Short communication - Machine learning, bootstrapping, null models and why we are still not 100% sure which bone surface modifications were made by crocodiles

Shannon P. McPherron[1], Will Archer[2,3], Erik R. Otárola-Castillo[4], Melissa G. Torquato[4] & Trevor L. Keevil[4]

[1]Department of Human Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, [2]Max Planck Partner Group, Department of Archaeology and Anthropology, National Museum, Bloemfontein, South Africa, [3]Department of Geology, University of the Free State, Bloemfontein, South Africa, [4]Department of Anthropology, Purdue University, West Lafayette, Indiana, USA

1. Introduction

Data science and open science are two of the more interesting developments in recent years that influence how research is conducted and disseminated. Data science generally draws on sophisticated, newly accessible methods of quantitative analysis as applied to large data sets, and this field is rapidly evolving. Open science represents a drive to make the scientific process, from experimental design and data collection to analysis and publication, more transparent and accessible (Wilkinson et al. 2016). Here, we argue for the interdependence of these two developments by exploring a paper recently published in the on-going and often contentious debate over the interpretation of bone surface modifications. We show how an application of machine learning in this instance artificially inflated the success rate of classification (Domínguez-Rodrigo and Baquedano 2018) and obscured a far simpler explanation for the differentiation of marks based on their measurements. We do this by replicating their study, following the published descriptions of the methods. We simulated our own random and patterned data to generate expectations for the machine learning model provided by the study's authors and analyzed the results. Aside from what our findings might mean for the interpretation of who or what made marks on bones, we use this example to highlight the increasing importance of the open science emphasis on methodological transparency, as more sophisticated data science protocols are brought into paleoanthropology.

1.1 Background

Machine learning algorithms are becoming increasingly influential in many sciences, and their potential utility is being explored more frequently by archaeologists. In general, machine learning enables computers to solve scientific problems through learning from data and by improving this learning (often self-correctively), in a semi-autonomous fashion. This approach allows for the semi-automated detection and modelling of patterns in existing data sets so as to make predictions for new data sets. Predictive modeling in general is not new, yet machine learning is particularly well-suited to working with data sets that contain large numbers of variables (e.g., tens or 100s of thousands in the case of images), and that may have complex interactions between them (e.g., recognizing a bicycle in an image). In these situations, many predictive approaches may over-fit. Overfitting occurs when a GLM is unnecessarily complex. In other words, when a model possesses too many parameters (and predictors) to be reliably estimated relative to the size of a given data set. In this context, a statistically ‘significant’ effect may describe a random error association rather than a true relationship in the data. Such significant relationships are spurious - usually only specific to the data in hand and not generalizable to the archaeological record. Consequently, overfitting often inflates the predictive accuracy of a model. Given the large number of variables and the potential for complex interactions between them, machine learning also requires large sample sizes to build (or train) the model. So, while the statistical fundamentals behind many of these algorithms have been around for several decades or more, it is only relatively recently that the combination of inexpensive computing, accessible software, and massive digital data collection have made possible their widespread use.

However, it is also true about machine learning that while it is straightforward to measure the predictive success of a model, it is much more difficult to understand or visualize how the model itself works, which variables are driving it, how they are interacting with one another, which cases have more leverage, etc. This is especially problematic in the context of scientific enquiry where understanding causality is as important as predictive success. Given this situation of a powerful new statistical technique being difficult to describe and understand, in our opinion, the way forward is greater sharing of data (where ethically appropriate) and of the code used to analyze these data (Barnes 2010; Hoffman 2016; Wilkinson et al. 2016; Marwick 2017). The combination of data and code allow the results to be reproduced (Marwick 2017) and explored in ways that can lead to a better understanding of issues like causality. The sharing of code can also lessen the burden of a paper’s methods section to fully and precisely describe the steps taken. To illustrate these points, we consider the recent publication by Domínguez-Rodrigo and Baquedano (2018) on the interpretation of bone surface modifications with our own re-analysis of their published data that also includes the code for doing so (see Supplementary Online Material [SOM] for the text and statistical source code used to make this document).

Their paper is largely a reaction to a high-profile paper by Sahle, El Zaatari, and White (2017) that questions the current emphasis on developing new, sophisticated, quantitative techniques of mark analysis in contrast to building larger comparative samples through new fieldwork and experimentation. In their paper, Domínguez-Rodrigo and Baquedano (2018) apply machine learning algorithms to their own data set of experimental surface modifications on bones produced by trampling, crocodile consumption, and butchery using two types of stone tools (simple and retouched) to show that, with the right statistical techniques, the agent responsible for a surface modification can be identified 100% of the time. They also apply the same machine learning techniques to mark data published by Sahle, El Zaatari, and White (2017). In the end, they conclude that there is a “methodological basis to overcome a purported, no longer existent equifinality in the identification of cut marks and crocodile bite marks” (p. 5).

