Journal of Archaeological Science: Reports 46 (2022) 103698

Contents lists available at ScienceDirect

Journal of Archaeological Science: Reports

journal homepage: www.elsevier.com/locate/jasrep

Multiple approaches to predicting flake mass

Guillermo Bustos-Pérez [a,b,c,*], Javier Baena Preysler [a]

[a] Departamento de Prehistoria y Arqueología, Universidad Autónoma de Madrid, Madrid, Spain
[b] Institut Català de Paleoecologia Humana i Evolució Social (IPHES), Zona Educacional 4, Campus Sescelades URV (Edifici W3), 43007 Tarragona, Spain
[c] Àrea de Prehistoria, Universitat Rovira i Virgili (URV), Avinguda de Catalunya 35, 43002 Tarragona, Spain

ARTICLE INFO

Keywords:
Lithic technology
Experimental archaeology
Flake weight
Machine learning
Deep learning

ABSTRACT

Predicting original flake mass is a major goal of lithic analysis. Predicting original flake mass allows for researchers to make estimations of remaining mass, lost mass, and other features. All these measures relate to the organization of lithic technology by past societies. The present work tests three different models to predict log of flake mass: multiple linear regression, random forest regression, and artificial neural networks (ANN). Estimations of flake mass were performed using the remaining features of flakes from an experimental assemblage. This assemblage was obtained by the expansion of a previous dataset through the inclusion of bigger flakes, allowing the analysis to account for the effects of sample size and value distribution. Correlation results show a large/strong relation between predictions and real outcome ($r^2 = 0.78$ in the best case). Comparison of the models affords insights into variable importance for predicting flake mass. Results show that (for the present dataset) multiple linear regression still stands as the best method for predicting log of flake weight. Additionally, transformation of predicted values from the multiple linear regression and true values to the linear scale reinforces the linear correlation above the 0.8 threshold.

1. Introduction

"Curated" is a key concept for the analysis of lithic technological organization (Andrefsky, 2009; Binford, 1979; Nelson, 1991; Spry and Stern, 2016). Initially, "curated" was defined as encompassing a series of behavioral patterns related to provisioning strategies (Binford, 1979, 1973). Further authors included tool transport, utilization in a wide range of tasks, anticipated production, hafting, and recycling (after the original tool had been discarded) among the adaptive behavioral strategies that defined curation. Shott (1996, 1989) proposed an alternative interpretation of the term "curation" as the "ratio of realized to potential utility." This shift in the definition of "curation" has deep implications for lithic analysis and the study of lithic technological organization, since it transforms "curation" into a continuous variable (Shott, 1996). A conception of "curation" as a continuous variable usually implies usually implies the degree of reduction or maintenance undergone by a tool (Shott, 2007, 1996, 1989). Additionally, the understanding of curation as a continuum also extends to the reduction approach (Dibble, 1995, 1987; Rolland and Dibble, 1990; Shott, 2007), which considers processes of resharpening as a major factor driving the presence and frequency of tool types. Ethnographic studies also emphasize the role of retouch in resharpening dulled edges, in changes in morphology, or in variations in artifact use as morphology changes throughout reduction (Casamiquela, 1978; Gould, 1968; Nuevo Delaunay et al., 2017; Shott and Weedman, 2007; White, 1967).

Usually, two approaches are employed to estimate the reduction and curation undergone by a retouched artifact. The first approach focuses on estimations made through measurements of reductions directly made on retouch. This has led to the proposal of several indexes that use different measurements, such as height of retouch, length of retouched edge, or projection of original angle (Bustos-Pérez and Baena, 2019; Eren et al., 2005; Hiscock and Clarkson, 2005; Kuhn, 1990; Morales et al., 2015). Although proposed indexes derived from this broad approach usually return high correlation values, they are conditioned by flake morphology, direction of retouch, or tool type (laterally retouched scrapers, endscrapers, bifacial products, etc.). Dibble (1995) noted the "flat flake problem" when applying Kuhn's (1991) general index of unifacial reduction (GIUR). The "flat flake problem" states that a flake with a trapezoidal cross section (where the dorsal face is mainly flat) will promptly reach maximum values of GIUR although reduction continues. The effects of the "flat flake problem" do not seem to be particularly severe on the GIUR (Hiscock and Clarkson, 2005), but they exemplify

* Corresponding author.
E-mail address: guillermo.bustos@uam.es (G. Bustos-Pérez).

https://doi.org/10.1016/j.jasrep.2022.103698
Received 30 June 2021; Received in revised form 13 September 2022; Accepted 20 October 2022
Available online 27 October 2022

the possible limitations that these indexes may possess as a result of flake morphology. Shott's (2005) extensive review of methods outlines the strengths and limitations derived from geometry, flake morphology, and assemblage suitability faced by each of the indexes.

The second approach aims to estimate original flake mass based on remaining features. This approach has the advantage of not being conditioned by tool type, direction of retouch, or flake morphology. Estimating original mass and comparing it with remaining mass can provide highly useful measures, such as percentage of mass remaining, amount of mass lost, and other features. All these measures are in keeping with the curation concept as a continuous and with the reduction approach. Initial controlled experiments showed highly promising results in the ability to predict flake mass from remaining features (Dibble and Pelcin, 1995). However, subsequent experiments based on the replication of knapping methods failed to obtain such high levels of correlation (Davis and Shea, 1998; Shott et al., 2000). Additionally, on some occasions, estimated original mass was lower than mass of flake after retouch (Davis and Shea, 1998). This posed an important drawback since, as Dibble (1998) states and Shott et al. (2000) reiterate: controlled experiments are useful only if results and variable relationships are extendible to the archaeological record. Further research has explored the estimation of flake mass through the combination of several variables (Dogandžić et al., 2015; Shott and Seeman, 2017) and the determination of the best variables with which to perform estimations (Bustos-Pérez and Baena, 2021).

Hiscock and Tabrett (2010) state the logical and analytical characteristics desirable for an index: inferential power; directionality; comprehensiveness; sensitivity; versatility; blank diversity; and scale independence. Following these characteristics, it can be stated that the first approach mentioned above is strong in inferential power, directionality, comprehensiveness, and sensitivity. On the other hand, present systems to estimate flake mass are strong in inferential power, comprehensiveness, sensitivity, versatility, blank diversity, and scale independence.

Most analysis focuses on the use of linear regression (usually using platform surface area as a proxy of flake mass) or the combination of several variables in multiple linear regression. The generalization afforded by statistical programming software (R Core Team, 2019; RStudio Team, 2019) allows for the implementation of regression models beyond simple linear regression. The present study uses and evaluates three common machine learning regression models (artificial neural networks, multiple linear regression, and random forest) for the estimation of flake mass. Additionally, each model provides insights into variable importance.

2. Methods

2.1. Experimental assemblage

The sample for analysis was composed of 500 experimentally knapped flakes using hard hammers. The flakes are categorized according to 30 knapping sequences wherein a wide variety of knapping methods were employed—hierarchical (Levallois and hierarchical discoid), bifacial (discoid), and unipolar—to generate the experimental sample, ensuring a wide range of morphologies (Boëda, 1995a, 1995b, 1993; Casanova i Martí et al., 2009; Terradas, 2003). This constitutes an expansion of a previous dataset employed for similar purposes (Bustos-Pérez and Baena, 2021), which increases the range of dimensions and mass of the assemblage. Although termination type influences flake mass, its influence on predicting original flake mass is considered residual or nonsignificant (Clarkson and Hiscock, 2011; Shott et al., 2000). The experimental assemblage was dominated by flakes with feather terminations (89.8 %), although other types of terminations were present (Table 1). All selected flakes were complete.

A key requirement of experimentations designed to estimate flake mass is that they are independent of external factors. To satisfy this requirement, the flakes were knapped with a wide variety of hammerstones. The raw material of hammerstones varied widely (quartz, quartzite, sandstone, and limestone), which allowed for a diverse range of morphologies and potential active percussion areas.

Table 1
Terminations type for the complete experimental assemblage.

	Feather	Hinge	Inflexed	Plunging	Step
Frequency	449	42	2	2	5

Comparison of the experimental dataset with the one from the previous study (Bustos-Pérez and Baena, 2021) shows an increase in the size and average mass of experimentally knapped flakes (Table 2; Fig. 1). While in the previous study 50 % of the flakes had mass values between 4.15 g and 14.02 g (Bustos-Pérez and Baena, 2021), in the present study 50 % of the flakes weighed between 5.87 g and 26.96 g. This indicates that the expansion of the dataset was achieved by the inclusion of heavier and bigger flakes. Additionally, exploratory visual analysis of flake mass (Fig. 2) shows a highly skewed distribution, with flakes weighing between 10 g and 20 g the most frequent.

2.2. Variable selection

Previous work (Bustos-Pérez and Baena, 2021) employed best subset selection (Furnival and Wilson, 1974; Hocking and Leslie, 1967) to obtain the best model with the best explanatory variables. The present work maintains the previously selected variables and uses an expanded version of the dataset. Variables employed to predict flake mass are: average thickness, $\log_{10}$ of maximum thickness, number of scars, amount of cortex, external platform angle (EPA), $\log_{10}$ of platform size, and $\log_{10}$ of platform depth (Fig. 3).

- **Average thickness:** mean flake thickness measured at 0.25, 0.50 and 0.75 of flake length (Eren and Lycett, 2012).
- **$\log_{10}$ of maximum thickness:** $\log_{10}$ transformation of the highest of the three values of average thickness.
- **Number of scars:** number of scars bigger than 5 mm (Scerri et al., 2016).
- **Amount of cortex:** measured on an ordinal scale. A slightly modified version of the triple cortex typology (Andrefsky, 2005; Fig. 4), with categories being: cortical (1), more than 50 % covered by cortex (2), <50 % covered by cortex (3), residual presence of cortex (4), and no cortex (5).
- **External platform angle (EPA):** relation in degrees between the platform and the dorsal surface of the flake. Measured with a manual goniometer.
- **$\log_{10}$ of platform size:** $\log_{10}$ transformation of platform size measured in accordance with Muller and Clarkson (2016).
- **$\log_{10}$ of platform depth:** $\log_{10}$ transformation of platform depth. Platform depth corresponds to the measure presented by Muller and Clarkson (2016).

Flake mass (in grams) was recorded using a Sytech SY-BS502 scale with 0.01 precision. All dimensional measures were performed using

Table 2
Descriptive statistics of experimental assemblage.

