H43 Ancestry Estimation Using Random Forest Modeling

Joseph T. Hefner, PhD, Joint POW/MIA Accounting Command, Central Identification Laboratory, 310 Worchester Avenue, Building 45, Hickam AFB, HI 96583; Kate Spradley, PhD, Department of Anthropology, Texas State University, 601 University Drive, San Marcos, TX 78666; and Bruce E. Anderson, PhD, Forensic Science Center, Office of the Medical Examiner, 2825 East District Street, Tucson, AZ 85714*

After attending this presentation, attendees will be introduced to the use of Random Forest Modeling (RFM) and the performance of RFM in ancestry estimation.

This presentation will impact the forensic science community by providing an additional method for the estimation of ancestry.

Compared to other exploratory and classification methods used in anthropological research, for example, principal component analysis (PCA) and linear discriminant function analysis (IDFA), Random Forests may be more appropriately applied to datasets frequently encountered in forensic anthropology. The suitability of Random Forest models to

forensic anthropological data is due in large part to the rather rigid assumptions of parametric methods (i.e., observation independence, normal distribution, and homogeneity of variance), which do not hold for many of the datasets encountered during forensic anthropological research and method development. Fortunately, these assumptions are not required for nonparametric methods like Random Forest modeling. The most promising potential of Random Forest models is their ability to handle all variable types (categorical, continuous, count, etc.) seamlessly and to relate the various observations in highly non-linear ways to a response variable. Ancestry estimation as practiced by forensic anthropologists regularly incorporates both metric (continuous) and morphoscopic (categorical) data. In reality, most analysts prefer—or trust!—one method over the other. Only after one method (e.g., morphoscopic analysis) has provided results does the analyst turn to the next (e.g., metric analysis) for confirmation or refutation. Combining metric and morphoscopic predictor variables into a single classification analysis is generally not possible because of the differences in the distribution of the data. RFM avoids these issues using a nonparametric classification algorithm (a classifier consisting of a collection of tree- structured classifiers) relying on majority voting and bootstrapping to assign cases to a response class after the initial model is produced from a randomly selected training set. Further randomness is introduced during initial variable selection and tree construction by randomly selecting predictor variables, resulting in a 'forest' of trees contrasted of randomly selected individuals. A classification matrix (and various classification statistics) is then constructed to assess how well the model classifies all individuals in the dataset. Two supplementary measures produced during Random Forest analysis provide additional information: a measure of the importance of each predictor variable and a proximity measure (measure of the internal structure of the data). These statistics provide the analyst a great deal of information on the structure of the data (proximity measure) while identifying the most important variables— continuous and categorical, combined—to consider when estimating ancestry.

To examine the usefulness of Random Forest modeling in ancestry estimation, we applied the RFM classification algorithm to 34 standard cranial measurements and 16 standard morphoscopic traits collected from 149 crania. The sample represents modern American Whites (n = 72) and Blacks (n = 38) from the William M. Bass Donated Skeletal Collection in Knoxville, Tennessee and identified and unidentified border crossers representing Southwestern Hispanics (n = 39) from the Pima County Medical Examiner's Office in Tucson, Arizona. Using Random Forest, 89.5% of the cross-validated groups (by group: American Whites (AW) = 84.0%; American Blacks (AB) = 92.8%; Hispanic (H) = 92.6%) were correctly classified, substantially improving classifications compared to using traditional methods independently (craniometric = 76.1% [by group: AW = 81.0%, AB = 75.0%, and H = 69.2%]; morphoscopic = 72.7% [by group: AW = 70.0%, AB = 61.5%, and H = 85.7%]). Heuristically setting a threshold value at 0.50, thirty- four variables (seven morphoscopic, 27 craniometric) derived from the RFM variable importance measure were examined for underlying patterns to better understand their significance. The significant morphoscopic traits are all mid-facial (NAS, INA, IOB, NBC, NAW, ORB, and NSF), quantifying Brues (1990) assertion that the mid-facial skeleton is the most important area to consider when estimating ancestry, at least anthroposcopically. The significant craniometric variables are facial breadth (ZYB), orbital breadth (OBB), alveolar length (MAL), vault width (WFB, STB, ASB) and vault length (NOL, GOL), and alveolar prognathism (BPL). The metric variables do not follow the same pattern of the morphoscopic variables as they are not isolated to one specific area, but rather the craniometric variables seem to describe overall cranial morphology.

The results of the analysis using Random Forest modeling to estimate ancestry indicate that the combination of morphoscopic and craniometric datasets—which have for so long been diametrically opposed—greatly enhances the estimation of ancestry, allowing

 ** Presenting Author*

researchers to quantify the process of variable selection. In other words, the advantage of Random Forest modeling as a practicable classification alternative to traditional methods, such as morphoscopic trait lists and discriminant function analysis, is that analysts are freed from the obligation of defending method selection while maintaining the principle of ancestry estimation.

Forensic Anthropology, Ancestry Estimation, Quantitative Method

 ** Presenting Author*