Navega D, Coelho C, Vicente R, Ferreira MT, Wasterlain S, Cunha E. 2015. AncesTrees: ancestry estimation with randomized decision trees. International Journal of Legal Medicine, 129:1145–1153 (DOI 10.1007/s00414-014-1050-9).

Introduction

Forensic anthropology plays a major role in the analysis and identification of human skeletal remains helping to narrow the field of inquiry, facilitating a match between human remains and missing person report. A first step of this process is to establish the biological profile of an individual, namely, its sex, age-at-death, ancestry and stature. Continued use of race or ancestry in forensic anthropology is controversial and has been criticized mainly due to the fact that strong scientific evidences disprove the biological race concept as constructed within traditional physical anthropology. Nevertheless, ancestry is an important identification parameter and forensic anthropologists can estimate it because there is some degree of concordance between social constructions of ancestry and skeletal morphology [1].

Several methodological approaches can be used in anthropological ancestry estimation. One of the most recent and innovative, is the application of geometric morphometrics [2], which allows for rigorous and objective analysis of the cranial shape through 3D coordinate data. Several studies have also been recently conducted to put classical tools such as cranial morphoscopic and dental morphology analysis in a more scientific framework [3–10]. One of the most well-known and established methods of ancestry estimation relies on craniometry and application of statistical classification techniques. Introduced by Giles and Elliot [11], ancestry estimation by means of linear discriminant analysis (LDA) aims to classify a target individual (e.g. unidentified), through weighted combination of craniometric variables, into a reference ancestral group. Computer applications like FORDISC [12, 13] and CRANID [14, 15] make the use of LDA for ancestry estimation straightforward, and useful allocation indicators such as posterior and typicality probabilities are outputted to help interpretation and decision making.

When correctly applied and interpreted LDA is a valuable technique to perform ancestry estimation from metrical data [13]. However, even when all of its assumptions are met, LDA can normally be outperformed by modern machine learning classification algorithms [16–18].

A pioneering study by Hefner et al. [4, 9] illustrates the utility of a machine learning technique called random forest modeling to estimate ancestry from cranial metrical and morphoscopic data. Unlike LDA, random forest allows for this opposing type of approaches to be integrated in a robust computational way. Using 110 skulls from the William M. Bass Donated Skeletal Collection representing modern American Whites (n=72) and Blacks (n=38), and 39 skulls representative of Southwestern Hispanics from the Pima County Medical Examiner's Office (Tucson, Arizona), they obtained an accuracy rate of 89% (cross-validated).

Inspired by Hefner et al. [4, 9] and by the major contributions of FORDISC and CRANID developers to forensic anthropology, in this paper we present a computer program named, AncesTrees, developed and evaluated for ancestry estimation using the *random forest algorithm* as underlying classification technique.

Material and Methods

Material

Reference dataset

As reference sample 1734 individuals from the Howells Craniometric Series [19–22] were selected to integrate our program database (Table 1). Individuals were pooled in six ancestral groups: African, Austro-Melanesian, East Asian, European, Native American and Polynesian. A total of 23 craniometric variables were used in our approach (Table 2). The selected measurements describe overall cranial morphology (length, width, height), and specific regions of facial skeleton (e.g. nasal aperture, orbits). All can be easily collected using only spreading and sliding calipers.

Test datasets

Although our computer program have implemented a specific mechanism of validation of the model generated in each execution (see next section), the most reliable accuracy test will always be the one conducted on an independent test dataset. Therefore, we tested our program in 128 adult skulls, belonging to 32 individuals of African ancestry and 96 of European ancestry.

The individuals of African ancestry were selected from the African slaves' skeletal collection of *Valle da Gafaria*. These skeletal remains belong to African slaves discarded during the 15th-17th centuries in a waste disposal site at Lagos (Portugal) [23]. All individuals show Negroid characteristics at morphoscopic level [24]. Data was collected by one of the authors (CC).