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Length (mm)	16.50	36.30	45.90	48.25	59.60	100.90
Width (mm)	14.90	31.18	39.00	40.56	46.83	85.50
Mean thickness (mm)	1.80	6.06	8.52	9.25	11.28	26.50
Platform surface (mm^2)	2.59	31.35	62.93	93.25	116.12	620.00
Weight (g)	1.14	5.87	12.97	21.39	26.96	200.73

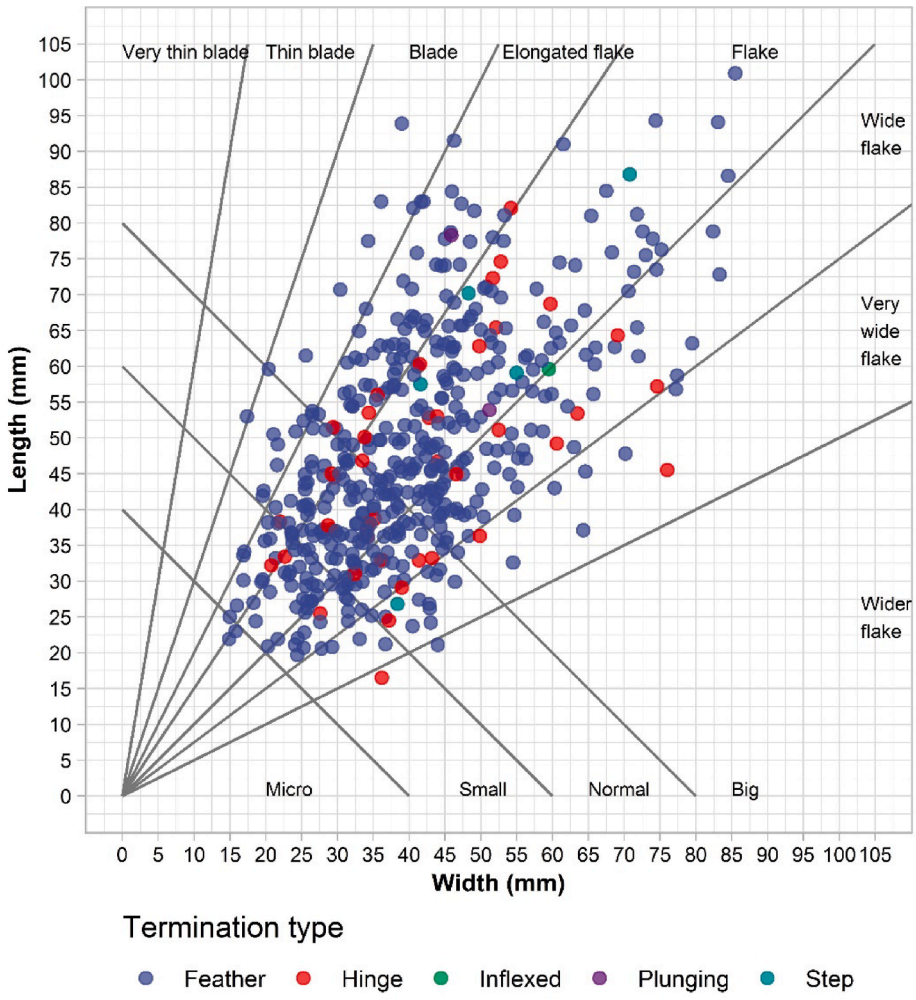

Fig. 1. Scatter plot of experimental assemblage (after Bagolini's [1968] scheme for plotting assemblage characteristics by length and width).

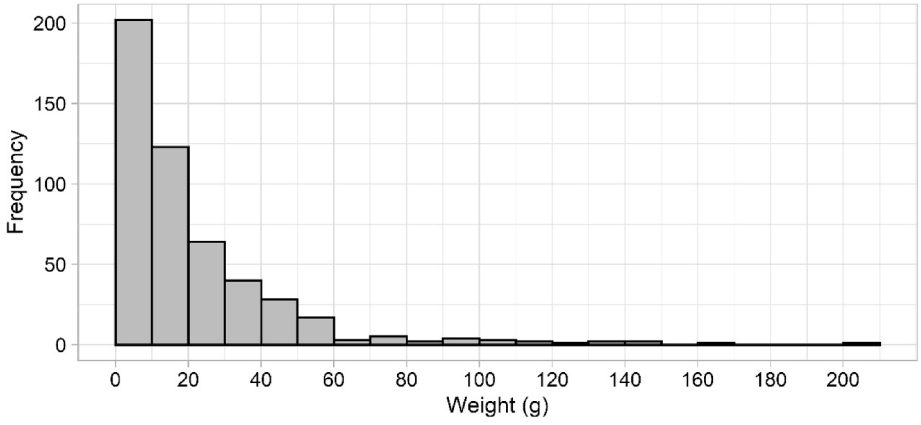

Fig. 2. Histogram distribution of flake weight. Bins represent intervals of 10 g.

digital calipers to 0.1 mm. Two different opinions exist on how EPA should be measured (Davis and Shea, 1998; Dibble and Pelcin, 1995), and the difficulty of obtaining accurate measurements when the platform or surface is curved is acknowledged. Two methods for recording flake platform exist. The first method (Andrefsky, 2005) uses the product of platform width and depth. The second method (Muller and Clarkson, 2016) first ascribes the general platform morphology to a geometric figure (rectangle, triangle, rhombus, trapezoid, or ellipse); the templated area of the geometric figures in combination with the corresponding measurements is then employed to calculate platform area. The second system has been shown to better approximate platform size when compared with measurements from scanning techniques and to not overestimate platform size (Muller and Clarkson, 2016). Additionally, previous studies have shown a clear preference for the second method as a variable for predicting flake weight (Bustos-Pérez and Baena, 2021). Thus, only measures of platform surface derived using the

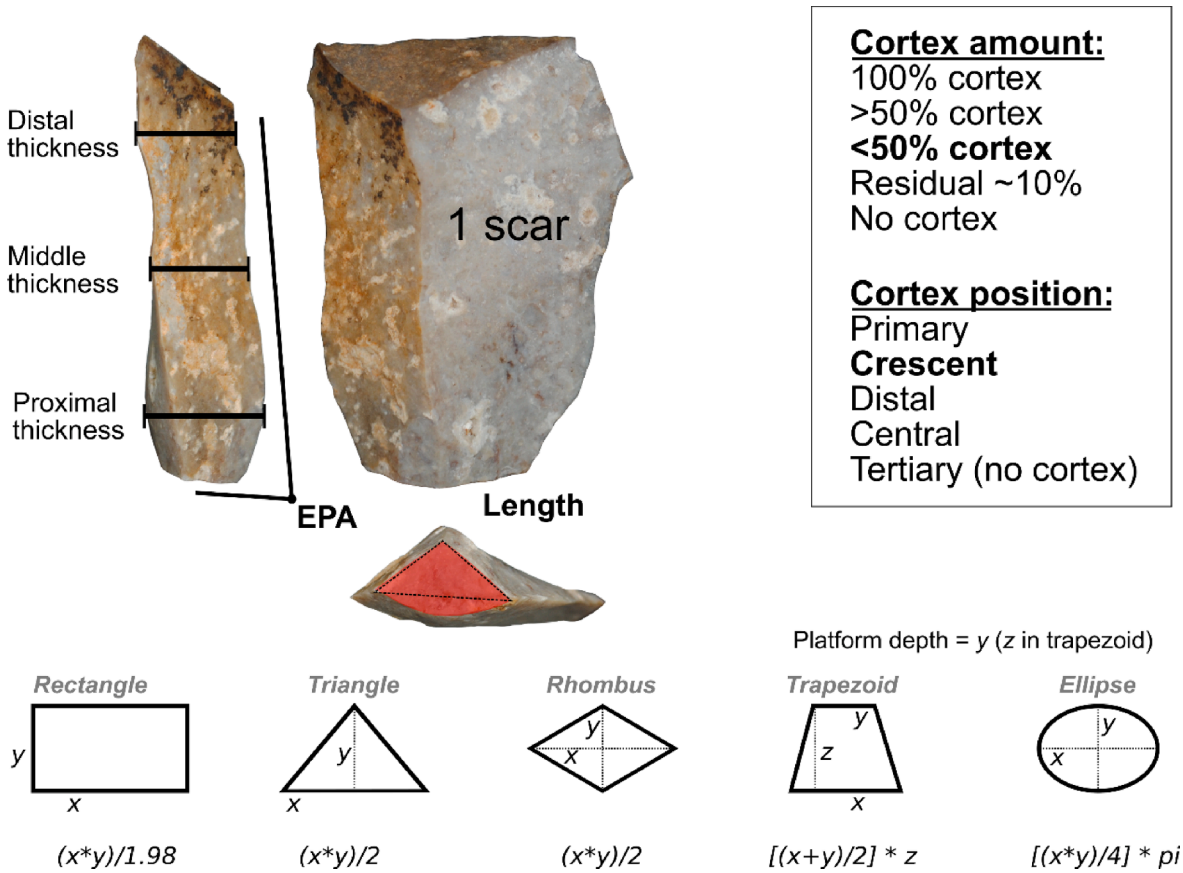

Fig. 3. Example of features employed in the present study: measurements of thickness, EPA, number of scars, relative amount of cortex, and platform surface following Muller and Clarkson (2016).

Fig. 4. Examples of experimental flakes with different amounts of cortex: 1) 100 % cortical; 2) >50 % cortical; 3) < 50 % cortical; 4) residual cortex; 5) no cortex.

second method (Muller and Clarkson, 2016) were employed.

Previous works have shown that it is easier to predict $\log_{10}$ of flake weight using $\log_{10}$ of platform size (Braun et al., 2008; Bustos-Pérez and Baena, 2021; Clarkson and Hiscock, 2011; Shott et al., 2000). Log transformations of variables are common, since they avoid negative results (necessary in the case of predicting flake weight), reduce skewed distributions, and can approximate parametric distributions (which favors the inferential power of models). In the present study, all logarithmic transformations refer to the common logarithm (base 10), and the target variable was the logarithmic transformation of flake weight.

Collinearity between predictors has previously been reported for platform surface and platform depth, and mean thickness and $\log_{10}$ of maximum thickness (Bustos-Pérez and Baena, 2021). For the present dataset, there is an important collinearity between $\log_{10}$ of maximum thickness and mean thickness ($r^2 = 0.879$) and an expected moderate/strong collinearity between platform depth and platform surface ($r^2 = 0.614$). Awareness of these collinearities is important, since collinearity affects variable importance (making it hard to separate the individual effect of a variable on the response), reduces the accuracy of the estimates in a multiple linear regression, and can result in counterintuitive estimates (James et al., 2013).

2.3. Regression methods

Three methods were employed to estimate $\log_{10}$ of flake mass: multiple linear regression, artificial neural networks (ANN), and random forest regression. The multiple linear regression extends the simple linear regression in such a way that it can directly accommodate multiple predictors (James et al., 2013, p. 71).

Artificial neural networks (ANN) are constituted by layers that are made on nodes (Fig. 5). Each ANN has an input layer made of input nodes (unprocessed features from the dataset) and an output layer made of output nodes. The output layer represents the target variable and can have one (in the case of regression when the target is to predict

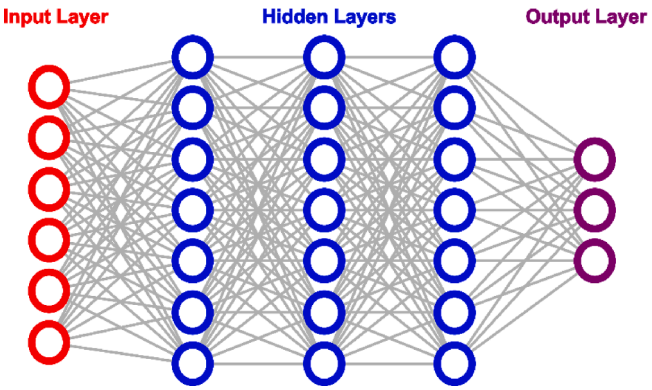

Fig. 5. Schematic representation of an ANN and its components.

numerical outputs) or several nodes (in the case of classification problems). Nodes between layers are connected with parameter values that are estimated when the ANN is fitted to the data. An ANN where nodes from the input layer are directly connected to the output layer (no hidden layers) is directly comparable to a multiple linear regression. To model more complex relationships, ANNs use hidden layers (each composed of a series of nodes) between the input and output layers that process signals from the input data and their interactions (Lantz, 2015). The structure of the ANN according to the number of hidden layers and number of nodes in each layer is referred to as its topology. The present work uses the R package *neuralnet* v.1.44.2 (Günther and Fritsch, 2010) to train the ANN with backpropagation (Rumelhart et al., 1986). For the present work, the ANN topology is limited to having only one or two hidden layers. The number of nodes of hidden layer 1 ranges between 1 and 4, while the number of nodes of hidden layer 2 ranges from 0 (no second hidden layer) to 4. All possible combinations were tested.

