JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, July 2014, Vol. 59, No. 4
doi: 10.1111/1556-4029.12421
Available online at: onlinelibrary.wiley.com

PAPER

ANTHROPOLOGY

Joseph T. Hefner,[1] *Ph.D.; and Stephen D. Ousley,*[2] *Ph.D.*

Statistical Classification Methods for Estimating Ancestry Using Morphoscopic Traits*,†

ABSTRACT: Ancestry assessments using cranial morphoscopic traits currently rely on subjective trait lists and observer experience rather than empirical support. The trait list approach, which is untested, unverified, and in many respects unrefined, is relied upon because of tradition and subjective experience. Our objective was to examine the utility of frequently cited morphoscopic traits and to explore eleven appropriate and novel methods for classifying an unknown cranium into one of several reference groups. Based on these results, artificial neural networks (aNNs), OSSA, support vector machines, and random forest models showed mean classification accuracies of at least 85%. The aNNs had the highest overall classification rate (87.8%), and random forests show the smallest difference between the highest (90.4%) and lowest (76.5%) classification accuracies. The results of this research demonstrate that morphoscopic traits can be successfully used to assess ancestry without relying only on the experience of the observer.

KEYWORDS: forensic science, forensic anthropology, morphoscopic traits, ancestry, classification statistics

As currently practiced by forensic anthropologists, ancestry assessments using cranial morphoscopic traits rely on subjective trait lists (1–3) and observer experience (4). In fact, there are currently very few, if any, empirically supported methodologies for assessing ancestry using morphoscopic traits. Unlike metric methods, morphoscopic traits have not been analyzed using the same statistical rigor, in part because of the difficulties encountered when working with categorical, rather than continuous, data (5), but also because forensic anthropologists have not necessarily been encouraged to quantify this traditional and seemingly effective method (6).

In light of the Daubert ruling (7–9) and other federal court rulings guiding judges on the evaluation of expert witness testimony (10), the methods used to establish the biological profile (i.e., age, sex, ancestry, and stature) require empirical support, estimated error rates, method standardization, and validation of the method through the peer-review process (7,8). The final component of the Daubert guidelines—general acceptance—has been met by morphoscopic traits; however, such acceptance without validation has been unfortunate. Morphoscopic traits were codified most notably in one publication (11), in which Rhine (3) examined 45 cranial morphoscopic traits in four groups (Whites, $n = 53$; Blacks, $n = 7$; Hispanics, $n = 15$; and, Amerindians, $n = 12$) and concluded that morphoscopic traits are useful for predicting race, despite less than ideal sample sizes and rather scant results. Rhine (3) acknowledges that his sample, particularly his sample of American Blacks ($n = 7$ [or three in many instances]), is small, yet his list of expected trait values remains in most forensic anthropology textbooks (6,12–14) and research articles (2,15). In 2007, Hefner et al. (4) published results of an experiment in which they attempted to identify whether and how forensic anthropologists use morphoscopic traits to assess ancestry by exploring several factors, including the analyst's education and experience. At the 2006 AAFS meeting, they conducted a survey and exercise to examine the methodological approaches to ancestry assessment. Exercise participants were asked to rank the techniques they use and to assess the ancestry of seven specimens. A total of 76 individuals participated in the survey, with education levels including Bachelor's, DDS, Master's, MD, and PhD degrees. Participants showed important ambiguities in trait terminology, interpretations, and application to ancestry assessment. Specific traits were referred to using various terms, and participants showed confusion between a trait and a character state for that trait (e.g., nasal aperture morphology, nasal sill, inferior nasal aperture morphology, nasal guttering, partial sill, etc.). In several cases, participants weighted certain traits more heavily than others, despite a declared preference for other traits during the initial survey. In other words, the relative importance of traits in assessing ancestry is adjusted based on the case under examination, likely reflecting *post hoc* trait selection after a general impression is formed, as suggested by Hefner and Ousley (16). Taken as a whole, their results imply that White crania may be more often assessed accurately: 92% of all survey participants correctly assessed ancestry for a White

[1]Joint POW/MIA Accounting Command, Central Identification Laboratory, 310 Worchester Ave. BLDG 45, Joint Base Pearl Harbor-Hickam, HI 96853-5530.

[2]Department of Anthropology/Archaeology, Mercyhurst University, 501 E 38th St., Erie, PA 16546.

*Presented in part during multiple annual meetings of the American Academy of Forensic Sciences.

†Funded in part by a Lucas Research Grant from the Forensic Sciences Foundation.

Received 26 Nov. 2012; and in revised form 3 May 2013; accepted 10 May 2013.

female. Southwestern Hispanics were the most difficult for the participants: only 11% of the participants answered correctly, a pressing problem which may be due to a lack of Hispanics in collections as well as their mixed ancestry.