The 96 individuals of European ancestry were selected from two Portuguese identified osteological collections. Sixty four individuals (32 males and 32 females) were randomly selected from the *Colecção de Crânios Identificados das Escolas Médicas* (Medical School Skull Collection). This collection, housed in the University of Coimbra, is composed by 585 skulls of individuals who died between 1895 and 1903 [25]. Data was collected by one of the authors EC [26].The remaining 32 individuals (18 males and 14 females) were randomly selected from the recently created *Colecção de Esqueletos Identificados Século XXI* (Identified Skeletal Collection of the 21st Century), whose individuals died between 1995 and 2007, and were exhumed between 2003 and 2013 [27]. Data was collected by two authors MTF and CC.

Descriptive statistics (mean, standard deviation and range) of the cranial measurements were computed in order to provide a characterization of test samples used in this study.

Methods

The random forest algorithm

This study evaluated the utility of the random forest algorithm [28] in ancestry estimation based on craniometric data. The random forest algorithm [28] belongs to a class of machine learning techniques denominated ensemble learners [29, 30]. An ensemble is a set of classifiers whose results are combined to produce a final classification. A key aspect to obtain a good ensemble is to have accurate and diverse classifiers. The random forest uses tree-structured classifiers as base learners to grow an ensemble. Classification trees are a simple, yet powerful, type of non-parametric classification model, constructed using a recursive top-down divide-and-conquer strategy, capable of learning complex interactions between variables and non-linear decision boundaries. Tree-based models can handle any type of input variable (numeric, nominal, ordinal), tackle multi-class problems, and automatically select the most relevant variables. Albeit very accurate, classification trees are very prone to over-fitting. Due to the nature of their training process, minor changes in the training data lead to different trees and consequent results, especially when trees are not pruned or simplified. For this reason, tree models are considered unstable classifiers [31, 32].

The random forest algorithm not only copes with the aforementioned problem of classification trees, but also exploits the instability of tree-based models to build a powerful ensemble composed of accurate and diverse classifiers. To generate diversity two stochastic mechanisms are employed during the forest growth. First, each tree, instead of the original training set, is induced using a bootstrap replicate of it. This concept is adapted from a previous ensemble learning technique designated bootstrap aggregation or bagging [33]. A bootstrap replicate is a sample drawn by re-sampling, with replacement, the original data. Each replicate have the same size of the original set and contains approximately 63% of the original data. This process makes each induced tree slightly different from each other and combining the generated trees generally leads to better results than using a single model. The second stochastic mechanism involves the process of tree construction itself. Normally, at each node of a tree all variables are tested to select the most relevant one to split the data and create two more branches for the tree. In the random forest algorithm, a subset of randomly selected variables is used instead. This key aspect of the random forest was influenced by previous works on tree randomization and stochastic discrimination [34–37]. The random forest algorithm uses classification trees grown to their full depth.

Unlike many other machine learning techniques that have many hyper-parameters that should be carefully tuned to ensure maximum performance, only two parameters should be specified in random forest model. One should specify the number of variables (p) to be randomly selected and evaluated in each split and the number of trees (T) to include in the ensemble. Many software packages use simple heuristic rules to select p and the user only needs to specify T.

The random forest algorithm is one of the most powerful and versatile algorithms in modern machine learning. It can tackle classification and regression tasks in a supervised learning framework, it can be used for data clustering, missing value imputation and novelty and outlier detection from an unsupervised learning perspective. One of its most interesting features is the fact that no cross-validation is required to get unbiased estimates of the model performance. The random forest model inherits from bagging the ability to produce out-of-bag error estimates. As mentioned before a bootstrap replicate of the data contains two-thirds of the original data, which means that about one-third is not used in the classification trees. For each example of the original data, an out-of-bag classification is produced by combining only the classification trees where that example was not present in the bootstrap replicate. Out-of-bag error provide a good and unbiased estimate about the generalization ability of the random forest and compares to K-fold cross-validation [32].

From tree-based classifiers random forest ensembles inherits the ability of automatically select the most discriminant predictors and also measure their importance for the model. One common measure of variable importance used in random forest is the mean decrease of Gini criterion. Gini criterion is a splitting criterion used to evaluate the quality of a variable when a new branch is to be created in a tree. Whenever a the value of the Gini criterion of a node's descent nodes is less than the node's Gini criterion, a new branch is added to the tree and the decrease of Gini criterion for that node and associated variable is recorded. All the decreases in Gini criterion due to a variable are summed and normalized by the number of trees in a forest. Higher values of mean decrease of Gini criterion are associated to the most important variables, and therefore variables can be ranked based on this measure.