We agree that new quantitative methods, along with field methods and continued experimentation, are an essential part of moving forward on the challenges presented by the variation exhibited by bone surface modifications (e.g., Harris et al. 2017; Maté-González et al. 2019; Otárola-Castillo et al. 2018; Pante et al. 2017; Yravedra et al. 2017; Byeon et al. 2019). Here, however, we raise some specific methodological considerations with the machine learning application conducted by Domínguez-Rodrigo and Baquedano (2018) that could account for the high predictive success rate. As we demonstrate, it is possible that their classification success rate remains extremely high even after these issues are addressed, but we use this example to illustrate our general point that the decreased transparency of machine learning can hinder scientific understanding, especially in the absence of data and code sharing.

Our main concern lies with one important aspect of their study: namely, how they created their statistical sample. Machine learning algorithms require very large samples to initially train the models to recognize the variability associated with a particular attribute (e.g., a crocodile tooth mark). However, in the field of archaeology where data sharing is still not common practice, and where there are few widely agreed upon protocols for data recording, building large samples is often prohibitively expensive and time consuming. In their own case, Domínguez-Rodrigo and Baquedano (2018) had data recorded from an already reasonably large experimental assemblage consisting of 633 marks, each with up to 17 variables describing them. Next, "in order to provide the modelling with large training and testing/validation sets, the sample was bootstrapped 10,000 times" (p. 5).

Bootstrapping is an extremely useful statistical technique when the underlying structure of variability in a data set is unknown or its population is difficult to sample (Efron and Tibshirani 1994). Typically, a large number of data sets is created by repeatedly, randomly sampling the original data set with replacement (meaning that individual cases in the original data set can be represented more than once in the bootstrapped data sets). This new set of assemblages then can be measured using standard descriptive statistics to develop some idea of the likelihood of various outcomes. For instance, given a set of faunal assemblages comprised of various species, bootstrapping new, hypothetical assemblages by drawing randomly from this original set can give some idea of how (un)likely finding an assemblage of only one of these species would be (e.g., were hominins targeting this particular species or is this simply chance). Importantly, however, bootstrapping cannot generate new types of observations (e.g., an instance of an assemblage dominated by a species not represented in the original data set). It can only resample and reshuffle what is already present. In the context of machine learning, bootstrapping can be useful for providing the model with new, additional training sets drawn at random from the original data set. These new training sets will not be larger (in terms of the number of cases) than the original data set from which they are drawn. If instead of making new data sets in this way, Domínguez-Rodrigo and Baquedano (2018) instead resampled the experimental data set to 10,000 cases to meet the need of machine learning for large data sets (as the wording of their paper indicates), prior to splitting training and validation samples, they amplified the existing structure of variability in the data and in doing so artificially increased the statistical power (and classification success rate) of the machine learning algorithms.

2. Methods and results

To assess the performance of Dominguez-Rodrigo and Baquedano's (2018) machine learning techniques, first, we evaluated the outcomes of bootstrapping random data. We then modeled their published experimental data using machine learning to compare the results with those of more conventional approaches.

2.1. Bootstrapping Random Data

We demonstrate the impact of bootstrapping as used by Domínguez-Rodrigo and Baquedano (2018) on classification success rate by replicating their study using a data set that machine learning should not be able to classify correctly. We generated a new data set that conforms to the structure of their data in terms of the number variables recorded for each bone surface modification and in terms of the types of observations made for each of these variables (see Supplementary Information [SI] in Domínguez-Rodrigo and Baquedano (2018)). However, rather than actual observations on bone surface modifications, we used entirely random data. So, for example, where Domínguez-Rodrigo and Baquedano (2018) categorize groove symmetry as symmetrical or asymmetrical, we assigned a random number of either 1 or 2 (treated as categorical in the analysis), respectively. We represent the categories of this variable, symmetric or asymmetric, by the numbers 1 and 2. However, our analyses effectively treat the variable as categorical. Of their 17 variables, 15 are of this type, where the observations fall into one of a few categories. Two variables, mark length and the number of marks per bone, required a different approach. For mark length, we assigned a random value from a uniform distribution ranging from 1 to 3 cm. For the number of marks per bone, we assigned a random value from a uniform distribution of 1 to 10. We then generated an initial data set of cases of random observations and assigned them to crocodile, trampling, simple cut-mark, and complex cut-mark agents, in proportion to the Domínguez-Rodrigo and Baquedano (2018) experimental sample (58, 224, 246, and 105 marks, respectively).

Doing our best to follow the protocol described by Domínguez-Rodrigo and Baquedano (2018), we then resampled or bootstrapped these data to generate a new data set of 10,000 cases. However, there are several ways this could be done, and how exactly they did it is not clear in their protocol description. Domínguez-Rodrigo and Baquedano (2018) state only that "the sample was initially bootstrapped with a function from the 'caret' R library that considers bootstrapping the sample in proportion to the variable representation to each of the factors of the outcome variable" (p. 5). Not knowing exactly which function they used but taking as our guide the emphasis on variable representation, we chose from the caret library (Kuhn 2020) the upSample() function which creates an equal number of samples in each category (here the type of mark) by randomly resampling the less well represented categories until they equal the best represented category (here simple cutmarked bones). Finally, we resampled this data set to 10,000 cases using a bootstrap with replacement. We note that bootstrapping directly to 10,000 cases without first using the upSample() step does not change the results reported here.