Random forest regressions select random samples of the data and build trees for prediction (Breiman, 2001). As a result, each tree is built from different data, and the average is used as prediction. This adds diversity, reduces overfit, and provides higher-resolution predictions (Lantz, 2015). Additional to the random selection of data, random forest can be further randomized by providing the number of trees to train, the number of possible variables to split at each node, and the minimal node size. These are hyperparameters, whose values need to be provided to the model before training. A Cartesian grid search (to test for all possible combinations) is performed on the abovementioned hyperparameters, with the number of trees to grow for each model ranging from 500 to 700 by 25; the number of possible variables to split at each node ranging from 1 to 5; and minimal node size ranging from 1 to 5. The R package *ranger* v.0.13.1 (Wright and Ziegler, 2017) was employed to train random forests.

2.4. Machine learning evaluation

With the exception of multiple linear regression, machine learning models and ANN are prone to overfit and need to be tested on data not previously seen by the models (Hastie et al., 2009; James et al., 2013). The present work employed a *k*-fold cross validation to estimate out-of-sample model performance. In *k*-fold cross validation, the dataset is randomly shuffled and divided into *k* folds. The first fold is employed as a test set, and the model is trained in the remaining folds. After this, the second fold is employed as a test set and the rest as a new training set. This process continues until all folds have served as a test set. Since the folds are shaped by the initial random shuffle, it is advisable to repeat this cycle a series of times. The present work employs a 10-fold cross validation (each fold having a sample of 50 elements) repeated 50 times.

Machine learning regression models are evaluated using proportion of variance explained (r^2 and *adjusted* r^2), visualization of regression plots, visualization of residuals (difference between actual and predicted value) plots, density plots of residuals, and descriptive statistics of residuals. Proportion of variance indicates how much of the observed variation is explained by the model (James et al., 2013). The addition of predictors results in an increasing r^2 irrespective of predictor contribution to the model and making it impossible to compare models with a different number of predictors. *Adjusted* r^2 is analogous to the r^2 but adjusted to the number of explanatory variables, thus making model comparison possible. *Adjusted* r^2 is required for the multiple linear regression model in order to make comparisons, while r^2 is required for the rest of the models.

Adjusted r^2 indicates how strongly predictions are related to the true value, but it does not indicate how far predictions fall from the true value (Lantz, 2015). Mean average error (MAE) and root mean squared error (RMSE) provide values of how far predictions fall from the true value (Lantz, 2015; James et al., 2013). MAE measures the average magnitude of errors, regardless of signal. RMSE also provides a measure of distance between predicted and actual values, although it punishes large errors. A perfect model will have MAE and RMSE values of 0. In general, better models will have lower values of MAE and RMSE.

A regression plot provides a scatter plot of predicted and true values along its regression line. In a good model, the regression line will pass through the center of all points, which will be evenly distributed above and below. The residuals plot provides a scatter plot of true values and residuals (difference between true value and predicted value), allowing for observation of whether there is systematic bias in the model. The residual plot of a good model will have the points evenly distributed on the zero value.

Collinearity of the abovementioned pairs of predictors is addressed by two means: first, by calculating variance inflation factor; and second, by comparing performance metrics values and residual distribution of the best models without collinear variables. Variance inflation factor provides a measure of correlation between predictors and their effects on the model. In the present study, variance inflation factor is calculated using the R package *car* v.3.1.0 (Fox and Weisberg, 2018). Thresholds for evaluating variance inflation factor values vary, although commonly, values between 1 and 10 are considered inconsequential, values between 10 and 30 are cause for concern, and values above 30 are considered seriously harmful (Marquardt, 1970; O'brien, 2007). At present, the package *car* only allows for the calculation of variance inflation factor for multiple linear regression. Although the different nature of the models can result in different effects of collinearity, results from calculating the variance inflation factor in the multiple linear regression can be extrapolated to the random forest and the ANN. While retrieving pairs of collinear variables allows for the determination of variable importance, it is important to keep in mind that collinearity between predictors does not affect predictions and the inferential power of a model (Alin, 2010; Paul, 2006).

The complete workflow was developed using the R language (v.4.0.2) in the RStudio IDE (v.1.4.1103; R Core Team, 2019; RStudio Team, 2019). The package *tidyverse* v.1.3.1 (Wickham et al., 2019) was employed for data manipulation and representation. The packages *leaps* v.3.1 (Lumley based on Fortran code by Alan Miller, T., 2020) and *lattice* v.0.20.45 (Sarkar, 2008) were employed additionally to the previously mentioned packages for model training. The package *caret* v.6.0.92 (Kuhn, 2008) was employed to set the validation methods and obtain the evaluation metrics of each model. All data, code complete workflow and models can be freely accessed in a Zenodo repository (https://zenodo.org/badge/latestdoi/432261142).

3. Results

3.1. Hyperparameter grid search

Fig. 6 presents the results of the hyperparameter Cartesian grid search for the random forest regression. In all cases, the hyperparameter of the number of variables to possibly split at each node was selected to

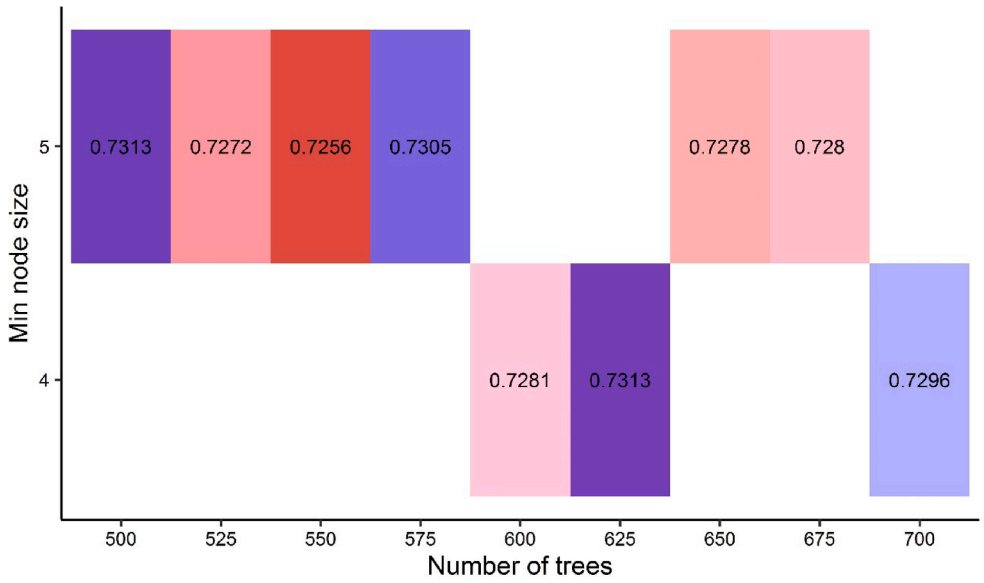

Fig. 6. Results of hyperparameter Cartesian grid search for the random forest. In all cases, the hyperparameter of the number of possible variables to split at each node was always selected to have a value of 2. Values represent the r^2 of each random forest model with the given combination of hyperparameters. Each value is obtained after a 10 × 50 cross validation.

have a value of 2. Linear correlation reaches its maximum at a minimum node size of 4 and 625 trees grown for the model ($r^2 = 0.73$). The second-best hyperparameter combination (minimum node size of 5 and 500 trees grown for the model) presents a marginally lower value of linear correlation (0.00005 lower).

The Cartesian grid search of the ANN topology (Fig. 7) indicates that increasing the number of nodes in the first hidden layer decreases linear correlation with the outcome and that increasing the number of layers and nodes results in lower values of r^2. Thus, the simplest ANN architecture (one hidden layer with one node) provides the highest correlation coefficient ($r^2 = 0.78$). The second-best topology (two hidden layers with one node at each layer) provides a marginally lower value (0.0005 lower).

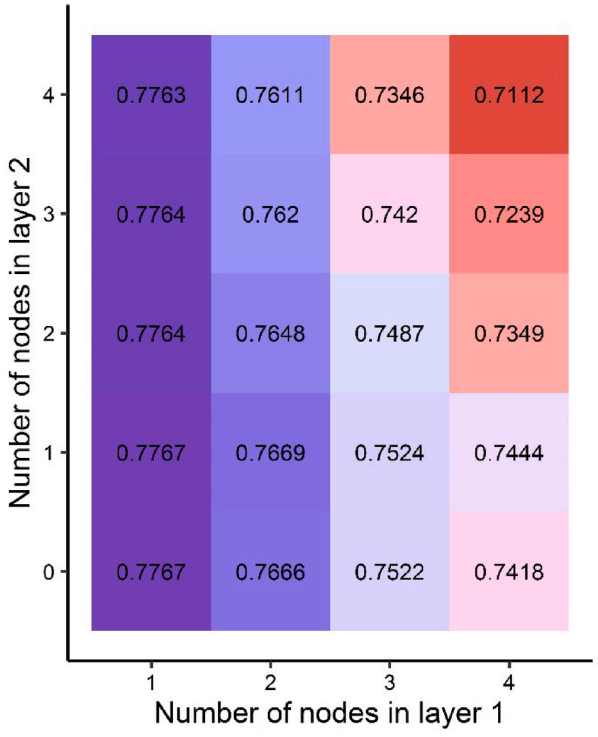

Fig. 7. Values of r^2 results for each combination of ANN topology. Each value is obtained after a 10 × 50 cross validation.

3.2. Model evaluation

Table 3 presents the precision metrics for each model. ANN and multiple linear regression perform similarly, with similar values of r^2 (0.78), RMSE (0.21), and MAE (0.17), although ANN performs slightly better. On the other hand, random forest regression performs slightly worse, with a lower value of r^2 (0.72) and higher values of RMSE (0.24) and MAE (0.19).

Visualization of regression plots for each model (Fig. 8) provides additional information on the performance of each model. The poor performance of random forest (lowest value of r^2) is reflected in a limited range of prediction, to between a minimum value of 0.55 and a maximum value of 1.76 for $\log_{10}$ of flake mass. As a result, data are not evenly distributed along the regression line. For the lowest values of prediction, most points fall below the regression line, while most data points fall above it for the highest values. ANN and multiple linear regression plots present similar patterns of distribution, with data evenly distributed along the regression line. Flakes with a $\log_{10}$ value of flake mass above 2 are more evenly distributed in the multiple linear regression than in the ANN.

Visual analysis of the scatter plot for observed and residual values (Fig. 9) allows the observation of model performance for different ranges of $\log_{10}$ of flake mass values. Residuals of the random forest present a systematic bias at the uppermost and lowest values of observed weight. In the case of flakes with a $\log_{10}$ value of 0.50, there is a systematic overestimation of size. In the case of flakes with a $\log_{10}$ value of 1.75, there is a systematic underestimation of values. ANN and multiple linear regression present very similar plots for observed values and

Table 3
Precision metrics of each model.