Clearly, the trait lists presented by Rhine (3) and others fundamentally disregard the reality of human variation. In critically examining the expression of five often-cited traits from Rhine (anterior nasal spine [ANS], nasal aperture width [NAW], inferior nasal aperture morphology [INS], interorbital breadth [IOB], and postbregmatic depression [PBD]), Hefner (17) demonstrated that the percentage of each group presenting all expected trait values from Rhine's (3) trait lists ranges from only 17 to 51 percent. Adding any more "typical" traits would only decrease the percentage. Due to normal human variation, traits can only be used probabilistically to estimate ancestry. There are no absolutes. In fact, even in the trait frequencies Rhine documented, percentages were as low as 0% and only up to "50% or more" for the expected group (3). Of the 45 traits listed, twenty showed at least one trait value expected in another group (3). Rhine suggested idiosyncratic variation and admixture as potential explanations for some of the lower frequencies, but "expected" trait values contradicting the actual frequencies are still reported. As a result of Rhine's influence, admixture is often claimed in papers and publications when human remains show a mixture of traits typical for different groups. We believe that these assertions are unfounded and have led to an invalid method of trait interpretation, namely that a particular trait state defines a group. The trait list approach is insidious, because it will often seem to work due to confirmation bias: When ancestry is known, there will always be traits in a long trait list that are consistent with the ancestry of the current case; the true believer merely chooses the traits *post hoc* (16). The trait list approach lacks scientific rigor and must be abandoned in favor of the scoring and analysis of trait expression.

Our greatest concerns are with the methods, or lack of methods, for objectively scoring and analyzing the morphoscopic traits, rather than with the traits themselves. Hefner (17) standardized scoring for eleven of the more common morphoscopic traits, an important first step in their use. Categorical data do not follow the same statistical distribution as continuous data, so specialized analytical methods are necessary to provide estimated error rates, termed "potential" error rates in the Daubert decision (18,19). Methods like the Rubison procedure (20), which take into account asymmetric cranial nonmetric trait expression, do not account for correlations among variables and are not as accessible as the more common classification methods used for cranial measurements, for example, the discriminant function analyses found in Fordisc (21). Robust and appropriate methods for categorical data have only recently been established (5), and the important products of applied statistical methods to morphoscopic data are explicit: large sample sizes, posterior probabilities for a specific case identifying how strongly an individual classifies, and estimated error rates.

The objective of this paper is to examine the utility of several of the more frequently cited morphoscopic traits through appropriate and novel statistical methods for classifying an unknown cranium into one of several reference groups.

Materials and Methods

Morphoscopic traits outlined in Hefner (17) were documented for 718 adult individuals curated at the National Museum of Natural History (NMNH), Smithsonian Institution, Washington, DC; the William M. Bass Donated Skeletal Collection, the Department of Anthropology, University of Tennessee, Knoxville, TN; and the Pima County Medical Examiner's Office, Tucson, Arizona. Table 1 presents the composition of the sample used herein, but for a more thorough discussion on the sample, see Hefner (17,22). Trait descriptions, illustrations, and character state frequency distributions are presented in Hefner (17) and Hefner et al. (23). To maximize sample sizes, males and females were pooled for analysis. Hefner (24,25) noted no differences in morphoscopic trait expression, with the exception of postbregmatic depression (PBD), which was slightly more common (although not significantly) in American Black females. All individuals used in this study are of adult age (16–99 years). This geographically and temporally diverse sample was selected to cover the range of casework seen in most forensic anthropology laboratories in the United States and to document morphoscopic traits in a sample of Hispanic individuals.

Analytical Approaches

The optimized summed scored attributes method, or OSSA, is a heuristic method using six morphoscopic traits (ANS, INA, IOB, NAW, NBC, and PBD) scored following Hefner (17). Each trait score is dichotomized (e.g., 1,2,3,4 transformed to 0, 1) to maximize the between-group differences in two populations, in this case American Blacks and Whites. A heuristic optimization of the compressed trait scores is accomplished by ordering the variables in a manner that maximizes the differences between groups. For this study, scores more common in American Blacks were scored "0", and those more common in American Whites to a score of "1". For example, the OSSA compression for the inferior nasal aperture (INA) morphology is presented in Table 2. The original character states are optimized in a manner that most efficiently separates the two values (0,1) between the two groups (between character states 3 and 4 for INA). Using this trait alone, we can expect to be correct almost 80% of the time for American Blacks and 72% for American Whites. Once all six traits have been converted into their new binary variable, the sum of all traits is calculated, producing individual OSSA scores ranging from 0 to 6. The distribution of OSSA scores is presented in Table 3 and is illustrated in Fig. 1.

Morphoscopic traits of an unknown cranium can now be scored using the original character states (original ordinal categories) and compressed to their respective OSSA state (Fig. 2). After compression, the scores are summed and the resulting value is compared with the sectioning point of 3 and below for American Blacks, and 4 and above for American Whites.