Following Domínguez-Rodrigo and Baquedano (2018), we applied the Random Forest machine learning algorithm to our simulated random and bootstrapped data. Random Forest is one of eight algorithms tested by Domínguez-Rodrigo and Baquedano (2018). Of the eight algorithms they tested, five achieved 100% classification success on the full data set, and of these, four performed equally well (each with 99% success) on a reduced variable data set. Random Forest is among these four. To use Random Forest on our bootstrapped random data, we first divided it into a training set and an independent validation set using a 70/30 respective split (all statistics used here were computed in R v. 4.0.5 (R Core Team 2021)). We then used the caret() library (Kuhn 2020) to create a model of the training sample (7000 cases). When we applied our Random Forest model to the validation data set, the classification success rate was 100% (Fig. 1a), which fully replicates the results reported by Domínguez-Rodrigo and Baquedano (2018).

The effect of the bootstrapping step on this result is clear when we apply the same machine learning algorithm to our initial data set of non-bootstrapped random data composed of cases. Here, after using the same 70/30 training/validation sampling, an average of 36.09% of the cases in the validation sample were correctly classified (Fig. 1a). This classification success rate is above what we would expect by chance alone (1:4 or 25%), because the mark types are unevenly proportioned in the original data. Naturally, marks that are better represented in the sample are correctly classified more of the time (Table 1). This sample case illustrates the potential for artificial inflation of group differences when bootstrapping is applied to increase sample sizes.

Table 1: Breakdown of classification success rate on the original-sized random data set. In this table, rows represent the actual mark type we assigned and columns represent the predicted (Pred.) mark type based on the Random Forest model. Mark types that are better represented in the random data set (e.g., simple cutmarks and trampling marks) more often correctly (i.e., have the lowest error rate).

	Pred. Croc.	Pred. CM Ret.	Pred. CM Simple	Pred. Trampling	Error Rate
Crocodile	0	2	27	29	1.00
Cutmark Retouched	0	3	61	41	0.97
Cutmark Simple	0	8	131	107	0.47
Trampling	3	6	120	95	0.58

Doing this may inadvertently generate misleading patterns when applied to the archaeological record. For example, having established that machine learning and bootstrapping successfully determined the agent of bone modification in their experimental data set, Domínguez-Rodrigo and Baquedano (2018) apply the same methodology to the archaeological data published by Sahle et al. (2017). Their conclusion is that crocodile marks can be distinguished from butchery marks and that marks on three of four fossil bones look more like they were made from butchery than by crocodiles. The published data set (Table 1 of the SI of Sahle et al. 2017) consists of nine cases (marks on four fossil bones, marks from four crocodile experiments, and marks from one butchery experiment) described by 12 variables, with no indications of variance or sample size. In this case a "…random forest (RF) was used on a slightly bootstrapped sample ($n = 100$) of the experimental data set of Sahle et al.´s [1] Table 1, and excluding the fossil bones" (SI, Domínguez-Rodrigo and Baquedano [2018]). In other words, five cases were resampled to 100 and "…this bootstrapped sample yielded a classification of crocodile [bone surface modifications] and butchery [bone surface modifications] with an accuracy of 100%…" (SI, Domínguez-Rodrigo and Baquedano [2018]). Domínguez-Rodrigo and Baquedano (2018) present multivariate plots (see their SI Fig. S2) indicating that mark maximum width ('mw') is the sole variable driving the distinction of crocodile tooth marks from stone tool butchery marks. When one considers the mark maximum width values for the four crocodile tooth-marked bones and the one butchery-marked bone in the Sahle, El Zaatari, and White (2017) data set, it is clear why: the experimental butchery marks are wider than any of the crocodile tooth marks, and bootstrapping to augment a sample with such limited variation allows a complete separation of the cases on this variable. Domínguez-Rodrigo and Baquedano (2018) go on to build a multivariate machine learning model using eight predictor variables from the five experimental cases. This model is then used to predict whether the marks on four fossil bones are the results of crocodiles or butchery activity. Whether resampling was used in this case is not stated, but the classification probabilities are less than 100% (roughly 80–20 in this case).

Importantly, when it comes to issues of sample size, bootstrapping existing data cannot be used as a substitute for collecting more actual data. Nonetheless, the methods as presented by Domínguez-Rodrigo and Baquedano (2018) do not demonstrate the efficacy of machine learning for improving the classification of bone surface modifications relative to more conventional predictive modelling approaches. In our view, given the difficulty of separating trampling marks from stone-tool cutmarks (both of which are made by sharp stones), 100% classification success rates seemed an unlikely possibility. Moreover, data scientists often suspect that overfitting, rather than a general model, might underlie such perfectly classified outcomes of machine and deep learning algorithms (Bilbao and Bilbao 5AD, @nichols_machine_2019). Consequently, we are still left with the question, can machine learning methods outperform more conventional statistical techniques? To answer this, we evaluated the effectiveness of machine learning over more conventional approaches such as Discriminant Function Analysis (DFA) and Multinomial Regression (a likelihood-based generalized linear model).