AncesTrees – a computer program of ancestry estimation

Our computer program, AncesTrees, performs the estimation of ancestry using the random forest algorithm for classification. The program consists of a spreadsheet file (.csv) and a script file. In the spreadsheet the user enters the measurements taken from a skull and selects which ancestral groups should be included in the model. Sex-specific models or the ability to estimate ancestry and sex at the same time are implemented. The ability to estimate sex using only individuals of one ancestral group, although not the focus of the program, is also permitted. The results of the algorithm are outputted in a separate sheet in the form of probabilities of membership to the ancestral groups included in the model. Out-of-bag accuracy rates, the cross-validation method inherent to bagging-based algorithms, are also outputted. These values can be interpreted as probabilities: the probability of correctly estimate the ancestry using the generated model or the probability of correctly detect individuals of a specific ancestry.

Like the FORDISC, fragmented skulls can be analyzed using our program because a new forest is fitted in each execution of the program according to the variables entered by the user. Beyond the variables introduced by the user, AncesTrees also uses a size-corrected version of those variables, which are added before the fitting procedure of a new random forest. These size-corrected variables are obtained by dividing each variable by the geometric mean of all cranial measurements available for an individual. This simple data transformation proposed by Darroch and Mossiman [38] removes the size component of linear distance measurements, leaving only a shape component. In addition to group membership and model accuracy information, the program also outputs information of variable importance as computed by random forest model.

In our implementation both size (normal variables) and shape (transformed variables) components are injected in the predictive model. Since tree-based classifiers have an automatic variable selection feature, the random forest will automatically give more importance and relevance to size or shape according to the estimation problem being posed by the user. If the user seeks to estimate sex for an individual of a specific ancestral group, the model will give more importance to the size variables given that sexual dimorphism, in this case, is mainly expressed as size differences. In a scenario where the user seeks to estimate ancestry but sex cannot be accurately known, which means creating a random forest model with both sexes pooled for each ancestral group, the shape component will be given more importance with more size-corrected variables used to create splits and new branches in each tree. In this case, shape is the most important component because ancestry variation is mainly expressed by shape differences. Giving more importance to shape also diminishes the risk of misclassifying an unusually small or large skull into an ancestral group where smaller or larger skulls are normally observed. It also allows the predictive model to achieve good levels of accuracy despite the fact that males and females are pooled together. When the user aims to estimate both sex and ancestry of a given skull, the model will seek a balance between the size and the shape component and the ranking of the most important variables will reflect that with a mixture of both size and shape variables as the most relevant for discrimination. Table 3 illustrate variable importance analysis in the three scenarios described.

Using AncesTrees 10,000 randomized classification trees are constructed in each execution of the program. The number of randomly selected variables to be used in each split is given by the squared root of the number of available variables, raw linear distances and their size-corrected version. AncesTrees is available upon request from the first author.

Results

Descriptive statistics of cranial measurements of the test samples are summarized in Tables 4 to 6.

The utility of the random forest algorithm in craniometric estimation of ancestry was tested under two scenarios: using a 6-way random forest model, where six ancestral groups were included and a 2-way random forest, to discriminate between individuals of African and European ancestry. In both models the sex was not considered. Results are summarized in Table 7 which also reports out-of-bag accuracy rates for both models using 23 cranial measurements, the most frequent situation during our test analysis.

In the 6-way random forest model, 75% of the individuals of African ancestry in the independent test sample were correctly identified, with the remaining 25% being misclassified as Austro-Melanesians. Seventy nine point two percent of the individuals of European ancestry were correctly identified, misclassified individuals are scattered across the remaining ancestral groups. The model involving only African and European ancestral groups performed much better, as expected. Ninety three point eight percent of all individuals in our test sample were correctly identified.