Model	Adjusted r^2	RMSE	MAE
ANN	0.777	0.209	0.166
Multiple linear regression	0.776	0.209	0.166
Random forest	0.721	0.239	0.192

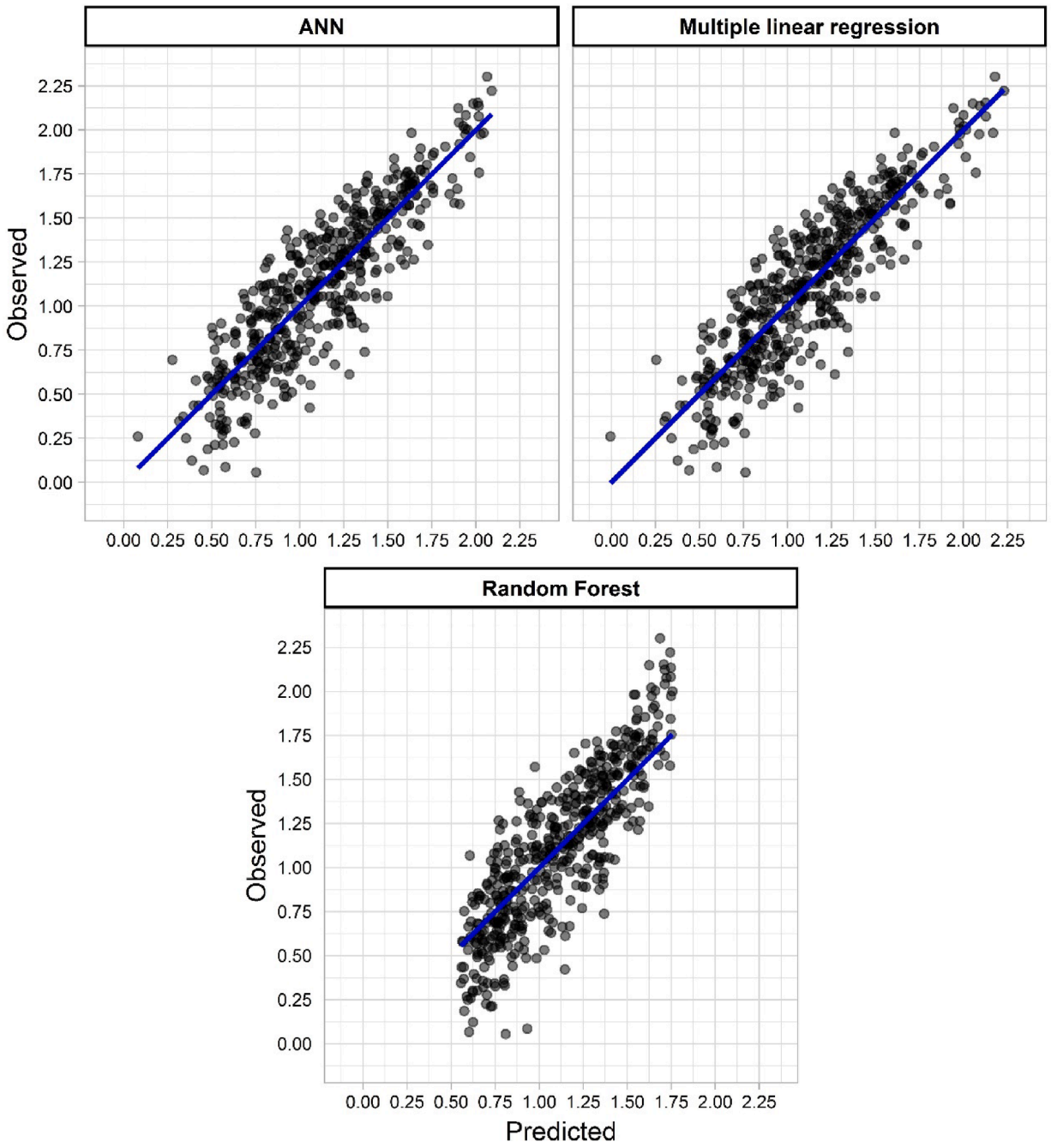

Fig. 8. Regression plots for each of the models for $\log_{10}$ of flake mass.

residuals. In both cases, residual values indicate a systematic overestimation of a $\log_{10}$ flake mass when the actual value is below 0.25.

Between values of 0.25 and 2, both models present a very similar performance, with residual values falling evenly on either side of the 0 value. ANN seems to present a slightly systematic underestimation of flakes with a $\log_{10}$ of flake mass above a value of 2. Multiple linear regression does seem to perform better for flakes with a $\log_{10}$ flake mass value above 2, with residual values falling evenly on either side of or very close to the cero value line.

Correlation between observed values and residuals allows for the evaluation of whether residuals increase along with increasing values of $\log_{10}$ of weight. ANN and multiple linear regression models present the same value of r^2 for correlation of observed values and residuals ($r^2 = 0.22$; $p < 0.01$) while random forest presents a higher value of correlation ($r^2 = 0.5$; $p < 0.01$).

Descriptive statistics of residuals (Table 4) and density plots (Fig. 10) allow for the evaluation of the dispersion range of residuals. All models present average and median residual values close to 0, with density curves peaking near this value (Fig. 10), which is indicative of good model performance. Fifty percent of residual values from the ANN model fall between the values of − 0.133 and 0.143, making for a distance of 0.276. Fifty percent of residual values from the multiple linear regression model fall between the values of − 0.137 and 0.134, making for a distance of 0.271. Fifty percent of residual values from the random forest model fall between the values of − 0.138 and 0.177, making for a distance of 0.315. This indicates that the multiple linear regression model concentrates 50 % of residual values in a slightly shorter range. This range is 0.005 shorter than the one from the ANN model. The random forest presents the highest dispersion range for 50 % of residual values.

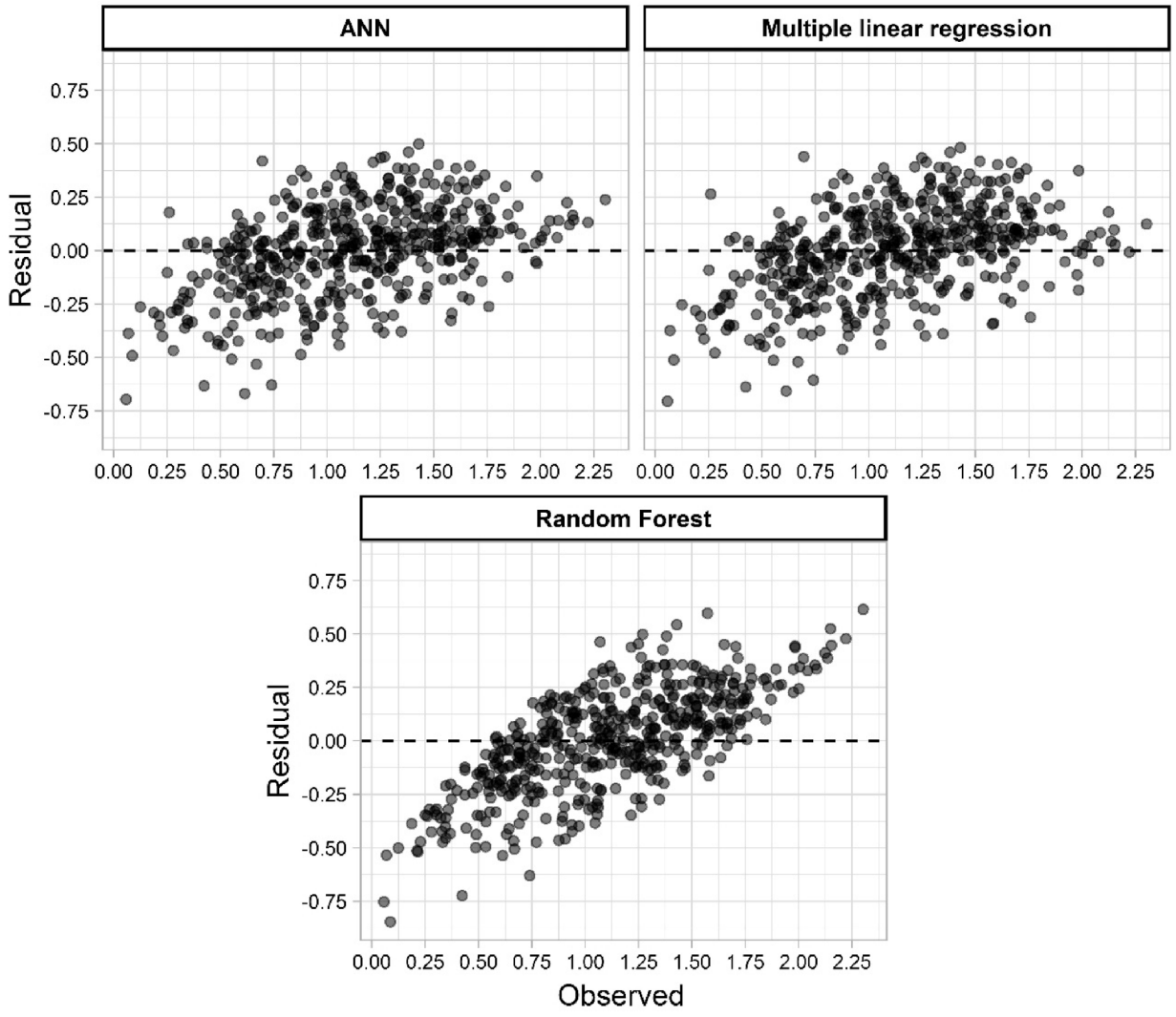

Fig. 9. Plots of observed and residual values for each of the models.

Table 4
Descriptive statistics of residuals for each model.

Model	ANN	Multiple linear regression	Random forest
Min.	−0.695	−0.705	−0.846
5th Percentile	−0.379	−0.371	−0.412
1st Quartile	−0.133	−0.137	−0.138
Mean	0.000	0.000	0.004
Median	0.024	0.020	0.013
3rd Quartile	0.143	0.134	0.177
95th Percentile	0.330	0.333	0.355
Max.	0.499	0.482	0.615

Ninety percent of residual values from the ANN model fall between the values of − 0.379 and 0.33, making for a distance of 0.709. Ninety percent of residual values from the multiple linear regression model fall between the values of − 0.371 and 0.333, making for a distance of 0.704. Ninety percent of residual values from the random forest model fall between the values of − 0.412 and 0.355, making for a distance of 0.767. Again, multiple linear regression concentrates 90 % of residuals in the shortest range. ANN presents a slightly wider range (a difference of 0.005), and random forest presents the widest range of the three models.

Exploratory data analysis of residuals according to termination type through box and violin plots shows possible differences in the distribution for the three models (Fig. 11). Comparison of residuals means according to termination type and for each model through *t*-test shows significant differences for the ANN model ($t = -2.5$; $p = .02$) and the multiple linear regression ($t = -2.52$; $p = .01$) but not for the random forest regression ($t = -1.82$, $p = .07$). In all models, the residuals mean of flakes with feather terminations fall near the 0 value (−0.007 in the case of ANN; −0.008 in the case of multiple linear regression; and − 0.002 in the case of random forest). Flakes with terminations other than feather tend to have a slightly higher mean of residuals values (0.07 in the case of ANN; 0.07 in the case of multiple linear regression; 0.06 in the case of random forest).

3.3. Linear transformation of predictions

Table 5 presents the performance metrics of each model after transforming true and predicted values back to the linear scale. ANN and multiple linear regression reinforce their correlation, while random forest decreases its r^2 value. Multiple linear regression provides the highest r^2 value ($r^2 = 0.813$), followed by ANN ($r^2 = 0.801$), indicating that multiple linear regression generalizes better to the linear scale. All models present lower RMSE values than the standard deviation value of weight of the experimental assemblage (24.83 g), which is indicative of good general performance.