2.2. Comparing machine learning and "conventional" methods

To compare the performance of machine learning to more conventional statistical methods, we re-analyzed a portion of the experimental data set used by Domínguez-Rodrigo and Baquedano (2018) and published by Dominguez-Rodrigo et al. (2009, their Table 5). We conducted our analyses using their machine learning methods, DFA and a likelihood-based GLM. Results of our re-analyses shows that near 100% classification success is still possible using machine learning and the more conventional methods. Because Dominguez-Rodrigo and Baquedano's (2018) original data were not available to us for re-analysis—beyond replication—we conducted a statistical reconstruction of the data set using their published summaries. With access to this information, we examined what is driving their results, to achieve a better causal understanding.

The data set for our re-analysis contains nearly all of Domínguez-Rodrigo and Baquedano's (2018) cases for the two types of butchery (with retouched and non-retouched tools) and trampling marks. However, the crocodile tooth mark cases are not published in comparable detail to the butchery cases and thus could not be included. Dominguez-Rodrigo et al. (2009, their Table 5) summarize the percentages of 14 of the 17 variables later used by Domínguez-Rodrigo and Baquedano (2018). Each variable is composed of multiple categories – sometimes two or three. For example, the variable named 'Groove Trajectory' is made up of three categories: straight, curvy, and sinuous. On the other hand, the variable "Barb" is comprised of two observable categories: its presence and absence. In addition, several variables are dependent on the state of another. For instance, “microstriation trajectory”, “shape of microstriation trajectory”, and “location of microstration”, depend on whether the variable named “internal microstriation” is “present” or “absent”. (Domínguez-Rodrigo et al. 2009, their Table 5) do not include three variables used by Domínguez-Rodrigo and Baquedano (2018): “number of conspicuous grooves”, “main groove length,” and “associated tooth pits on mid-shafts”, and thus we could not analyze them here. Domínguez-Rodrigo and Baquedano (2018) call these variables 5, 15, and 17. The authors describe these variables in text (Domínguez-Rodrigo et al. 2009) and their supplementary materials (Domínguez-Rodrigo and Baquedano 2018).

The reported percentages of the variables’ categories among the experimental unretouched, retouched, and trampled marks capture the major patterns in the data analyzed by Domínguez-Rodrigo and Baquedano (2018). Consequently, we modeled the outcomes of each variable’s categories as a multinomial random variable. The multinomial distribution is a natural statistical model of events determined by an underlying vector of proportions – such as these data. The data generation process and data reconstruction are described in SOM (BSMsimData_signal.R). Our reconstruction replicates a data matrix similar to that used by Domínguez-Rodrigo and Baquedano (2018) and is composed of a total 575 unretouched and retouched butchery and trampling cases (105, 246, and 224 respectively).

We then fit a Random Forest model to the reconstructed data. From this model, we made a dotchart using the ‘varImpPlot’ function from the randomForest package (Liaw and Wiener 2002) to assess the hierarchy of variable importance and noted that associated shallow striation (abrasion), groove shape and striae trajectory location (variables 16, 4, and 14) are overwhelmingly more important than the other 14 variables used by Domínguez-Rodrigo and Baquedano (2018) to build their model. We provide here the variable importance plot for illustrative purposes (Fig. 1b), and we note that one variable, abrasion, is ~80.6% successful at separating trampled bones from cut-marked bones in the data set. The two other variables, groove shape and striae location, separate marks from retouched and non-retouched tools. Thus, we fitted a new Random Forest model using just these three variables of abrasion, groove shape and striae location. With this drastically reduced Random Forest model we achieved a classification success rate of ~99.4%, calling into question the utility of the 11 other descriptive variables (and the three variables we were unable to model) used by Domínguez-Rodrigo and Baquedano (2018).

Figure 1. a) A comparison of the classification success rates between the random data set sized to match the actual number of cases in the Domínguez-Rodrigo and Baquedano (2018) data set and the same data set bootstrapped to 10,000 cases (left) and b) A dotchart showing the relative importance, as measured here by the Gini coefficient, of different types of mark observations on the Random Forest model’s predictive success