Discussion

The results obtained are very encouraging, especially because correct classification rates of Europeans and African individuals in our test sample are very similar to the expected classification rates, using the out-of-bag validation, the cross-validation technique used in AncesTrees. This is very important because it illustrates the generalization capacity, the ability to perform accurately other samples not used in the random forest model. If the out-of-bag estimates of accuracy, produced by our program, remain unbiased in other situations, as reported in this study, a 3-way random forest using 23 cranial measurements to discriminate, for example, between individuals of African ancestry, European and Native American, will have an expected overall accuracy of 88.6% which compares to results reported by Hefner et al. [4, 9], who used random forest modeling with metrical and morphoscopic data.

Despite being composed of archaeological specimens in its vast majority, the craniometric Howells Series proved to be very useful. The 32 individuals from the most recent Portuguese identified osteological collection were all correctly identified using the 2-way model and 93.8% of them were correctly allocated using the 6-way random forest.

Our results reflect the utility and advantages of the ensemble learning framework in such a complex domain like ancestry estimation. From a computational perspective, fitting an ensemble of classifiers instead of just one classifier, even if fully tuned and optimized with procedures such as stepwise variable selection, makes more sense and is clearly advantageous. Fitting a learning algorithm, like a classifier, can be viewed as a search problem where one seeks to find the best hypothesis, in this case the most accurate classifier [30]. Building a classifier for ancestry or sex estimation is in fact a search for the best hypothesis that allows us to explain *y*, ancestry or sex, as a function of *x*, the osteometric or morphoscopic data. What happens is that our data, what the learning algorithm uses to search for the best hypothesis, are small in comparison to size of the hypothesis space. The learning algorithm can find many hypotheses with the same accuracy on the training data. By combining all possible classifiers instead of selecting just one, choosing the wrong classifier can be avoided. This is particularly important when the classification model is to be applied to data that do not come from the same reference dataset. In other words, the cases of the unknown individuals that forensic anthropologists seek to positively identify. Forensic anthropologists work with finite reference or training data, which means that the learning algorithm will search only a limited set of hypotheses and stop looking when it finds a hypothesis that suits to the training dataset. Two problems arise as consequence of it. The algorithm may choose a hypothesis that is too specific of the reference data and is therefore over-fitted, or in more anthropological terms too specific of the population/reference sample in which it was developed.

In the worst case scenario the true function is not present in the hypothesis space. Ensemble learning methods such as bagging and the random forest handle these two problems by expanding the hypothesis space. The simple use of bootstrap replicates of the training data, instead of the training data itself, allows each classifier to be searched in a different hypotheses space, and when all classifiers are combined they represent solutions found in the vast hypotheses space. This key aspect of ensemble methods is crucial to tackle problems such as the population-specificity of methods used in biological profile estimation.

From what have been exposed, is easy to infer that the biggest advantage of the random forest is the fact that it is a balanced algorithm from a bias-variance tradeoff perspective [ref]. Its proprieties make it very effective at capturing regularities of the training data (low bias) and to generalize learned rules to unseen data with only residual and neglectable losses in predictive accuracy (low variance). Its low variance is its most valuable feature for anthropological applications. Since it is one of the most overftting-resistant algorithms, it is less likely to misclassify a skull simply because it is not very similar to the overall configuration of the database used to train the model or because the model is overtuned and captured noise from the training data. While ensemble models are not guaranteed to always provide the most accurate solution in terms of bias, it is very difficult for a single classifier to compete with an ensemble in terms of generalization. Ensembles are much more expressive in the hypothesis space they represent. In fact, what makes some type single classifiers so accurate (in terms of bias), such as stepwise variable selection procedures or optimization-based training (i.e. artificial neural networks), is what explains why they more often fail so dramatically in terms of generalization in unseen data.

A clear disadvantage of ensemble techniques like the random forest is the loss in interpretability of the generated model when compared to a single classifier. Ensemble methods operate as "black-boxes": in one side enters a set of measurements, and in the other is outputted a classification based on hundreds or thousands of classifiers. Although variable importance analysis is outputted for each new forest created by the program, we advise to proceed with caution if the user, for example, want to remove the least important variables and create a new model. Variable importance analysis is a useful tool but it reflects the importance of a variable for the ensemble model. A variable can be very discriminative in some of the trees composing the forest but its overall contribution small. Forest models are very effective at capturing variable interactions, a variable is important due to its individual discriminant power but also to its interactions. It is recommend to collect as much measurements as possible and let the program operate with all of them. As already illustrated it is capable of automatically select the most discriminative variables for the type of model requested by the user. The fact that the random forest uses a random subset of variables in each split makes it one of the most effective techniques for high-dimensional problems [39]. A large number of cranial measurements does not represent a problem using random forest models.

