A88 Missing Data Imputation Methods Using Morphoscopic Traits and Their Performance in the Estimation of Ancestry

Michael W. Kenyhercz, PhD, University of Tennessee, 250 S Stadium Hall, Knoxville, TN 37996; Nicholas V. Passalacqua, PhD, 1559 Mount Vernon, East Lansing, MI 48823; and Joseph T. Hefner, PhD, Michigan State University, Dept of Anthropology, 355 Baker Hall, East Lansing, MI 48824*

After attending this presentation, attendees will understand the effects of missing data imputation on the estimation of ancestry from the human cranium using morphoscopic variables.

This presentation will impact the forensic science community by demonstrating which missing data imputation methods have the highest accuracy when working with morphoscopic traits for the determination of ancestry.

Missing data is 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 peri-mortem trauma or taphonomic alterations that damage or destroy bony morphology. Often in cases with missing data, the variables in question are just excluded from analysis; however, the avoidance of analytical processes when missing data are present have the potential to significantly decrease the ability to reliably estimate aspects of the biological profile, limiting the number of variables used or decreasing the power of the estimations.

While different methods function differently, the goal of missing data imputation methods is to accurately estimate the missing values, using the other observed values. Unfortunately, the analysis of datasets with missing values receives little attention and listwise deletion is the most common form of handling cases with missing values; however, this is often not possible in forensic anthropology, where the entire sample is a single case that may present varying levels of missing data. Because of this, the ability to accurately impute missing data in forensic anthropology is paramount. The goal of this project is to quantify the accuracy of morphoscopic data in conditions with moderate (25%) and severe (50%) amounts of missing data.

Four data imputation techniques were selected to examine which of the missing data imputation methods performed best: Hot-Deck, Iterative Robust Model (IRMI), K nearest-neighbor (kNN=5), and the variable medians. A subset of Hefner's Macromorphiscopic Databank was used. The full sample consisted of 688 individuals from three population groups (Black=292, Hispanic=186, and White=210). Six commonly used cranial macroscopic variants were scored in accordance with Hefner, and Hefner and Ousley.[1,2] Two versions of the dataset were then created wherein values were randomly deleted from each variable so that 25% and 50% of the data were considered missing. The same data subsets were used for each of the imputation techniques, and the efficacy of each technique was based on absolute agreement using the Intra-Class Coefficient (ICC). Correct classification rates and Mahalanobis D^2 values were calculated for the original dataset with actual measures and each of the imputed datasets in order to examine the effects of imputed data on biodistance and classification. Results suggest that Hot-Deck imputation is the most accurate method with 25% missing data, with this imputation method consistently performing at the highest levels of agreement (up to ICC=0.97), with the least impact to the D^2 and, ultimately, classification. In severe instances of missing data (50%), IRMI consistently produced the highest levels of agreement. Kenyhercz and Passalacqua examined the effects of missing data using metric cranial data and found that when dealing with continuous variables, the kNN imputation method had the best performance in both moderate (25%) and severe (50%) missing data conditions.[3]

Reference(s):

1. Hefner J.T. Cranial nonmetric variation and estimating ancestry. *J Forensic Sci* 2009;54:985-995.
2. Hefner J.T., Ousley S.D. Statistical classification methods for estimating ancestry using morphoscopic traits. *J Forensic Sci.* 2014;4:883-890.
3. Kenyhercz M.W., Passalacqua N.V. Missing data imputation methods and their performance with biodistance analyses. *Program of the Am Assoc Phys Anthropol 84th Annual Meeting*, 2015, St. Louis, MO.

Non-Metric Data, Imputation, Missing Data

*Presenting Author