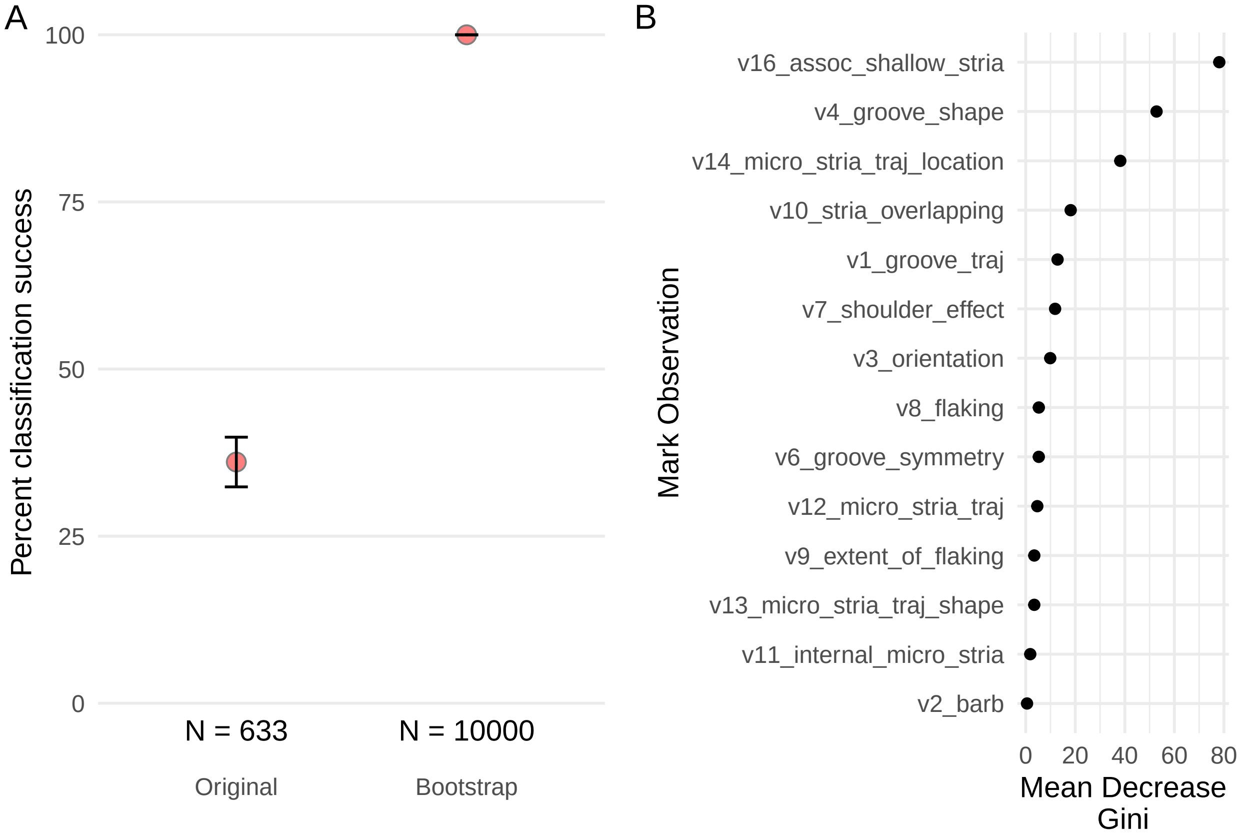

Figure 1: a) A comparison of the classification success rates between the random data set sized to match the actual number of cases in the Domínguez-Rodrigo and Baquedano (2018) data set and the same data set bootstrapped to 10,000 cases (left) and b) A dotchart showing the relative importance, as measured here by the Gini coefficient, of different types of mark observations on the Random Forest model's predictive success (right). The Gini coefficient is essentially a measure of the average importance of a given variable in splitting the mark types relative to the agents that produced them across all the decision trees in the model. Here these are sorted by the most important variables in the Random Forest model at the top and the variables that contribute the least to the success of the model at the bottom. From this plot it is clear that three variables (abrasion, groove shape and striae location) most strongly influence the model.

(right). The Gini coefficient is essentially a measure of the average importance of a given variable in splitting the mark types relative to the agents that produced them across all the decision trees in the model. Here these are sorted by the most important variables in the Random Forest model at the top and the variables that contribute the least to the success of the model at the bottom. From this plot it is clear that three variables (abrasion, groove shape and striae location) most strongly influence the model.

As an alternative to machine learning, given that the structure of variability in the data is simpler than initially imagined, we also modeled these data using standard multivariate statistical models. To do so, we again started by subsampling the reconstructed data set into 70/30 training/testing data sets. To create predictive models of mark type, we used traditional LDA using the lda function of the MASS package (Venables and Ripley 2002) in R (R Core Team 2021). We also modeled mark type using a Multinomial GLM. In the GLM, Maximum Likelihood achieved parameter estimates. To generate the GLM, we used the multinom function from the nnet package which, although it is built to construct artificial neural networks, defaults to a maximum likelihood fit, thus making it a GLM by definition (Venables and Ripley 2002). We created the LDA and GLM models using only the training data set. We then used the independent test data and the estimated coefficients of each model to predict the type of mark based on their individual combination of 14 variables. The LDA model was able to predict the marks from the independent testing sample with ~95.3% accuracy (Figure 2). The GLM performed slightly better, achieving an accuracy of 100%. The accuracy of these models is quite similar to the machine learning results reported by Domínguez-Rodrigo and Baquedano (2018) and partially replicated here. Lastly, as with our machine learning replication, we applied the models on a reduced set of variables. After carefully considering the influence of each of the 14 variables, we included only two variables: abrasion and striae location. These simpler models predicted mark type with ~94.2% (LDA) and ~94.2% (GLM) accuracy.