Conclusion

The AncesTrees base learning algorithm, random forest, proved to be very successful both in global and binary ancestry discrimination. Although the specificities of the Howells Craniometric Series (within each ancestral group there are individuals from different geographic regions and time periods), the program provided excellent classification results both for contemporaneous Portuguese European individuals and archaeological African slaves.

Like every predictive system, AncesTrees is not guaranteed to always provide a correct estimate, it would be completely unrealistic to have such expectation. It is reasonable to say that using craniometric measurements the maximum accuracy one can expect to achieve in ancestry estimation without incurring in overfitting is around 97%. As we already mentioned, the main advantage of random forest is not in its absolute accuracy but in its ability to generalize to unseen data. Computer programs like AncesTrees, FORDISC and CRANID are designed as decision support systems to aid the user arrive at a scientifically informed decision. Due to the complexity inherent of ancestry estimation we think that most effective strategy is to think of existent methods as an ensemble classifier. Instead of look at different methods as opposing approaches, one should aim to combine their results as in an ensemble model. Metrical and morphological data being used under different computational approaches such discriminant functions, K-nearest neighbors, random forests or support vector machines. AncesTrees is a relevant contribution for such endeavor.

The results obtained through this program are promising although more tests in samples from different geographic origins and ancestries are necessary in order to prove its utility in Forensic Sciences and to allow, in the future, robust comparisons with other methods. AncesTrees is available upon request from the first author.

Acknowledgements

The authors thank to Centro de Ciências Forenses and to Centro de Investigação em Antropologia e Saúde.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Ousley S, Jantz R, Freid D (2009) Understanding race and human variation: Why forensic anthropologists are good at identifying race. Am J Phys Anthropol 139:68–76

2. Slice D, Ross AH (2009) 3D-ID: geometric morphometric classification of crania for forensic scientists.

3. Hefner JT (2009) Cranial Nonmetric Variation and Estimating Ancestry*. J Forensic Sci 54:985–995

4. Hefner JT, Spradley K, Anderson BE (2011) Ancestry estimation using random forest modelling. Proc. Am. Acad. Forensic Sci. Chicago, IL, pp 352–353

5. Hefner JT, Ousley SD, Dirkmaat DC (2012) Morphoscopic Traits and the Assessment of Ancestry. In: Dirkmaat DC (ed) Companion Forensic Anthropol., First. Wiley-Blackwell, West Sussex, UK, pp 287–310

6. Edgar HJH (2005) Prediction of race using characteristics of dental morphology. J Forensic Sci 50:269–273

7. Edgar HJH (2009) Testing the utility of dental morphological traits commonly used in the forensic identification of ancestry. Front Oral Biol 13:49–54

8. Edgar HJH (2013) Estimation of ancestry using dental morphological characteristics. J Forensic Sci 58 Suppl 1:S3–8

9. Hefner JT, Spradley MK, Anderson B (2014) Ancestry Assessment Using Random Forest Modeling, ,. J Forensic Sci 59:583–589

10. Hefner JT, Ousley SD (2014) Statistical Classification Methods for Estimating Ancestry Using Morphoscopic Traits,. J Forensic Sci n/a–n/a

11. Giles E, Elliot O (1962) Race identification from cranial measurements. J Forensic Sci 7:147–157

12. Ousley SD, Jantz RL (2005) FORDISC 3.0: Personal Computer Forensic Discriminant Functions. Universty of Tennesse

13. Ousley SD, Jantz RL (2012) ForDisc 3 and Statistical Methods for Sex and Ancestry Estimation. In: Dirkmaat DC (ed) Companion Forensic Anthropol., First. Wiley-Blackwell, West Sussex, UK, pp 311–329

14. Wright R (1992) Correlation between Cranial Form and Geography in Homo Sapiens: CRANID - A Computer Program for Forensic and Other Applications. Archaeol Ocean 27:128–134