Ten additional multivariate statistical classification methods that provide multi-group classifications were also tested using the morphoscopic dataset. These methods each generally take correlations among variables into account and provide estimated accuracy rates. All of these methods have different requirements

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

	Sex			
Group	Male	Female	Indeterminate	Total
Black	169	87	0	256
Hispanic	106	37	101	244
White	98	120	0	218
Total	373	244	101	718

TABLE 2—*Distribution of the inferior nasal aperture (INA) with original ordinal scores and OSSA scores.*

Original Score	American Black $n = 218$ Cum.%	n	%	American White $n = 146$ n	%	Cum.%	OSSA Score
1	29.3	64	29.3	1	0.7	100	0
2	58.2	63	28.9	5	3.4	99.3	0
3	79.8	47	21.6	35	24	95.9	0
4	93.1	29	13.3	60	41.1	71.9	1
5	100	15	6.9	45	30.8	30.8	1

TABLE 3—*Distribution of OSSA summed scores between American Blacks and American Whites.*

	OSSA Score							
Ancestry	0	1	2	3	4	5	6	Total
Black	20	57	36	29	17	3	2	164
White	0	1	6	10	31	45	23	116
Total	20	58	42	39	48	48	25	280

for optimal classification accuracy of individuals, and some methods require that the data have a multivariate normal distribution. Discrete data such as nonmetric trait scores would likely fail most tests for multivariate normality.

Discriminant methods are intuitive because an unknown individual will be classified into the group to which it is most similar based on group means, and relationships can be easily graphed. Both linear discriminant function analysis (LDFA) and quadratic discriminant function analysis (QDFA) have a requirement that the discriminant scores are more or less normally distributed, which is most always the case when the original data are normally distributed, as with many skeletal measurements. Using ordinal data with few classes, such as morphoscopic data, is clearly bending or even breaking the rules, but nonetheless, the classification performance can be very good; however, caution must be exercised in interpreting the posterior probabilities. Ordinal data can often be made more normally distributed by converting them to principal component scores. LDFA is the best known discriminant technique and classifies best when the level of variation is more or less the same in all groups. When this is not true, QDFA can be used, but classification accuracy is often lower than in LDFA, and the sample size requirements for QDFA are higher. Another option, when the data are not normally distributed, is a semi-parametric method known as logistic regression (LR). Logistic regression directly estimates probabilities of group membership and has fewer requirements than DFA: the data need not be normally distributed and the level of variation in each group can differ (26). Naïve Bayesian is a probabilistic classifier based on Bayes theorem in the form of a conditional model. Naïve Bayesian analysis assumes independence of the features of interest.

Two additional methods of statistical classification are termed nonparametric, because they use individual similarities rather than group similarities to classify unknown individuals. K-nearest-neighbor analysis (kNN) classifies an individual based on the k most similar individuals from known reference samples, often using a majority rule. For instance, if the ancestries of the three most similar individuals to an unknown individual are two Black individuals and one Hispanic, then the kNN method would classify the unknown individual as a Black. Kernel probability density (KPD) classifies an unknown individual using a probability density function calculated from reference group individuals. An unknown individual is classified into the reference group with the highest cumulative probability calculated from the group's individuals.

The newest classification methods depend on the speed and power of computers and are generally called "machine learning" methods. They include such methods as artificial neural networks (aNNs), decision trees (DTs), random forest models (RFMs), and support vector machines (SVMs), which do not work directly with group parameters such as means and standard deviations. Machine learning methods involve tuning thousands of random cutoff points in the sample to find the most accurate ways of pooling individuals into groups (27,28). However, in machine learning, there are many possible data transformations that can be applied to the original data, which can affect classification accuracy. Machine learning methods have been used to varying levels of success in biological anthropology (29–32), but only recently, and never to our knowledge, for ancestry estimation.

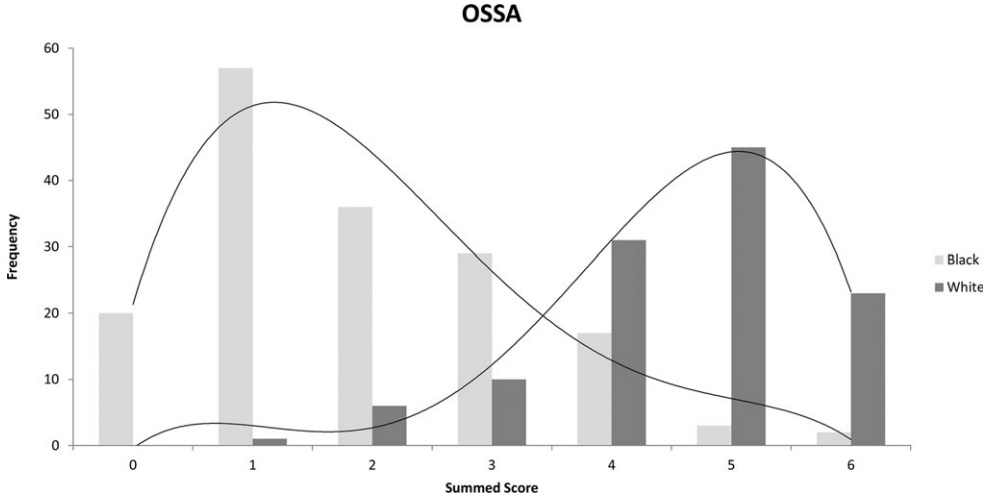

FIG. 1—*Distribution of OSSA scores in a sample of American Blacks and American Whites.*