Figure 2. Graphical representation of Linear Discriminant Analysis (LDA) showing separation of our reconstructed bone surface modification data. The LDA correctly recognized the identity of each bone surface modification with an approximate accuracy of 95% (whether they were marks created by retouched or unretouched stone tools, or trampling). Overall, retouched (green triangles) and unretouched (red circles) cut marks exhibit greater similarity to one another than to trampling marks (blue squares). As such, the cut marks are more difficult to differentiate from one another than from trampling marks. The different colored regions highlight the class separation decision boundaries of the LDA model.

3. Discussion and conclusions

We have tried to show here that the relative novelty of machine learning methods in archaeology, and their

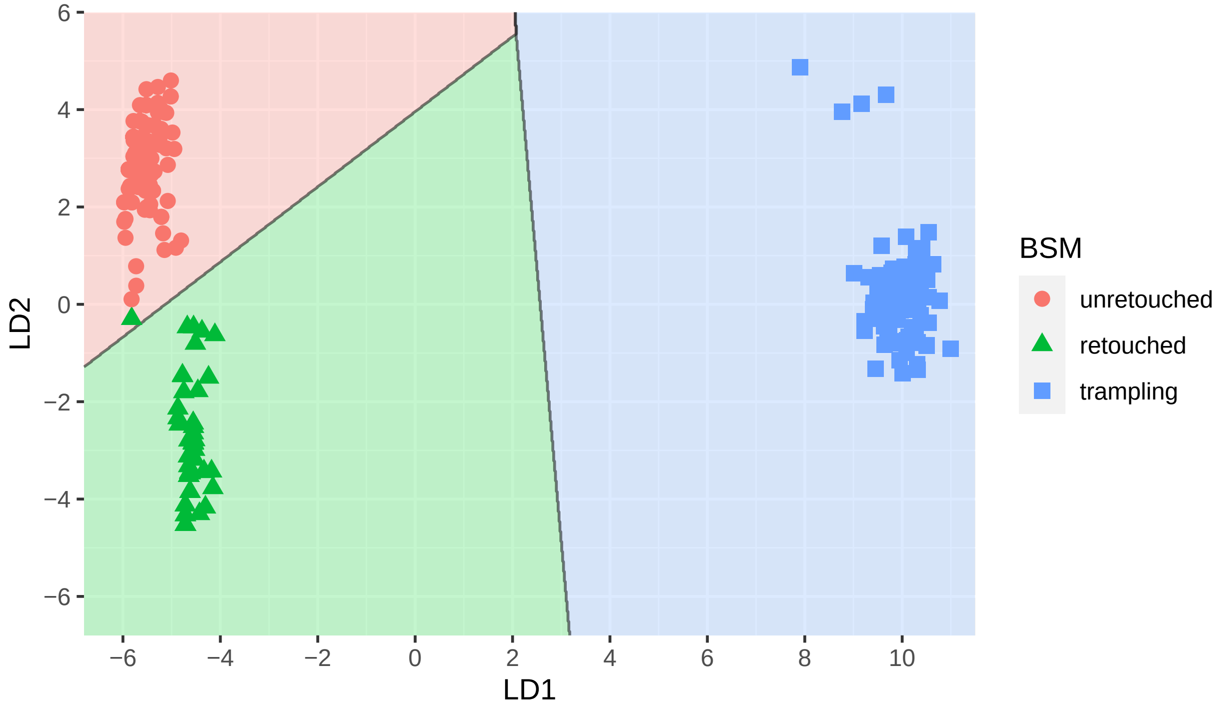

Figure 2: Graphical representation of Linear Discriminant Analysis (LDA) showing separation of our reconstructed bone surface modification data. The LDA correctly recognized the identity of each bone surface modification with an approximate accuracy of 95% (whether they were marks created by retouched or unretouched stone tools, or trampling). Overall, retouched (green triangles) and unretouched (red circles) cut marks exhibit greater similarity to one another than to trampling marks (blue squares). As such, the cut marks are more difficult to differentiate from one another than from trampling marks. The different colored regions highlight the class separation decision boundaries of the LDA model.

complexity in general, make it difficult to evaluate the appropriateness of their application. It is not our intention to call into question the general use of machine learning but rather to emphasize that moving forward, the efficient evaluation and replication of these types of papers will, in many cases, require access to the original data sets in addition to the detailed associated methods descriptions (i.e., the scripts or a description thereof), which are now more commonly published alongside archaeology papers (e.g., Clarkson et al. 2015; McPherron 2018; Miller-Atkins and Premo 2018; Calandra et al. 2019; Reeves et al. 2019; Mraz et al. 2019; Coco, Holdaway, and Iovita 2020; McPherron et al. 2020).