Visualization of regression plots (Fig. 12) also supports the better

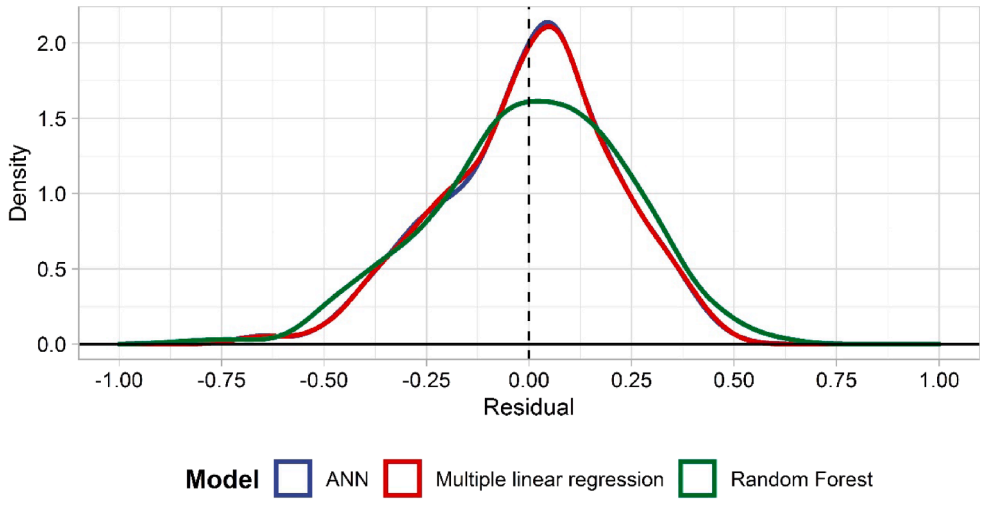

Fig. 10. Density plot of residuals of each model.

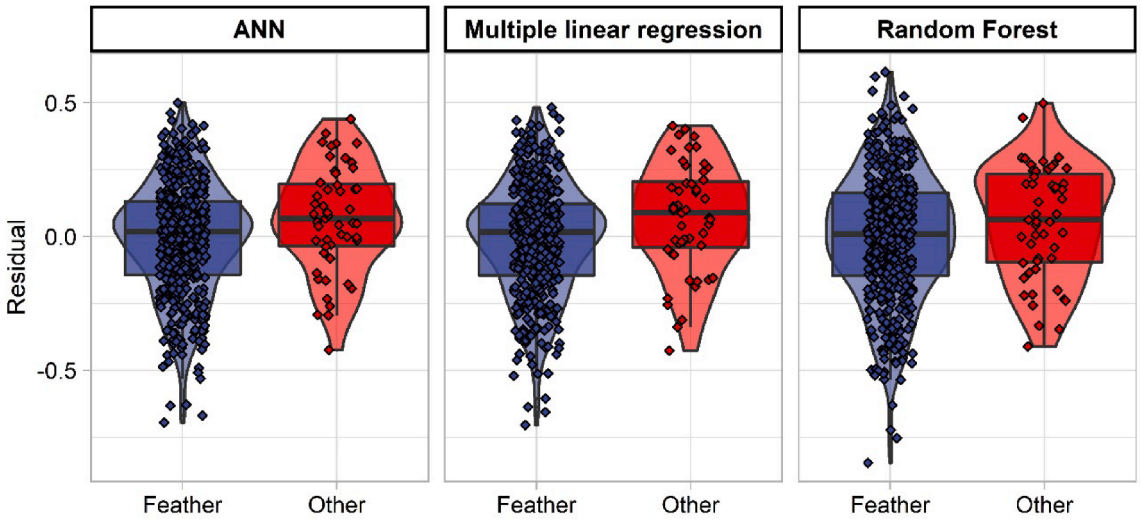

Fig. 11. Box and violin plots of residuals distribution according to termination.

Table 5
Performance metrics of each model in the linear scale.

Model	Adjusted r^2	RMSE	MAE
ANN	0.801	11.344	6.942
Multiple linear regression	0.813	10.853	6.793
Random forest	0.660	16.996	8.700

generalization of multiple linear regression to the linear scale. Random forest limits its maximum prediction to 57.2 g, resulting in a poor generalization to the linear scale. Due to this, residuals from the random forest (Table 6) indicate significant underestimations of flake weight, with an average underestimation of 4.6 g. Fifty percent of the residuals of the random forest range between overestimations of 2.64 g and underestimations of 7.06 g. Ninety percent of the residuals from the random forest range between overestimations of 10.35 g and underestimations of 29.74 g. Visual representation of residuals of the random forest by density plot (Fig. 13) shows that despite peaking on the 0 value, it presents a long tail of positive residuals as a result of underestimations of predictions.

ANN generalizes better to the linear scale (Fig. 12), with a higher range of predictions that reach a maximum value of 123 g. A density plot of residuals from the ANN presents a concentrated peak on the 0 value (Fig. 13) with a mean value of 1.82 g. Despite this, ANN residuals still present a slightly long tail of positive values for residuals as a result of some underestimations. Fifty percent of residuals from ANN range between overestimations of 2.52 g and underestimations of 5.55 g. Ninety percent of the residuals from ANN range between overestimations of 13.18 g and underestimations of 18.79 g.

As previously mentioned, multiple linear regression generalizes better to the linear scale (Fig. 13), with a maximum predicted value of 170 g. Residuals (Table 6) present an average 1.4 g value, with the density plot peaking near the 0 value and tails to the positive and negative values of similar length (Fig. 13). Fifty percent of residuals from multiple linear regression range between overestimations of 2.42 g and underestimations of 5.73 g. Ninety percent of residuals from multiple linear regression range between overestimations of 13.18 g and underestimations of 18.79 g. Thus, multiple linear regression presents the shortest range in a 90 % concentration of residuals.

3.4. Collinearity and variable importance

Table 7 presents the variance inflation factor of each of the predictors in the multiple linear regression model. Although mean thickness and $\log_{10}$ of maximum thickness present the highest values (8.43 and 8.88

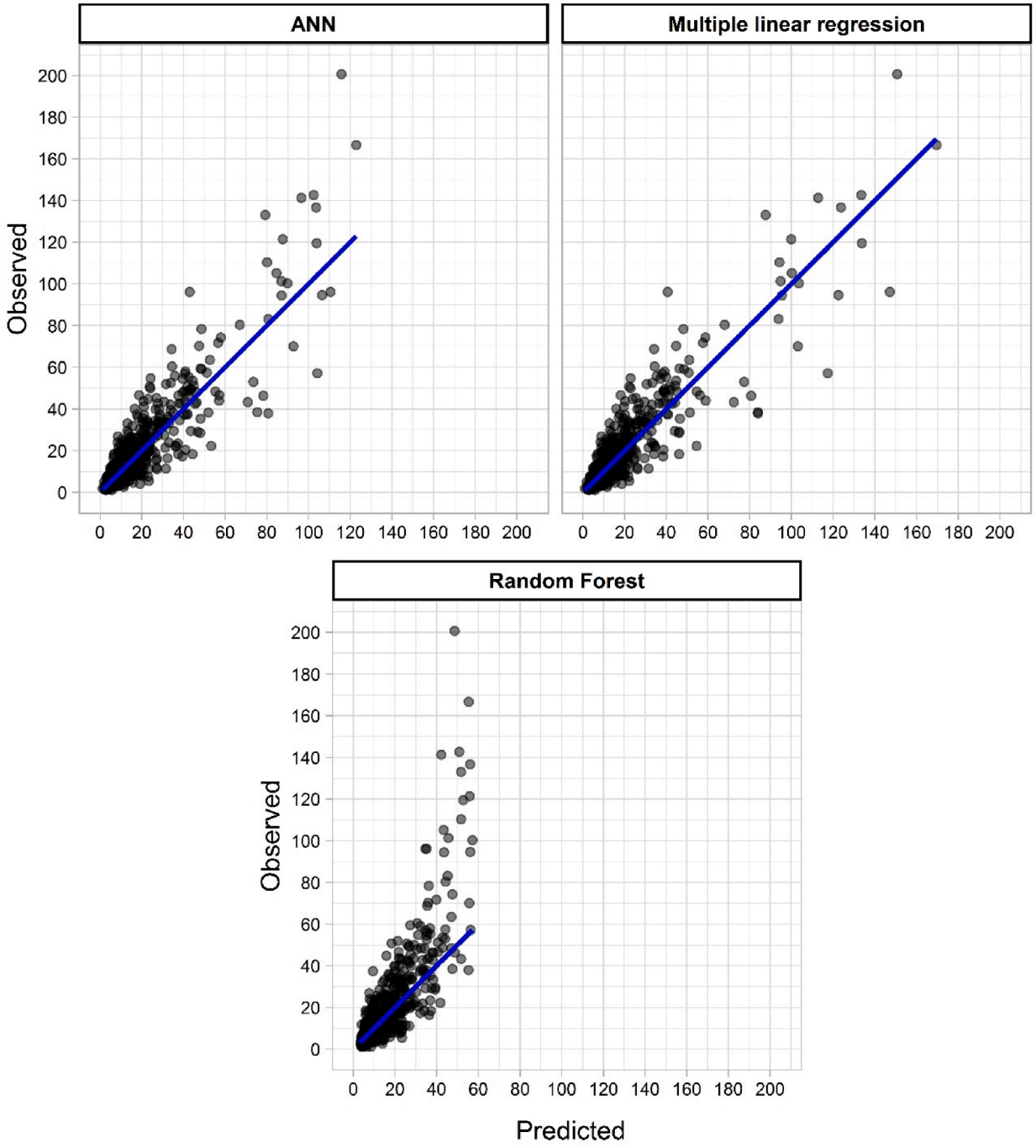

Fig. 12. Regression plots of predicted and true values transformed back into the linear scale.

Table 6
Descriptive statistics of residuals in the linear scale for each model: weight (g).

Model	ANN	Multiple linear regression	Random forest
Min.	−47.083	−60.223	−20.066
5th Percentile	−13.790	−13.182	−10.351
1st Quartile	−2.516	−2.422	−2.641
Mean	1.816	1.400	4.612
Median	0.476	0.331	0.334
3rd Quartile	5.553	5.725	7.060
95th Percentile	19.809	18.791	29.742
Max.	84.989	55.585	152.076

respectively), neither of the predictors presents a value above 10, indicating that collinearity is irrelevant.

Table 8 presents model performance metrics of the three tested methods when collinear variables are retrieved. For multiple linear regression and ANN, performance metrics were the best when mean thickness and $\log_{10}$ of platform depth were retrieved, and these models presented the lowest performance values when average thickness and $\log_{10}$ of platform depth were kept as predictive variables. In the case of random forest regression, performance values were lowest when mean thickness and $\log_{10}$ of platform surface were excluded as predictive variables. However, these values are similar to the ones obtained when $\log_{10}$ of maximum thickness and $\log_{10}$ of platform surface were excluded when training the random forest regression.

All models with the best combinations of noncollinear variables presented slightly lower performance metrics (r^2, RMSE, MAE), but

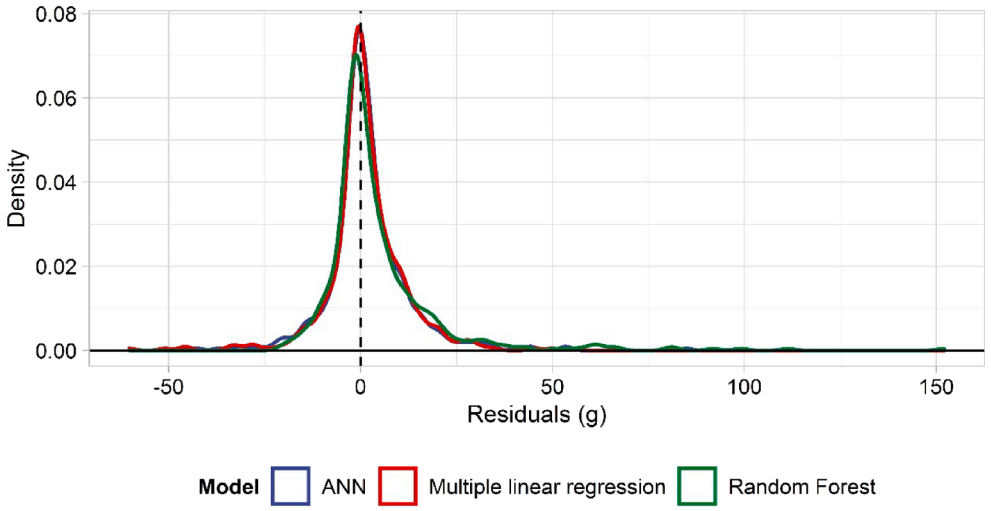

Fig. 13. Density plot of residuals in the linear scale of each model.