Anterior Nasal Spine (ANS)				**Nasal Aperture Width (NAW)**			
State	OSSA			State	OSSA		
1	0	Slight		1	1	Narrow	
2	1	Intermediate		2	1	Medium	
3	1	Marked	ANS State =	3	0	Broad	NAW State =
			OSSA SCORE =				OSSA SCORE =
Inferior Nasal Aperture (INA)				**Nasal Bone Structure (NBS)**			
State	OSSA			State	OSSA		
1	0	Pronounced Slope		0	0	Low/Round	
2	0	Moderate Slope		1	0	Oval	
3	0	Straight		2	1	Marked Plateau	
4	1	Partial Sill		3	1	Narrow Plateau	
5	1	Sill	INA State =	4	1	Triangular	NBS State =
			OSSA SCORE =				OSSA SCORE =
Interorbital Breadth (IOB)				**Post-Bregmatic Depression (PBD)**			
State	OSSA			State	OSSA		
1	1	Narrow		0	1	Absent	
2	1	Intermediate		1	0	Present	
3	0	Wide	IOB State =				PBD State =
			OSSA SCORE =				OSSA SCORE =
							SUMMED OSSA SCORE =
Race Assessment: _____ Percent Correct (for sample): _____ %							

FIG. 2—*OSSA scoring sheet.*

Artificial neural networks use a search algorithm to examine multiple subsets of the predictor variables with various random weights assigned to each (this avoids finding locally optimal results that cannot be generalized to individuals outside of the training set). A large number of these "competing" models are generated and then compared with assess which model best fits the data. Using the prediction and the actual outcome to assign relative variable weights is known as forward feedback propagation, one of the most common aNNs and the method we applied in this study. Decision trees are also known as classification and regression trees and employ a sequential series of rules to estimate group membership starting with the most effective rule, or node, that separates individuals into two or more subsamples that are most accurately classified according to group membership. For instance, for these data, the rule that is most effective at dividing the total sample into the three original groups is "INA < 1.5". This rule divides the entire sample into two "branches", one branch having individuals with an inferior nasal aperture score ≤ 1.5, and one branch having individuals with an INA score > 1.5. Further nodes divide the branches, sometimes using rules that split a branch into more than two branches, until the divisions cannot separate the original groups any better. Figure 3 provides an example of such trees. Each rule is assessed from top to bottom; when a terminal node is reached, the classification for the unknown is presented in bold above the frequency of each ancestry in that node. For example, in the terminal node below INA (<1.5), 96 American Blacks, 9 Hispanics, and 3 Whites were all classified as Black (see "B" in Fig. 3). Random forest classification uses many random subsets of the variables and repeated sampling of the original data to produce hundreds of decision trees, called an ensemble, and the consensus of the ensemble is used to determine the best classification rules. Random forests can generally tolerate a large number of variables simultaneously, including "noisy" ones. Support vector machine classification identifies boundaries between individuals that lie near boundaries separating groups and then manipulates variable weights to produce linear boundaries that best separate those individuals from different groups. Support vector machines can classify especially well when there are many variables with nonlinear relationships to each other.

In evaluating statistical classification functions, especially in the forensic context, the most important measure is the classification accuracy rate, which is estimated from the known reference samples. Classification accuracy is the best measure of a statistical method's validity. Because estimates of classification accuracy are biased when the same individuals are used to calculate and evaluate classification performance, statistical methods must incorporate objective validation procedures. The most widely used validation procedure in traditional classification methods is leave-one-out cross-validation (33), whereby each individual in the total reference sample is sequentially removed and the classification methods are followed using the rest of the sample and applied to the removed individual. The accuracy rate is calculated from the classifications of each removed individual. Logistic regression classification accuracy is often based on an algorithm and can be biased optimistically (34). Machine learning classifiers divide the total sample into a training or calibration sample to calculate rules for classification, usually between 50% and 70% of the total sample; a tuning or testing sample for further adjustment of classification rules, between 5% and 10% of the total sample; and a holdout sample, a completely separate validation sample to estimate accuracy, between 10% and 30% of the total sample. For these results, the training, test, and validation samples were 70%, 10%, and 20% of the total sample. Additionally, there are a great many data transformations and method-specific options which can be adjusted that influence the performance of each method (28). Machine learning methods must employ more rigorous cross-validation methods, because they will otherwise report optimistically biased results. As a