In the absence of this methodological detail, we have presented another approach to gaining insights into the appropriateness of a particular statistical method by building a null expectation using random data. In doing so, we have demonstrated that, as it was described, Domínguez-Rodrigo and Baquedano's application of machine learning to their experimental data produces results that are indistinguishable from the null model, and we think that because of the small sample size of the data set (nine rows of data) in the Sahle, El Zaatari, and White (2017), resampling as applied by Domínguez-Rodrigo and Baquedano (2018) is likely to produce an inaccurate result, similar to what we demonstrated here with random data.

For archaeologists, the main goal of predictive modelling is to develop a tool that can generalize reliably to unknown archaeological cases, in other words, a model that can make good predictions. This is conventionally assessed with a so-called out-of-bag test sample of data, which is left aside for validation purposes when the predictive model is built with a training sample of data. In the approach of Domínguez-Rodrigo and Baquedano (2018), our understanding is that the test portion of data was subjected to the same process of resampling as the training data, artificially inflating the success of prediction, and making the out-of-sample performance of the model challenging to evaluate.

Through our statistical reconstruction of Domínguez-Rodrigo and Baquedano's data set, we also demonstrated that the pattern in their experimental data is driven by only two or three variables, rather than 14, with the presence or absence of abrasion having the highest importance. In other words, the application of machine learning in this case obscured our understanding of how bone surface modifications are linked to agents. Minimally, the classification success rate on the original data set of cases and a variable importance plot would have been important statistics for Domínguez-Rodrigo and Baquedano (2018) to report. In their absence, once the full data set is published, we will have a better idea how sure we can be that it was a crocodile that marked the bone and why we think so.

Acknowledgements

We thank David Braun, Jordy Orellana Figueroa, Philipp Gunz, Jacob Harris, Curtis Marean, Luke Premo, Jonathan Reeves, Claudio Tennie and others for reading and providing their insights into previous versions of this manuscript. Manuel Domínguez-Rodrigo and Enrique Baquedano shared their data with us to replicate their original analyses and engaged in open discussions and correspondence about this paper prior to submission. While it is normal to thank the reviewers and editors for their help with a manuscript, in this case we want to particularly acknowledge the efforts made by these individuals. From detailed edits to thought provoking comments and suggestions, they helped make this paper much better than where it started. As always, though, all mistakes remain our own.

References

Barnes, Nick. 2010. "Publish Your Computer Code: It Is Good Enough." *Nature* 467 (7317): 753–53. https://doi.org/10.1038/467753a.

Bilbao, I., and Bilbao. 5AD. "Overfitting Problem and the over-Training in the Era of Data: Particularly for Artificial Neural Networks." In *2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS)*, 173–77. https://doi.org/10.1109/INTELCIS.2017.8260032.

Byeon, Wonmin, Manuel Domínguez-Rodrigo, Georgios Arampatzis, Enrique Baquedano, José Yravedra, Miguel Angel Maté-González, and Petros Koumoutsakos. 2019. "Automated Identification and Deep Classification of Cut Marks on Bones and Its Paleoanthropological Implications." *Journal of Computational Science* 32 (March): 36–43. https://doi.org/10.1016/j.jocs.2019.02.005.

Calandra, Ivan, Lisa Schunk, Alice Rodriguez, Walter Gneisinger, Antonella Pedergnana, Eduardo Paixao, Telmo Pereira, Radu Iovita, and Joao Marreiros. 2019. "Back to the Edge: Relative Coordinate System for Use-Wear Analysis." *Archaeological and Anthropological Sciences* 11 (11): 5937–48. https://doi.org/10.1007/s12520-019-00801-y.

Clarkson, Chris, Mike Smith, Ben Marwick, Richard Fullagar, Lynley A. Wallis, Patrick Faulkner, Tiina Manne, et al. 2015. "The Archaeology, Chronology and Stratigraphy of Madjedbebe (Malakunanja II): A Site in Northern Australia with Early Occupation." *Journal of Human Evolution* 83 (June): 46–64. https://doi.org/10.1016/j.jhevol.2015.03.014.

Coco, Emily, Simon Holdaway, and Radu Iovita. 2020. "The Effects of Secondary Recycling on the Technological Character of Lithic Assemblages." *Journal of Paleolithic Archaeology* 3 (3): 453–74.

https://doi.org/10.1007/s41982-020-00055-4.

Domínguez-Rodrigo, Manuel, and Enrique Baquedano. 2018. “Distinguishing Butchery Cut Marks from Crocodile Bite Marks Through Machine Learning Methods.” *Scientific Reports* 8 (1): 5786. https://doi.org/10.1038/s41598-018-24071-1.

Domínguez-Rodrigo, M., S. de Juana, A. B. Galán, and M. Rodríguez. 2009. “A New Protocol to Differentiate Trampling Marks from Butchery Cut Marks.” *Journal of Archaeological Science* 36 (12): 2643–54. https://doi.org/10.1016/j.jas.2009.07.017.

Efron, Bradley, and Robert J Tibshirani. 1994. *An Introduction to the Bootstrap*. CRC press.

Harris, Jacob A., Curtis W. Marean, Kiona Ogle, and Jessica Thompson. 2017. “The Trajectory of Bone Surface Modification Studies in Paleoanthropology and a New Bayesian Solution to the Identification Controversy.” *Journal of Human Evolution* 110 (September): 69–81. https://doi.org/10.1016/j.jhevol.2017.06.011.