Table 7
Variance inflation values of each of the predictors in the multiple linear regression model.

Mean Thick.	Cortex	No. of Scars	EPA	Log. Max. Thick.	Log_{10} Plat. Size	Log_{10} Plat. Depth
8.43	1.97	1.84	1.18	8.88	4.76	5.14

Table 8
Descriptive statistics of model performance after retrieving collinear variables.

Model	Variables excluded to avoid collinearity	r^2	RMSE	MAE
Multiple linear regression	Mean Thickness & Log_{10} of Plat. Depth	0.761	0.216	0.173
Multiple linear regression	Mean Thickness & Log_{10} of Plat. Surf.	0.744	0.224	0.178
Multiple linear regression	Log_{10} of Max. Thick. & Log_{10} of Plat. Depth	0.755	0.219	0.175
Multiple linear regression	Log_{10} of Max. Thick. & Log_{10} of Plat. Surf.	0.736	0.227	0.182
ANN	Mean Thickness & Log_{10} of Plat. Depth	0.764	0.215	0.171
ANN	Mean Thickness & Log_{10} of Plat. Surf.	0.748	0.222	0.177
ANN	Log_{10} of Max. Thick. & Log_{10} of Plat. Depth	0.765	0.214	0.170
ANN	Log_{10} of Max. Thick. & Log_{10} of Plat. Surf.	0.747	0.222	0.177
Random forest	Mean Thickness & Log_{10} of Plat. Depth	0.707	0.248	0.200
Random forest	Mean Thickness & Log_{10} of Plat. Surf.	0.695	0.253	0.204
Random forest	Log_{10} of Max. Thick. & Log_{10} of Plat. Depth	0.714	0.246	0.197
Random forest	Log_{10} of Max. Thick. & Log_{10} of Plat. Surf.	0.698	0.251	0.201

similar to the ones from models including collinear variables. When the predicted values from models with collinear variables are compared to the predicted values of models with no collinear variables, no significant differences are present for the multiple linear regression ($t = -0.002$, $p = .998$), the ANN ($t < 0.001$, $p = 1$), or the random forest ($t = -0.08$, $p = .936$). Table 9 presents performance metrics when predictions and observations from the best models without collinear variables are transformed into the linear scale. Again, the exclusion of collinear variables results in slightly increased values of RMSE and MAE, and a lower r^2 value. Additionally, when collinear variables are excluded, RMSE and MAE indicate that ANN generalizes better to the linear scale. Both results of performance metrics (in the logarithmic and linear scales) indicate that the predictive power of models is slightly diminished when collinear variables are excluded from model training, but this diminution is not significant.

Table 9
Performance metrics of best models without collinear variables in the linear scale.

Model	r^2	RMSE	MAE
ANN	0.777	12.011	7.222
Multiple linear regression	0.779	12.445	7.349
Random forest	0.619	18.022	8.902

Retrieving collinear variables allows for the evaluation of variable importance for each model (Fig. 14). All models consider measures of thickness (mean thickness in the case of random forest, and log_{10} of maximum thickness in the case of multiple linear regression and ANN) to be of maximum importance. Both ANN and multiple linear regression consider the importance of the rest of the variables to sit in the same order: relative amount of cortex is considered the second most important variable, followed by number of scars and log of platform size. The order of importance of variables does change in the random forest regression, with log of platform size being the second most important variable, followed by relative amount of cortex and number of scars. EPA is considered the variable of least importance by all models. ANN does attribute some importance to the prediction of log_{10} of flake weight. A 0 value of importance for EPA is obtained in the random forest and multiple linear regression models (Fig. 14). However, coefficient estimates of the multiple linear regression do consider EPA a significant predictor ($p = 0.015$).

4. Discussion

The present study has expanded a previous dataset (Bustos-Pérez and Baena, 2021) with bigger and heavier flakes and applied three common machine and deep learning regression algorithms (multiple linear regression, ANN, and random forest) to determine log_{10} of flake mass based on previously selected variables (Bustos-Pérez and Baena, 2021). Additionally, predicted results and true values have been transformed back to the linear scale to explore further relations. ANN and multiple linear regression present similar r^2 values (0.78 in both cases), performance metrics, and residual distributions in the logarithmic scale.

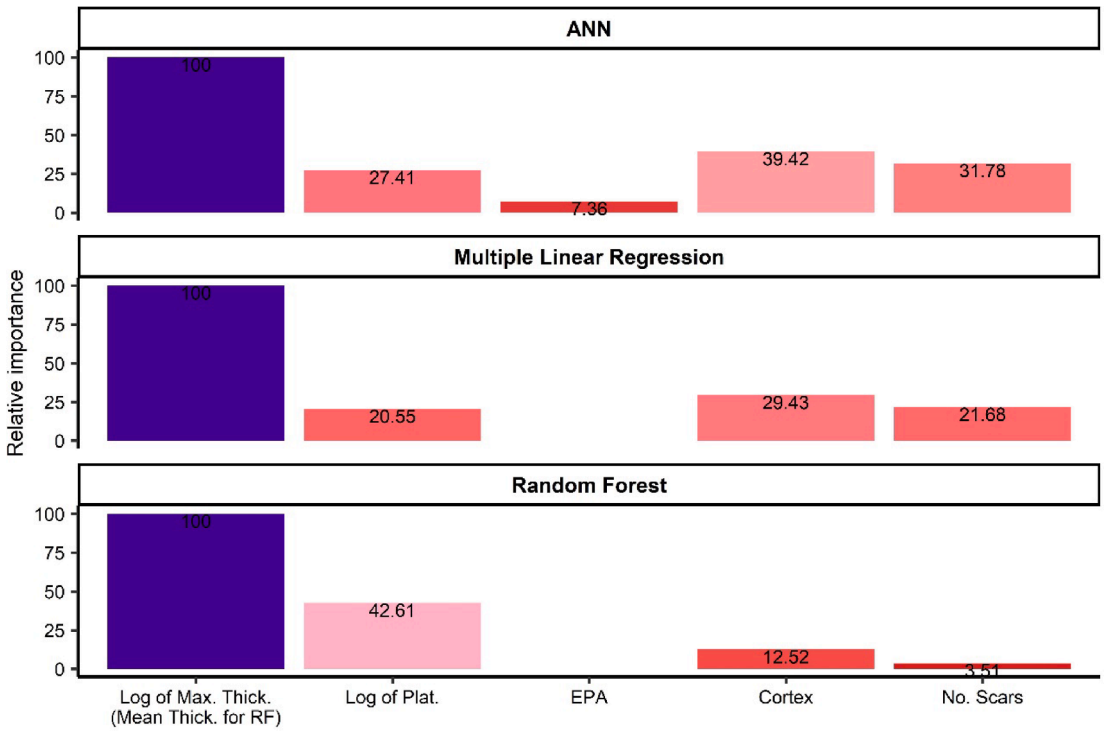

Fig. 14. Scaled variable importance (0–100) of models with the best combination of noncollinear variables.

Comparatively, random forest performed poorly with a lower r^2 (0.72), worse performance metrics, and clearly biased distributions of residuals. Transformation of predicted and true values back to the linear scale slightly reinforced an increase in ANN's and multiple linear regression's r^2 values (0.8 and 0.81 respectively), while generating a decrease in the random forest r^2 (0.66). Results from residuals analysis and distribution, performance metrics, and regression plots indicate that multiple linear regression is the model that best generalizes to the linear scale when collinear variables are included.

Results obtained by removing collinear variables allowed for the evaluation of predictor impact on model performance and reliability of predictions. Results from variance inflation factor analysis and comparison of residuals distribution indicate that no statistical difference exists between predictions that include collinear variables and predictions that exclude them. This indicates that predictions from models including collinear variables are as reliable as the ones that do not include collinear variables. However, when collinear variables are excluded from model training, ANN generalizes slightly better to the linear scale than multiple linear regression.

Removing collinear variables also allowed for a better evaluation of variable importance for each model. In all cases, measures of thickness (average thickness in the case of random forest and $\log_{10}$ of maximum thickness in the case of ANN and multiple linear regression) are considered of maximum importance for predicting $\log_{10}$ of flake weight. After measures of thickness, the order of importance of the variables is the same for ANN and multiple linear regression, although their relative values differ. These variables and their order are: relative amount of cortex, number of scars, and $\log_{10}$ of platform size. However, for these variables, importance values are much lower in the case of multiple linear regression, suggesting that ANN is diversifying the importance of predictors, while multiple linear regression relies more on $\log_{10}$ of maximum thickness. Random forest (the model with the worst performance metrics) considered mean thickness as the key feature for determining $\log_{10}$ of flake mass, followed by $\log_{10}$ of platform surface. Amount of cortex and number of scars were considered as variables of minor importance, and EPA was given no importance at all by random forest. ANN is the only model to assign a value of relative importance to EPA (although being the least important variable for that model), while multiple linear regression accords it a 0 value of importance, although it is considered statistically significant in the model coefficients. These results indicate that although EPA might be a significant predictor, its importance for predicting $\log_{10}$ of flake mass is minimal when compared with other predictors.

Results from the present study show significant differences in residual distribution according to flake termination when determining flake weight. This suggests that flake termination plays a significant role when predicting original weight. Although not included as a predictor in the present study, further research might benefit from coding termination type (as "feather" or "other") to determine original flake weight.

Several works have addressed the estimation of flake mass from remaining variables. The first element of comparison is with the previous version of this dataset (Bustos-Pérez and Baena, 2021). The expansion of the dataset through the inclusion of bigger flakes has resulted in an increased linear correlation for multiple linear regression and ANN. A possible interpretation for the increased value of r^2 is that as flake mass increases, the importance of variables shifts, and more variance is captured by the model. Thus, it can be inferred that smaller flakes with relatively low values of mass have a higher variability among the selected variables, resulting in a lower r^2. Additionally, the inclusion of bigger flakes and the addressing of collinearity has also resulted in changes in variable importance. The previous version of the dataset (Bustos-Pérez and Baena, 2021) considered cortex amount and number of scars as the third and fourth most significant variables of the model (behind $\log_{10}$ values of maximum thickness and platform size). In the present study, all three models emphasize even more markedly the importance of thickness measures (especially $\log_{10}$ of maximum thickness). Cortex amount and number of scars are respectively considered the second and third most important variables (although with values of importance considerably lower than measures of thickness), while $\log_{10}$ of platform size is considered the fourth most important variable (second in the case of random forest regression). EPA is the variable most heavily affected by the increasing size of flakes, since it goes from being a significant variable (Bustos-Pérez and Baena, 2021) to being considered almost irrelevant by most of the models. Possible sources for the lack of

resolution when measuring EPA can be attributed to the use of manual goniometers. Previous studies have shown high variability when obtaining angle measurements from manual goniometers (Dibble and Bernard, 1980; Morales et al., 2015), and this lack of resolution can also be applied to EPA measurements. Previous studies have acknowledged the difficulty of measuring EPA with manual goniometers or have produced different interpretations of the flake exterior surface (Davis and Shea, 1998; Dibble and Pelcin, 1995; Shott et al., 2000). This could be a possible explanation for the low importance attributed to EPA by the models of the present study. Four variables (average thickness, $\log_{10}$ of maximum thickness, number of scars, and relative amount of cortex) employed in the present study can be altered by extensive processes of retouch. Therefore, caution and an overall estimation of the integrity of the predictors are highly advisable before applying the model.