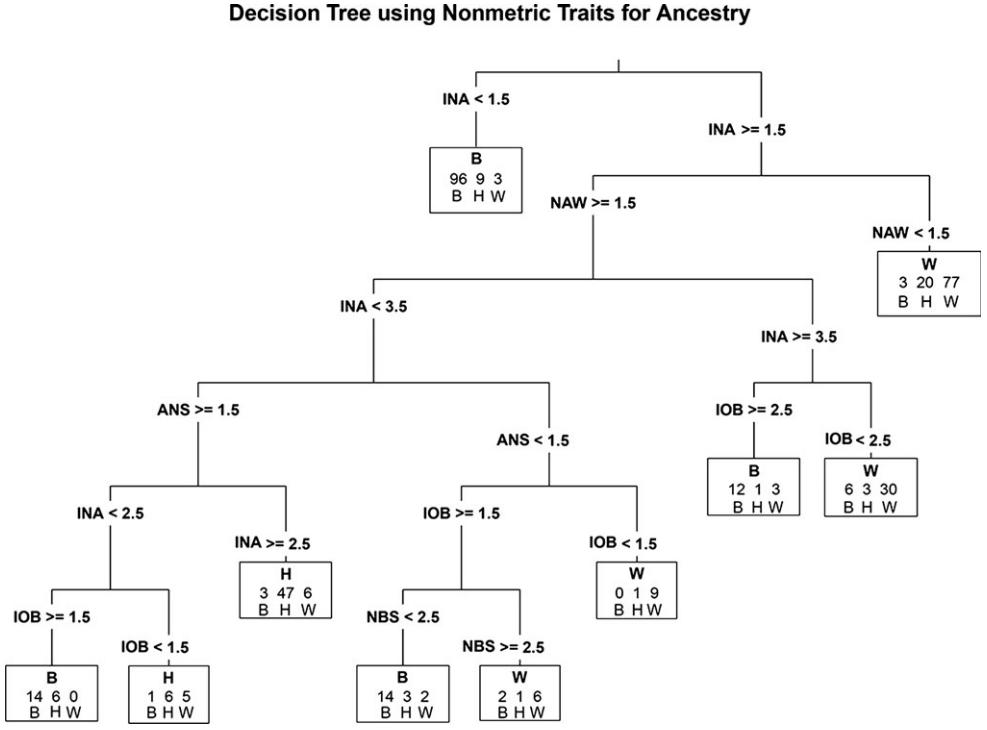

FIG. 3—*Decision tree using all five nonmetric traits to classify into ancestry. The total sample is divided into bifurcating branches according to the listed criteria. The ends of the branches are known as leaves and are shown as boxes with the number of individuals classified in the training sample into that leaf from each group, (B) Black, (H) Hispanic, and (W) White. Bold text above numbers represents the group into which an unknown is classified.*

result, generally, they need somewhat larger sample sizes to best estimate classification accuracy.

The overall correct classification rate is important, but so are the correct rates for each group. In multi-group analyses, certain groups will often be more similar to each other than to other groups, so maximizing correct classification rates may be challenging. Some groups classify at higher rates than other groups, and some groups may tend to be misclassified into specific groups. An additional challenge comes from the fact that using fewer variables will often classify reference groups better than using more variables, most often due to overfitting data (having too many variables relative to individuals), or because some variables do not differ among groups and are considered "noise" as far as classification is concerned. Fordisc 3.1 (19), R (35), SAS version 9.0 (36), and SYSTAT for Windows version 11 (37) were used for all statistical analyses.

Results

Classification results for all methods are shown in Table 4. The best overall methods are aNN, OSSA, SVM, and RFM, with mean classification accuracies of at least 85%. Of these methods, the aNNs show that highest overall classification rate (87.8%) and random forests show the smallest difference between the highest (90.4%) and lowest (76.5%) classification accuracies. Logistic regression shows the lowest overall classification accuracy, because each group shows the lowest classification accuracy of all methods. Every method but the kernel method shows the highest classification percentage for Blacks and the lowest classification percentage for Hispanics. The kernel method classifies Hispanics the most accurately and classifies them more accurately than Blacks or Whites. The OSSA method classifies Blacks and Whites about as well as the three-group analyses. The decision tree rules are easy to follow for these data and are illustrated in Fig. 3. We can follow the classification decisions by observing the bifurcating branches corresponding to the tests at each node; each classification is known as a leaf and is shown by a box. At the very top, the first sectioning point, we branch left if the INA score is 1, and classify the individual as Black; that leaf classified 96 Black, 9 Hispanic, and 3 White individuals from the test sample as Black. If the INA score is higher than 1, we proceed further depending on whether the NAW score is 1 or more than 1. Following these branches results in classifications for an unknown individual, and the estimated classification accuracy comes from the holdout sample.

Discussion

Overall, morphoscopic traits have some advantages over metric analysis. For instance, they are easily observed and do not require instruments (with the exception of a contour gauge); the key is having a consistent scoring strategy, which is now available (17). The next step—making the most of the data in statistical classification—has been explored here.

The results of this research demonstrate that morphoscopic traits can be used to assess ancestry accurately/successfully, with estimated error rates and without relying only on the experience of the observer. Moreover, these results demonstrate that subjective trait lists do not capture the true range of variation.

