd ed. New York: Macmillan; 1946.

Kenyhercz MW, Klales AR, Rainwater CW, Fredette SM. The optimized summed scored attributes method for the classification of U.S. Blacks and Whites: A validation study. *Journal of Forensic Sciences* 2017;62(1):174–180.

Kenyhercz MW, Passalacqua NP. Missing data imputation methods and their performance with biodistance analyses. In: Pilloud MA, Hefner, JT, eds. *Biological Distance Analysis: Forensic and Bioarchaeological Perspectives*. San Diego: Academic Press; 2016:181–194.

Kindt R, Coe R. Tree diversity analysis: A manual and software for common statistical methods for ecological and biodiversity studies. World Agroforestry Centre, Nairobi; 2005.

Kowarik A, Templ M. Imputation with the R Package VIM. https://cran.r-project.org/web/packages/VIM/index.html. Created April 11, 2017. Accessed May 18, 2018.

Legendre P, Legendre LFJ. *Numerical Ecology*. 2nd ed. Amsterdam: Elsevier; 1998.

Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. Hoboken, NJ: Wiley-Interscience; 2002.

Mealli F, Rubin DB. Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika* 2016;103(2):491.

Meire M, Ballings M,Van den Poel D. *Predictive Analytics for Analytical Customer Relationship Management Using SAS, Oracle and R*. Springer; forthcoming.

Passalacqua NV, Zhang Z, Pierce SJ. Sex determination of human skeletal populations using latent profile analysis. *American Journal of Physical Anthropology* 2013;151(4):538–543.

R Core Team. R: A language and environment for statistical computing, version 3.3.0. R Foundation for Statistical Computing, Vienna, Austria; 2016. Accessed March 23, 2016.

Rhine S. Nonmetric skull racing. In: Gill G, Rhine S, eds. Skeletal attribution of race: Methods for forensic anthropology. Maxwell Museum of Anthropological Papers No. 4. Albuquerque: University of New Mexico; 1990:9–20.

Schopfhauser D, Templ M, Kowarik A, Prantner B. VIMGUI: Visualization and imputation of missing values. R package version 0.9.0. 2014. Accessed April 20, 2016.

Templ M, Alfons A, Filzmoser P. Exploring incomplete data using visualization techniques. *Advances in Data Analysis Classification* 2012;6(1):29–47

Templ M, Kowarik A, Filzmoser P. Iterative stepwise regression imputation using standard and robust methods. *Computational Statistics & Data Analysis* 2011;55(10);2793–2806.