Hoffman, Joseph I. 2016. “Reproducibility: Archive Computer Code with Raw Data.” *Nature* 534 (7607): 326–26.

Kuhn, Max. 2020. *Caret: Classification and Regression Training*. Manual.

Liaw, Andy, and Matthew Wiener. 2002. “Classification and Regression by randomForest.” *R News* 2 (3): 18–22.

Marwick, Ben. 2017. “Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation.” *Journal of Archaeological Method and Theory* 24 (2): 424–50.

Maté-González, Miguel Ángel, Lloyd A. Courtenay, Julia Aramendi, José Yravedra, Rocío Mora, Diego González-Aguilera, and Manuel Domínguez-Rodrigo. 2019. “Application of Geometric Morphometrics to the Analysis of Cut Mark Morphology on Different Bones of Differently Sized Animals. Does Size Really Matter?” *Taphonomic New Technologies* 517 (May): 33–44. https://doi.org/10.1016/j.quaint.2019.01.021.

McPherron, Shannon P. 2018. “Additional Statistical and Graphical Methods for Analyzing Site Formation Processes Using Artifact Orientations.” *PLoS One* 13 (1): e0190195. https://doi.org/10.1371/journal.pone.0190195.

McPherron, Shannon P., Aylar Abdolahzadeh, Will Archer, Annie Chan, Igor Djakovic, Tamara Dogandžić, George M. Leader, et al. 2020. “Introducing Platform Surface Interior Angle (PSIA) and Its Role in Flake Formation, Size and Shape.” *PLOS ONE* 15 (11): e0241714. https://doi.org/10.1371/journal.pone

.0241714.

Miller-Atkins, Galen, and L. S. Premo. 2018. “Time-Averaging and the Spatial Scale of Regional Cultural Differentiation in Archaeological Assemblages.” *STAR: Science & Technology of Archaeological Research* 4 (1): 12–27. https://doi.org/10.1080/20548923.2018.1504490.

Mraz, Veronica, Mike Fisch, Metin I. Eren, C. Owen Lovejoy, and Briggs Buchanan. 2019. “Thermal Engineering of Stone Increased Prehistoric Toolmaking Skill.” *Scientific Reports* 9 (1): 14591. https://doi.org/10.1038/s41598-019-51139-3.

Nichols, James A., Hsien W. Herbert Chan, and Matthew A. B. Baker. 2019. “Machine Learning: Applications of Artificial Intelligence to Imaging and Diagnosis.” *Biophysical Reviews* 11 (1): 111–18. https://doi.org/10.1007/s12551-018-0449-9.

Otárola-Castillo, Erik, Melissa G. Torquato, Hannah C. Hawkins, Emma James, Jacob A. Harris, Curtis W. Marean, Shannon P. McPherron, and Jessica C. Thompson. 2018. “Differentiating Between Cutting Actions on Bone Using 3D Geometric Morphometrics and Bayesian Analyses with Implications to Human Evolution.” *Journal of Archaeological Science* 89: 56–67. https://doi.org/10.1016/j.jas.2017.10.004.

Pante, Michael C., Matthew V. Muttart, Trevor L. Keevil, Robert J. Blumenschine, Jackson K. Njau, and Stephen R. Merritt. 2017. “A New High-Resolution 3-D Quantitative Method for Identifying Bone Surface Modifications with Implications for the Early Stone Age Archaeological Record.” *Journal of Human Evolution* 102 (January): 1–11. https://doi.org/10.1016/j.jhevol.2016.10.002.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Manual. Vienna, Austria: R Foundation for Statistical Computing.

Reeves, Jonathan S., Shannon P. McPherron, Vera Aldeias, Harold L. Dibble, Paul Goldberg, Dennis Sandgathe, and Alain Turq. 2019. “Measuring Spatial Structure in Time-Averaged Deposits Insights from Roc de Marsal, France.” *Archaeological and Anthropological Sciences* 11 (10): 5743–62. https://doi.org/10.1007/s12520-019-00871-y.

Sahle, Yonatan, Sireen El Zaatari, and Tim D. White. 2017. “Hominid Butchers and Biting Crocodiles in the African PlioPleistocene.” *Proceedings of the National Academy of Sciences USA* 114 (50): 13164. https://doi.org/10.1073/pnas.1716317114.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth Edition. New York: Springer.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

Yravedra, José, Miguel Ángel Maté-González, Juan Francisco Palomeque-González, Julia Aramendi, Verónica Estaca-Gómez, María San Juan Blazquez, Elena García Vargas, et al. 2017. “A New Approach to Raw Material Use in the Exploitation of Animal Carcasses at BK (Upper Bed II, Olduvai Gorge, Tanzania): A Micro-Photogrammetric and Geometric Morphometric Analysis of Fossil Cut Marks.” *Boreas* 46 (4): 860–73. https://doi.org/10.1111/bor.12224.