The OSSA method is a simple method with several weaknesses. First, OSSA does not adequately take the multivariate relationships among traits into account, because each trait is given equal weight in the OSSA score. Some traits help classify better, and joint distributions of traits generally provide better classification information than single traits. More sophisticated methods of analyzing traits are needed and are provided in the

TABLE 4—*Classification accuracies using OSSA and ten classification methods, ordered by decreasing accuracy.*

	Neural Network			Support Vector Machine			OSSA			Random Forest		
Group	Black (n = 256)	Hispanic (n = 244)	White (n = 218)	Black (n = 225)	Hispanic (n = 125)	White (n = 193)	Black (n = 164)	Hispanic NA	White (n = 116)	Black (n = 225)	Hispanic (n = 125)	White (n = 193)
Black	95.9	2.5	1.6	90.4	9.6	0	86.6	–	13.4	90.4	9.6	0
Hispanic	6	76.8	17.2	17.6	64.7	17.6	–	–	–	5.9	76.5	17.6
White	5.7	3.5	90.8	2.4	7.3	90.2	14.7	–	85.3	2.4	14.6	82.9
% Correct	87.8			86.4			86.1			85.5		

	Naïve Bayesian			Decision Tree			k-Nearest Neighbor			Quadratic DFA		
	Black (n = 164)	Hispanic (n = 95)	White (n = 116)	Black (n = 225)	Hispanic (n = 125)	White (n = 193)	Black (n = 225)	Hispanic (n = 125)	White (n = 193)	Black (n = 225)	Hispanic (n = 125)	White (n = 193)
Black	82.9	6	14.6	91.1	2.2	6.7	86.2	8	5.8	83.6	10.2	6.2
Hispanic	3.1	65.6	31.3	16.7	58.3	25.0	13.6	64.8	21.6	12.8	66.4	20.8
White	12.9	6	81.1	4.3	15.2	80.4	7.8	16.6	75.7	4.7	17.6	77.7
% Correct	80.4			80.0			77.5			77.5		

	Kernel Probability Density			Linear DFA			Logistic Regression		
	Black (n = 225)	Hispanic (n = 125)	White (n = 193)	Black (n = 225)	Hispanic (n = 125)	White (n = 193)	Black (n = 225)	Hispanic (n = 125)	White (n = 193)
Black	81.78	13.78	4.44	82.2	12.4	5.3	80.4	11.8	7.9
Hispanic	8.8	83.2	8	8.8	65.6	25.6	20.6	41.6	37.8
White	5.18	27.98	66.84	4.1	25.9	69.9	9.5	24.1	66.3
% Correct	76.8			74.0			66.4		

other methods examined. Second, preliminary research suggests that postbregmatic depression may not be very useful for classification purposes (38). Third, all six traits must be present for the analysis, so fragmentary crania missing the mid-facial skeleton or the vault should not be scored. Fourth, because the OSSA scores are discrete, limited, and overlap between the groups, the choice of the sectioning point – the score to decide whether to classify into White or Black – determines which group will be classified lower, because many individuals are near the sectioning point; however, we found that a score ≤ 3 for Black is the optimal cutoff value for this sample, and an initial independent validation study supports that cutoff as well (39). The cutoff point in the OSSA method (3 and below for American Blacks) is important, because it indicates that even with a binary scoring system, only half of the traits that are "supposed" to be present in American Blacks (as described by trait lists) are actually present among American Blacks. Fifth, the OSSA method does not calculate posterior probabilities, so it is difficult to assess how strong the classification is, although the associated sensitivity, specificity, and predictive values may be used. Finally, at this time, the OSSA method is limited to American Blacks and Whites, but the method is still useful, particularly in the United States where the majority of forensic casework involves these two groups. The OSSA method looks quite promising for the bulk of forensic cases encountered in the United States and is well suited for work at the Joint POW/MIA Accounting Command Central Identification Laboratory, where nearly 90% of the analyzed remains are either Black or White Americans.

Morphoscopic data require different statistical approaches compared with metric data. Fortunately, there has been substantial recent progress in the analysis of categorical data. Using more recent methods on ordinal morphoscopic data more accurately classifies an unknown than "bending the rules" when applying LDFA to the data. However, LDFA performed better than logistic regression, which has looser assumptions. Using the artificial neural network is preferred due to its higher accuracy, but several of the other methods perform nearly as well. There are plenty of choices when analyzing categorical data.

In this paper, we explored eleven approaches to statistical classification using morphoscopic traits. Several of the methods are quite recent, making the most of computing power, and more methods are sure to come. In fact, even now, there is a seemingly unlimited number of ways the data can be transformed for further analysis, and ensemble methods – which employ several machine learning methods for a consensus – are being developed. The best way to employ these methods and morphoscopic traits is to integrate them into a computer program, which is planned. No matter which statistical method is used, the differences between these methods and the trait list method as embodied in Rhine (3) highlight the value of the Daubert guidelines: Our group sample sizes are much larger; we describe explicit methods of analysis and classification rather than general patterns of variation; the criterion of general acceptance has been met in the statistical literature using other data; and estimated error rates for each statistical method provide estimates of validity. The Daubert guidelines also remind us how to choose among new methods based on their validity. Classification accuracy is the best measure of validity, as long as sample sizes are large and representative, and interobserver concerns are thoroughly addressed (i.e., reliability is established).

We have also illustrated that scoring trait expression is much more valuable than looking for specific traits to estimate ancestry accurately. Because there is variation in trait expression and traits are more valuable in combination, statistical methods are necessary to optimize accuracy. Trait scores analyzed using optimized and objective statistical methods provide far greater accuracy in estimating ancestry.