Most previous works employ a single linear regression to estimate flake mass in the $\log_{10}$ scale or the linear scale. Shott et al. (2000) estimate $\log_{10}$ value of flake mass based on $\log_{10}$ values of platform area measured with digital calipers, obtaining an r^2 of 0.67. Braun et al. (2008) obtain a similar linear correlation value of $\log_{10}$ of flake mass ($r^2 = 0.66$) for their sample of flakes, employing digital calipers to measure platform area and using $\log_{10}$ of platform area. This value increases drastically when they measure platform area of the same flakes with digital photographs ($r^2 = 0.865$). However, this highly promising result was nuanced by Clarkson and Hiscock's (2011) estimations of flake mass using 3D measures of platform surface. Clarkson and Hiscock (2011) report an r^2 of 0.49 for their total experimental sample and indicate that diverse flake assemblages reduce the capability of estimating flake mass based on platform area. Maloney (2020) estimates flake mass (g) based on 3D measures of platform surface and obtains an r^2 value of 0.411 for the complete sample. Orellana Figueroa et al. (2021) use virtual knapping and neural networks to predict flake shape (width and length) along with flake volume, reporting an $r^2 = 0.771$ and an RMSE = 0.763. Dogandžić et al. (2015) approach the estimation of flake mass using a multiple linear regression that uses platform width and depth, EPA, and blank thickness, obtaining an r^2 value of 0.75 for the cubic root of weight. Although this is a high degree of correlation, they express concerns about the model accuracy (Dogandžić et al., 2015). Shott and Seeman (2017) follow in the steps of Dogandžić et al. (2015) and use a multiple linear regression with platform surface (as the product of platform width and depth), flake thickness, and EPA as variables, resulting in an r^2 value of 0.73. Archer et al. (2018) and Morales et al. (2015) employ 3D scanning techniques to estimate original flake mass (using geometric morphometrics or geometrical relationships), obtaining respective r^2 values of 0.879 and 0.891. However, these resources are not as widespread among archaeologists and are hard to apply to numerous collections, making desirable the estimation of flake mass by the use of non–laser scanning techniques.

In the present study, the transformation back to the linear scale of $\log_{10}$ values of observed and predicted mass resulted in changes in the coefficients of determination for the three models. ANN and multiple linear regression reinforced their correlation values above the 0.8 threshold, while random forest suffered a decrease. It should not be surprising that ANN and multiple linear regression behave similarly, since an ANN with one node in only one hidden layer is considered to be a distant cousin of multiple linear regression (Lantz, 2015). Ideally, predictions of flake weight would be done in the linear scale, since they are easier to interpret. Changes in the correlation coefficient when shifting from the logarithmic to the linear scale can be considered a result of the distribution of residuals, original data, and the nature of the logarithmic scale. Here, a possible explanation resides in the skewed distribution of flake mass of the present dataset, which is directly associated with predictions of the model. The higher the value of $\log_{10}$ of flake mass, the more imperative it becomes that care is taken to ensure accuracy of prediction. This can be illustrated with the following example: a flake with a $\log_{10}$ value of mass of 1 (10 g) and a $\log_{10}$ predicted value of 1.1 (12.59 g) will result in a residual of 2.59 g in the linear scale. A flake with a $\log_{10}$ value of mass of 2 (100 g) and a predicted $\log_{10}$ value of 2.1 (125.89 g) will result in a residual of 25.89 g. Thus, although in the logarithmic scale the residuals have the same value, in the linear scale they result in a difference 10 times bigger. Despite this drawback, the multiple linear regression model generalizes relatively well to the linear scale but requires further evaluation.

Reproducibility is a key issue for archaeology (and all sciences), and the existence of independent researches reaching similar results on correlations and variable importance is key for the validation of a method (Marwick, 2017). The present study differs slightly in variable importance from previous studies that use multiple linear regression and measures from flakes (Dogandžić et al., 2015; Shott and Seeman, 2017). Measures of flake thickness (either mean flake thickness or $\log_{10}$ of maximum thickness) have been previously acknowledged as important variables for estimating original flake mass (Dogandžić et al., 2015; Shott and Seeman, 2017). The present study reinforces this relationship, since the evaluation of variable importance when collinearity is excluded results in $\log_{10}$ of maximum thickness being considered the key variable for estimating original flake mass. Log_{10} values of platform surface have also been acknowledged as an important predictor of flake mass (Braun et al., 2008; Bustos-Pérez and Baena, 2021; Clarkson and Hiscock, 2011; Davis and Shea, 1998; Shott et al., 2000). However, in the present study, although $\log_{10}$ of platform size remains as an important variable, its relative importance is dwarfed due to the heavy importance of $\log_{10}$ of maximum thickness. Log_{10} of platform surface does gain considerable importance in the random forest, the only model that uses mean thickness. This is possibly indicating that the importance of $\log_{10}$ of platform size was perhaps overemphasized in studies where mean thickness was employed as a predictor instead of $\log_{10}$ of maximum thickness (Braun et al., 2008; Bustos-Pérez and Baena, 2021; Clarkson and Hiscock, 2011; Davis and Shea, 1998; Shott et al., 2000).

Hiscock and Tabrett (2010) advocate for the use of indexes presenting logical and analytical qualities. These are: inferential power (advocating for those indexes with an r^2 above 0.8); directionality (increasing values as reduction proceeds); comprehensiveness (capability of operating at all levels of reduction); sensitivity; versatility (applicable to different types and positions of retouch); blank diversity; and scale independence.

Estimation of original flake mass would serve as an ideal index to satisfy these logical and analytical criteria, since it would allow for the comparison of remaining mass with original mass (being able to estimate derived measures such as amount of mass lost, percentage of mass remaining, etc.) and it could be applied to different types of blanks. In the present study, the estimation of original flake mass using multiple linear regression in the linear scale provided an r^2 above the 0.8 threshold, although as previously mentioned, caution is required. Theoretically, this would fulfill the seven logical and analytical requirements put forward by Hiscock and Tabrett (2010). However, Davis and Shea (1998) showed how estimations of flake mass might result in lower values than those of the flake after undergoing retouch. This drawback violates the logical and analytical principle of directionality of indexes and requires further evaluation before applying the present model. One of the reasons for this drawback might reside in Hiscock and Tabrett's (2010) advocation for the use of r^2. Use of r^2 might not be such a good indicator of model performance in predicting original flake mass, for two reasons. First, and as previously mentioned, r^2 does not indicate how far (on average) predictions fall from the true value (which is provided by MAE and RMSE). Evaluating models according to how far predictions fall from the true value can help avoid the contradiction pointed out by Davis and Shea (1998), in which the estimated flake mass is lower than the mass of the same retouched flake. Second, it has been well demonstrated that different distributions of data might result in similar or identical values of r^2 (Anscombe, 1973; Chatterjee and Firat, 2007), outlining the need for graphical evaluation of regression plots and distribution of residuals.

Research on the previous version of this dataset (Bustos-Pérez and

Baena, 2021) also provided slightly higher, but very similar, values of RMSE (0.217 vs 0.209) and MAE (0.178 vs 0.166). This comparison, along with performance metrics and plots from previous similar research (Bustos-Pérez and Baena, 2021; Dibble and Pelcin, 1995; Dogandžić et al., 2015; Maloney, 2020; Shott et al., 2000; Shott and Seeman, 2017) seems to delineate a limit in the ability to predict original flake mass based on remaining flake attributes, even with the inclusion of new algorithms. This limitation might be the result of variables that do not survive the archaeological record, or be due to the need to include new variables for analysis. The unexplained portion of variance might be related to variables that barely survive or are hard to determine in the archaeological record (such as hammerstone speed, morphology, and density).

Several variables have the potential to increase the inferential power of models in estimating original flake mass and can be explored in further research. Bradbury and Carr (1999) used scar density per flake surface (which can act as a replacement of simple scar count), showing promising results. McPherron et al. (2020) introduced the platform surface interior angle (PSIA), which is also showing promising results in estimating original flake mass. Additional variables such as height of retouch (Bustos-Pérez and Baena, 2019; Kuhn, 1990) in combination with remaining flake mass might also have the potential to increase the inferential power of models. The system for recording cortex amount might be a source of disincentive for applying the present study models. Systems for recording amount of cortex might vary among lithics analysts, with the added complication of being an ordinal variable used in regression analysis. Thus, further research might benefit from excluding this variable and then evaluating model performance.

Other indexes, such as GIUR (Hiscock and Clarkson, 2005; Kuhn, 1990), the estimated reduction percentage (ERP; Eren et al., 2005), 3DERP (Morales et al., 2015), or the combination of retouched edge length and average retouched height known as the AvtL (Bustos-Pérez and Baena, 2019), guarantee directionality and sensitivity and are reported to have higher inferential power in most cases.

5. Conclusion

The present research deals with the estimation of flake mass using the remaining features of a flake. Estimating original flake mass is key for the estimation of curation and for making inferences on the organization of the lithic technology of past societies. The experimental sample employed to estimate flake mass was obtained after expanding a previously existing dataset (Bustos-Pérez and Baena, 2021) by the inclusion of bigger flakes. The inclusion of bigger flakes resulted in a higher correlation value (r^2), although measures of distance between predictions and true values (RMSE and MAE) did not vary substantially from the previous version of the dataset. Addressing collinearity ensured the quality of predictions and variable importance. Predictions from models that include collinear variables do not statistically differ and are as reliable as predictions from models without collinear variables. Log_{10} of maximum thickness stands out as the most important variable for predicting flake mass. Multiple linear regression and the simplest ANN have been shown to be the best models for estimating log_{10} of flake mass in the present dataset.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data and code are freely accessible.

Acknowledgments

The authors wish to thank the co-editor and the three anonymous reviewers for their comments and suggestions. This article is the result of the research projects “Como, Quien Y Donde?: Variabilidad De Comportamientos En La Captación Y Transformación De Los Recursos Liticos Dentro De Grupos Neandertales 2” (HAR2016-76760-C3-2-P) financed by Agencia Estatal de Investigación (AEI), Fondo Europeo de Desarrollo Regional (FEDER); and “En Los Limites De La Diversidad: Comportamiento Neandertal En El Centro Y Sur De La Penisula Iberica” (ID2019-103987 GB-C33) financed by the Programa Estatal de Generación de Conocimiento y Fortalecimiento Científico y Tecnológico del Sistema de I + D + i y de I + D + i Orientada a los Retos de la Sociedad, del Plan Estatal de Investigación Científica y Técnica y de Innovación (2017–2020). Development of the experimentation and analysis of the materials were undertaken at the Laboratory of Experimental Archaeology (Universidad Autónoma de Madrid). This work has been carried out with the financial support of the Generalitat de Catalunya, AGAUR agency (2017SGR1040 Research Group), the Universitat Rovira i Virgili (2021PFR-URV-126), and the Spanish Ministry of Science and Innovation (MICINN/FEDER project PID2021-122355NB-C32).The Institut Català de Paleoecologia Humana i Evolució Social (IPHES-CERCA) has received financial support from the Spanish Ministry of Science and Innovation through the “María de Maeztu” program for Units of Excellence (CEX2019-000945-M).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jasrep.2022.103698.