15. Wright R (2008) Detection of Likely Ancestry Using CRANID. In: Oxenham M (ed) Forensic Approaches Death Disaster Abuse. Australian Academic Press, Sydney, Australia, pp 111–122

16. Du Jardin P, Ponsaillé J, Alunni-Perret V, Quatrehomme G (2009) A comparison between neural network and other metric methods to determine sex from the upper femur in a modern French population. Forensic Sci Int 192:127.e1–6

17. Mahfouz M, Badawi A, Merkl B, Fatah EEA, Pritchard E, Kesler K, Moore M, Jantz R, Jantz L (2007) Patella sex determination by 3D statistical shape models and nonlinear classifiers. Forensic Sci Int 173:161–170

18. Moss GP, Shah AJ, Adams RG, Davey N, Wilkinson SC, Pugh WJ, Sun Y (2012) The application of discriminant analysis and Machine Learning methods as tools to identify and classify compounds with potential as transdermal enhancers. Eur J Pharm Sci Off J Eur Fed Pharm Sci 45:116–127

19. Howells WW (1973) Cranial Variation in Man: A Study by Multivariate Analysis of Patterns of Difference Among Recent Human Populations. Harvard University Press

20. Howells WW (1989) Skull Shapes and the Map: Craniometric Analyses in the Dispersion of Modern Homo. Peabody Museum of Archaeology and Ethnology, Harvard University

21. Howells WW (1995) Who's Who in Skulls: Ethnic Identification of Crania from Measurements. Peabody Museum of Archaeology and Ethnology, Harvard University

22. Howells WW (1996) Howells' craniometric data on the Internet. Am J Phys Anthropol 101:441–442

23. Neves MJ, Almeida M, Ferreira MT (2011) História de um arrabalde durante os séculos XV e XVI: O "poço dos negros" em Lagos (Algarve, Portugal) e o seu contributo para o estudo dos escravos africanos em Portugal. In: Matos AT, Costa JPO (eds) Herança Infante História Arqueol. E Museol. Em Lagos. Câmara Municipal de Lagos, Lagos, Portugal, pp 29–46

24. Coelho C (2012) Uma Identidade perdida no mar e reencontrada nos ossos: avaliação das afinidades populacionais de uma amostra de escravos dos séculos XV-XVI. Dissertation, University of Coimbra

25. Cunha E, Wasterlain S (2007) The Coimbra identified osteological collections. In: Grupe G, Peters J (eds) Skelet. Ser. Their Socio-Econ. Context. Verlag Marie Leidorf, GmbH, Rahden/Westf, Germany, pp 23–33

26. Cunha E (1989) Cálculo de Funções Discriminantes para a Diagnose Sexual do Crânio. Dissertation, University of Coimbra

27. Ferreira MT, Navega D, Vicente R, Cunha E (2013) A Colecção de Esqueletos Identificados Século XXI. 12º Congr. Nac. Med. Leg. E Ciênc. Forenses

28. Breiman L (2001) Random Forests. Mach Learn 45:5–32

29. Dietterich TG (2000) An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Mach Learn 40:139–157

30. Dietterich TG (2000) Ensemble Methods in Machine Learning. Mult. Classif. Syst. Springer Berlin Heidelberg, pp 1–15

31. Mitchell TM (1997) Machine Learning. McGraw Hill, Burr Ridge, IL

32. Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer

33. Breiman L (1996) Bagging predictors. Mach Learn 24:123–140

34. Ho TK (1995) Random decision forests. Proc. Third Int. Conf. Doc. Anal. Recognit. 1995. pp 278–282 vol.1

35. Ho TK (1998) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20:832–844

36. Amit Y, Geman D (1997) Shape Quantization and Recognition with Randomized Trees. Neural Comput 9:1545–1588

37. Kleinberg EM (1996) An overtraining-resistant stochastic modeling method for pattern recognition. Ann Stat 24:2319–2349

38. Darroch JN, Mosimann JE (1985) Canonical and principal components of shape. Biometrika 72:241–252

39. Yang P, Hwa Yang Y, B. Zhou B, Y. Zomaya A (2010) A Review of Ensemble Methods in Bioinformatics. Curr Bioinforma 5:296–308