The results of this study suggest that classifying an individual to "mixed" ancestry based on discordant trait values is likely incorrect. Thus, what are forensic anthropologists to do when

confronted with Hispanics—a population often described as a hybrid group evincing morphologies shared between American Whites, Native Americans, and Africans—if these seemingly isolated populations also present discordant trait values? The artificial neural network and kernel methods classify Hispanics very well, others do not, but as with any classification statistic, an appropriate reference sample is needed. Individuals of mixed ancestry, those with different social race labels for each parent, are still relatively rare, but are becoming more common as interracial marriage is practiced more and more. Whether these traits will show blending or dominance in skeletal tissues is currently unknown.

Conclusions

The statistical methods presented above account for variation in trait expression in individuals as well as differences in empirical trait frequencies among groups and have cross-validated classification accuracies nearing 90%. Individuals show a "mosaic" of trait expressions, rather than extreme trait values seen in older trait lists; classification statistics account for group trait frequencies and the relationships among those traits within groups to derive better classification rules for individuals. The benefits of a statistical approach to nonmetric traits are obvious. First, the importance placed on the subjective experience of the observer is reduced, an attractive attribute in light of the Daubert ruling. Secondly, the calculated variable weights and sectioning points in analyses optimize classification accuracies, while accounting for the relationships among variables and the true nature of biological variation. Finally, sample size and composition are known and should express the underlying range of variability in the populations.

Morphoscopic trait analysis remains an essential element in the assessment of ancestry because of the emphasis, and importance forensic anthropologists have historically placed on these slight variations in cranial form. However, when the actual distribution of these traits is understood, the discordance of multiple traits should come as no surprize and should not be treated as evidence of admixture. On the contrary, trait discordance is evidence against the typological approach to ancestry prediction and represents the true nature of the distribution of morphoscopic traits among human groups.

Now is a critical time in forensic anthropology. Rigorous validation of forensic anthropological techniques and methods is needed more than ever. The *Daubert* decision does not technically apply outside of the courtroom, but following the *Daubert* guidelines when assessing ancestry and other aspects of the biological profile will increase the probability of drawing correct conclusions. In any event, samples will be described, well-described methods should be replicable, and error rates will be estimated. Good science dictates understanding the inherent validity, or accuracy, associated with a method. We must evaluate each method for accuracy and choose the most accurate method. Finally, methods with unknown accuracies and clear problems due to confirmation bias, like the traditional trait lists, should not be used (21,40).

The classification methods evaluated here have a firmer empirical basis than the traditional approach used in ancestry assessment. There is a definite need for scientific rigor in ancestry assessment methods that are empirically supported by the data at hand. The validity of any method, whether novel or long established, should be tested and refined rather than performed simply due to tradition or subjective/personal experience.

Acknowledgments

This research benefited greatly from discussions with Richard Jantz, Dennis Dirkmaat, Kate Spradley, Paul Emanovsky, and Nick Passalacqua. We would also like to thank Lee M. Jantz (Department of Anthropology, University of Tennessee), David Hunt (National Museum of Natural History, Smithsonian Institution), and Bruce Anderson (Pima County MEO) for providing access to skeletal collections. We would also like to acknowledge Kandus Linde for editorial comments on draft versions of this manuscript.

References

1. Bass WM. Human osteology: a laboratory and field manual, 3rd edn. Columbia, MO: Missouri Archaeological Society, 1987.
2. Hughes C, Juarez CA, Hughes TL, Galloway A, Fowler G, Chacon S. A simulation for exploring the effects of the "trait list" method's subjectivity on consistency and accuracy of ancestry estimations. J Forensic Sci 2011;56(5):1094–106.
3. Rhine S. Morphoscopic skull racing. In: Gill GW, Rhine S, editors. Skeletal attribution of race: methods for forensic anthropology. Albuquerque, NM: Maxwell Museum of Anthropology, 1990;7–20.
4. Hefner JT, Emanovsky PD, Byrd J, Ousley SD. The value of experience, education, and methods in ancestry prediction. Proceedings of the 59th Annual Meeting of the American Academy of Forensic Sciences, 2007 Feb 19–24; San Antonio, TX. Colorado Springs, CO: American Academy of Forensic Sciences, 2007.
5. Krzanowski WJ. Principles of multivariate analysis: a user's perspective. London, U.K.: Oxford University Press, 2002.
6. Klepinger LL. Fundamentals of forensic anthropology. Hoboken, NJ: Wiley-Liss, 2006.
7. Christensen A. The impact of *Daubert*: implications for testimony and research in forensic anthropology (and the use of frontal sinuses in personal identification). J Forensic Sci 2004;49:1–4.
8. Christensen AM, Crowder CM. Evidentiary standards for forensic anthropology. J Forensic Sci 2009;54(6):1211–216.
9. Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 1993.
10. Grivas CR, Komar DA. *Kumho*, *Daubert*, and the nature of scientific inquiry: implications for forensic anthropology. J Forensic Sci 2008;53 (4):771–6.
11. Gill GW, Rhine S. Skeletal attribution of race: methods for forensic anthropology. Albuquerque, NM: Maxwell Museum of Anthropology, 1990.
12. Burns K. Forensic anthropology training manual. Upper Saddle River, NJ: Prentice-Hall, 1999.
13. Byers SN. Forensic anthropology. Boston, MA: Pearson Education LTD, 2008.
14. Reichs K. Forensic osteology: advances in the identification of human remains, 2nd edn. Springfield, IL: Charles C. Thomas, 1998.
15. L'Abbé EN, Van Rooyen C, Nawrocki SP, Becker PJ. An evaluation of non-metric cranial traits used to estimate ancestry in a South African sample. Forensic Sci Int 2011;209:1–3.
16. Hefner JT, Ousley SD. Morphoscopic traits and the statistical determination of ancestry II. Proceedings of the 58th Annual Meeting of the American Academy of Forensic Sciences, 2006 Feb 20–25; Seattle, WA. Colorado Springs, CO: American Academy of Forensic Sciences, 2006.
17. Hefner JT. Cranial nonmetric variation and estimating ancestry. J Forensic Sci 2009;54(5):985–95.
18. Christensen AM, Crowder CM, Ousley SD, Houck MM. Error and its meaning in forensic science. J Forensic Sci 2014;59(1):123–26.
19. Ousley SD, Hollinger E. The pervasiveness of *Daubert*. In: Dirkmaat DC, editor. A companion to forensic anthropology. Malden, MA: Wiley-Blackwell, 2012;654–65.
20. Finnegan M, McGuire SA. Classification systems for discrete variables used in forensic anthropology. Am J Phys Anthropol 1979;51:547–53.
21. Jantz RL, Ousley SD. FORDISC 3: computerized forensic discriminant functions. Version 3.1.292. Knoxville, TN: The University of Tennessee, 2005.
22. Hefner JT. Cranial morphoscopic traits and the assessment of American Black, American White, and Hispanic Ancestry. In: Ta'ala S, Berg G, editors. Biological affinity in forensic identification of human skeletal remains: beyond black and white. Boca Raton, FL: Taylor & Francis Group, LLC. In press.