References

Alin, A., 2010. Multicollinearity. Wiley Interdiscip. Rev. Comput. Stat. 2, 370–374. https://doi.org/10.1002/wics.84.
Andrefsky, W., 2005. Lithics macroscopic approaches to analysis, Second ed. Cambridge Manuals in Archaeology. Cambridge University Press, Cambridge.
Andrefsky, W., 2009. The analysis of stone tool procurement, production, and maintenance. J. Archaeol. Res. 17, 65–103. https://doi.org/10.1007/s10814-008-9026-2.
Archer, W., Pop, C.M., Rezek, Z., Schlager, S., Lin, S.C., Weiss, M., Dogandžić, T., Desta, D., McPherron, S.P., 2018. A geometric morphometric relationship predicts stone flake shape and size variability. Archaeol. Anthropol. Sci. 10, 1991–2003. https://doi.org/10.1007/s12520-017-0517-2.
Binford, L.R., 1973. Interassemblage variability - the Mousterian and the ’functional’ argument. In: Renfrew, C. (Ed.), The Explanation Of Culture Change. Models in Prehistory, Duckworth, Gloucester, pp. 227–254.
Binford, L.R., 1979. Organization and formation processes: Looking at curated technologies. J. Anthropol. Res. 35, 255–273.
Boëda, E., 1993. Le débitage discoïde et le débitage Levallois récurrent centripède. Bulletin de la Société Préhistorique Française 90, 392–404. https://doi.org/10.3406/bspf.1993.9669.
Boëda, E., 1995a. Caractéristiques techniques des chaînes opératoires lithiques des niveaux micoquiens de Külna (Tchécoslovaquie). Paléo. Supplément 1, 57–72. https://doi.org/10.3406/pal.1995.1380.
Boëda, E., 1995b. Levallois: A Volumetric Construction, Methods, A Technique. In: Dibble, H.L., Bar-Yosef, O. (Eds.), The Definition and interpretation of Levallois Technology, Monographs in World Archaeology. Prehistory Press, Madison, Wisconsin, pp. 41–68.
Braun, D.R., Rogers, M.J., Harris, J.W.K., Walker, S.J., 2008. Landscape-scale variation in hominin tool use: Evidence from the Developed Oldowan. J. Hum. Evol. 55, 1053–1063. https://doi.org/10.1016/j.jhevol.2008.05.020.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.
Bustos-Pérez, G., Baena, J., 2019. Exploring volume lost in retouched artifacts using height of retouch and length of retouched edge. J. Archaeolog. Sci.: Rep. 27, 101922 https://doi.org/10.1016/j.jasrep.2019.101922.
Bustos-Pérez, G., Baena, J., 2021. Predicting flake mass: a view from machine learning. Lithic Technology 46, 130–142. https://doi.org/10.1080/01977261.2021.1881267.
Casamiquela, R.M., 1978. Temas patagónicos de interes arqueológico. La talla del vidrio. Relaciones de la Sociedad Argentina de Antropología 12, 213–223.
Casanova i Martí, J., Martínez Moreno, J., Mora Torcal, R., de la Torre, I., 2009. Stratégies techniques dans le Paléolithique Moyen du sud-est des Pyrénées. L’Anthropologie 113, 313–340. 10.1016/j.anthro.2009.04.004.
Clarkson, C., Hiscock, P., 2011. Estimating original flake mass from 3D scans of platform area. J. Archaeol. Sci. 38, 1062–1068. https://doi.org/10.1016/j.jas.2010.12.001.

Davis, Z.J., Shea, J.J., 1998. Quantifying lithic curation: an experimental test of dibble and Pelcin's original flake-tool mass predictor. J. Archaeol. Sci. 25, 603–610. https://doi.org/10.1006/jasc.1997.0255.
Dibble, H.L., 1987. The interpretation of Middle Paleolithic scraper morphology. Am. Antiq. 52, 109–117.
Dibble, H.L., 1995. Middle paleolithic scraper reduction: background, clarification, and review of the evidence to date. J. Archaeol. Method Theory 2, 300–368.
Dibble, H.L., 1998. Comment on "Quantifying Lithic Curation: An Experimental Test of Dibble and Pelcin's Original Flake-Tool Mass Predictor", by Zachary J. Davis and John J. Shea. J. Archaeol. Sci. 25, 611–613. https://doi.org/10.1006/jasc.1997.0254.
Dibble, H.L., Bernard, M.C., 1980. A comparative study of basic edge angle measurement techniques. Am. Antiq. 45, 857–865.
Dibble, H.L., Pelcin, A., 1995. The effect of hammer mass and velocity on flake mass. J. Archaeol. Sci. 22, 429–439. https://doi.org/10.1006/jasc.1995.0042.
Dogandžić, T., Braun, D.R., McPherron, S.P., 2015. Edge length and surface area of a blank: experimental assessment of measures. Size predictions and utility. PLoS ONE 10, e0133984.
Eren, M.I., Domínguez-Rodrigo, M., Kuhn, S.L., Adler, D.S., Le, I., Bar-Yosef, O., 2005. Defining and measuring reduction in unifacial stone tools. J. Archaeol. Sci. 32, 1190–1201. https://doi.org/10.1016/j.jas.2005.03.003.
Eren, M.I., Lycett, S.J., 2012. Why levallois? A morphometric comparison of experimental 'preferential' levallois flakes versus debitage flakes. PLoS ONE 7, e29273.
Fox, J., Weisberg, S., 2018. An R companion to applied regression, Third. ed. Sage publications, Thousand Oaks.
Furnival, G.M., Wilson, R.W., 1974. Regressions by Leaps and Bounds. Technometrics 16, 499–511.
Gould, R.A., 1968. Living archaeology: the Ngatatjaraof Western Australia. Southwestern J. Anthropol. 24, 101–122.
Günther, F., Fritsch, S., 2010. neuralnet: training of neural networks. The R J. 2.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Data Mining, Inference, and Prediction, Second Edition. ed, Springer Series in Statistics. Springer.
Hiscock, P., Clarkson, C., 2005. Experimental evaluation of Kuhn's geometric index of reduction and the flat-flake problem. J. Archaeol. Sci. 32, 1015–1022. https://doi.org/10.1016/j.jas.2005.02.002.
Hiscock, P., Tabrett, A., 2010. Generalization, inference and the quantification of lithic reduction. World Archaeol. 42, 545–561. https://doi.org/10.1080/00438243.2010.517669.
Hocking, R.R., Leslie, R.N., 1967. Selection of the best subset in regression analysis. Technometrics 9, 531–540.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning, Springer Texts in Statistics. Springer New York, New York, NY. 10.1007/978-1-4614-7138-7.
Kuhn, S.L., 1990. A Geometric index of reduction for unifacial stone tools. J. Archaeol. Sci. 17, 583–593.
Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28.
Lantz, B., 2015. Machine Learning with R, Seond Edition. ed. Packt Publishing Ltd., Birmingham.
Lumley based on Fortran code by Alan Miller, T., 2020. leaps: Regression Subset Selection.
Maloney, T.R., 2020. Experimental and archaeological testing with 3D laser scanning reveals the limits of I/TMC as a reduction index for global scraper and point studies. J. Archaeolog. Sci.: Rep. 29, 102068 https://doi.org/10.1016/j.jasrep.2019.102068.
Marquardt, D.W., 1970. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics 12, 591–612.
Marwick, B., 2017. Computational reproducibility in archaeological research: basic principles and a case study of their implementation. J. Archaeol. Method Theory 24, 424–450. https://doi.org/10.1007/s10816-015-9272-9.
Morales, J.I., Lorenzo, C., Vergès, J.M., 2015. Measuring retouch intensity in lithic tools: a new proposal using 3D scan data. J. Archaeol. Method Theory 22, 543–558. https://doi.org/10.1007/s10816-013-9189-0.
Muller, A., Clarkson, C., 2016. A new method for accurately and precisely measuring flake platform area. J. Archaeolog. Sci.: Rep. 8, 178–186. https://doi.org/10.1016/j.jasrep.2016.06.015.
Nelson, M.C., 1991. The study of technological organization. Archaeol. Method Theory 57–100.
Nuevo Delaunay, A., Belardi, J.B., Carballo Marina, F., Saletta, M.J., De Angelis, H., 2017. Glass and stoneware knapped tools among hunter-gatherers in southern Patagonia and Tierra del Fuego. Antiquity 91, 1330–1343. 10.15184/aqy.2017.125.
O'brien, R.M., 2007. A Caution Regarding Rules of Thumb for Variance Inflation Factors. Qual Quant 41, 673–690. 10.1007/s11135-006-9018-6.
Orellana Figueroa, J.D., Reeves, J.S., McPherron, S.P., Tennie, C., 2021. A proof of concept for machine learning-based virtual knapping using neural networks. Sci. Rep. 11, 19966. https://doi.org/10.1038/s41598-021-98755-6.
Paul, R.K., 2006. Multicollinearity: Causes, effects and remedies. IASRI, New Delhi 1, 58–65.
R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Rolland, N., Dibble, H.L., 1990. A new synthesis of Middle Paleolithic variability. Am. Antiq. 55, 480–499.
RStudio Team, 2019. RStudio: Integrated Development for R. RStudio, Inc., Boston, MA.
Sarkar, D., 2008. Lattice: Multivariate Data Visualization with R. Springer, New York.
Scerri, E.M.L., Gravina, B., Blinkhorn, J., Delagnes, A., 2016. Can lithic attribute analyses identify discrete reduction trajectories? A quantitative study using refitted lithic sets. J. Archaeol. Method Theory 23, 669–691. https://doi.org/10.1007/s10816-015-9255-x.
Shott, M.J., 1989. On tool-class use lives and the formation of archaeological assemblages. Am. Antiq. 54, 9–30. https://doi.org/10.2307/281329.
Shott, M.J., 1996. An exegesis of the curation concept. J. Anthropol. Res. 52, 259–280.
Shott, M.J., 2007. The role of reduction analysis in lithic studies. Lithic Technology 32, 131–141.
Shott, M.J., Bradbury, A.P., Carr, P.J., Odell, G.H., 2000. Flake size from platform attributes: predictive and empirical approaches. J. Archaeol. Sci. 27, 877–894. https://doi.org/10.1006/jasc.1999.0499.
Shott, M.J., Seeman, M.F., 2017. Use and multifactorial reconciliation of uniface reduction measures: a pilot study at the nobles pond paleoindian site. Am. Antiq. 82, 723–741. https://doi.org/10.1017/aaq.2017.40.
Shott, M.J., Weedman, K.J., 2007. Measuring reduction in stone tools: an ethnoarchaeological study of Gamo hidescrapers from Ethiopia. J. Archaeol. Sci. 34, 1016–1035. https://doi.org/10.1016/j.jas.2006.09.009.
Shott, M.J., 2005. The Reduction Thesis and its Discontents: Overview of the Volume, in: Clarkson, C., Lamb, L. (Eds.), Lithics "Down Under": Australian Perspectives on Lithic Reduction, Use and Classification, BAR International Series. British Archaeological Reports, pp. 109–125.
Spry, C., Stern, N., 2016. Technological Organization. In: Jackson, J.L. (Ed.), Oxford Bibliographies in "Anthropology". Oxford University Press, New York.
Terradas, X., 2003. Discoid flaking method: conception and technological variability. In: Peresani, M. (Ed.), Discoid Lithic Technology. Advances and Implications, BAR International Series. Archaeopress, Oxford, pp. 19–32.
White, J.P., 1967. Ethno-archaeology in New Guinea: Two Examples. Mankind 6, 409–414.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the Tidyverse. Journal of Open Source Software 4.
Wright, M.N., Ziegler, A., 2017. ranger: a fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. 77.