23. Hefner JT, Dirkmaat DC, Ousley SD. Morphoscopic traits and the assessment of ancestry. In: Dirkmaat DC, editor. A companion to forensic anthropology. Malden, MA: Wiley-Blackwell, 2012;287–310.
24. Hefner JT. Assessing morphoscopic cranial traits currently used in the forensic determination of ancestry [thesis]. Gainesville (FL): The University of Florida, 2003.
25. Hefner JT. The statistical determination of ancestry using cranial nonmetric traits [dissertation]. Gainesville (FL): The University of Florida, 2007.
26. Tabachnick BG, Fidell LS. Using multivariate statistics, 5th edn. Boston, MA: Allyn and Bacon, 2001.
27. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York, NY: Springer, 2001.
28. Williams G. Data mining with rattle and R. New York, NY: Springer, 2011.
29. McBride DG, Dietz MJ, Vennemeyer MT, Meadors SA, Benfer RA, Furbee NL. Bootstrap methods for sex determination from the os coxae using the ID3 algorithm. J Forensic Sci 2001;46(3):427–31.
30. Konigsberg LW, Herrmann NP, Wescott DJ. Commentary on: McBride DG, Dietz MJ, Vennemeyer MT, Meadors SA, Benfer RA, Furbee NL. Bootstrap methods for sex determination from the os coxae using the ID3 algorithm. J Forensic Sci 2001;47(2):424–7.
31. Moore MK, Schaefer E. A comprehensive regression tree to estimate body weight from the skeleton. J Forensic Sci 2011;56(5):1115–22.
32. Love JC, Derrick SM, Wiersema JM, Peters C. Validation of tool mark analysis of cut costal cartilage. J Forensic Sci 2012;57(2):306–11.
33. Lachenbruch PA, Mickey MR. Estimation of error rates in discriminant analysis. Technometrics 1968;10:1–11.
34. Peng C, So T. Logistic regression analysis and reporting: a primer. Understanding Statistics 2002;1(1):31–70.
35. R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2011.
36. SAS Institute Inc., SAS 9.1.3 Help and documentation. Cary, NC: SAS Institute Inc., 2000-2004.
37. SPSS, Inc. SYSTAT 11. Chicago, IL: SPSS, Inc., 2009.
38. Ousley SD, Hefner JT. The statistical determination of ancestry. Proceedings of the 58th Annual Meeting of the American Academy of Forensic Sciences, 2005 Feb 21–26; Seattle, WA. Colorado Springs, CO: American Academy of Forensic Sciences, 2005.
39. Emanovsky PD. Analytical test method selection and validation of laboratory-based methods. Proceedings of the 65th Annual Meeting of the American Academy of Forensic Science, Washington DC, 2013 Feb 18–23; Washington, DC. Colorado Springs, CO: American Academy of Forensic Sciences, 2013.
40. Ousley SD, Billeck WT, Hollinger RE. Federal repatriation legislation and the role of physical anthropology in repatriation. Yearb Phys Anthropol 2005;48:2–32.

Additional information and reprint requests:
Joseph T. Hefner, Ph.D.
JPAC - Central Identification Laboratory
310 Worchester Avenue, Bldg. 45
Joint Base Pearl Harbor-Hickam, HI 96853-5530
E-mail: joseph.hefner@jpac.pacom.mil