Age Estimation in Forensic Anthropology: methodological considerations about the validation studies of prediction models

Andrea Valsecchi, Javier Irurita Olivares, Pablo Mesejo

International Journal of Legal Medicine manuscript No.
(will be inserted by the editor)

Age Estimation in Forensic Anthropology: methodological considerations about the validation studies of prediction models

Andrea Valsecchi · Javier Irurita Olivares · Pablo Mesejo

the date of receipt and acceptance should be inserted later

Abstract There is currently no clear consensus on how to calculate, express and interpret the error when validating methods for age estimation in forensic anthropology. For this reason, it is likely that researchers are commonly drawing erroneous or confusing conclusions about the existence of population differences or the need to design new and increasingly complex estimation methods. In recent years, many researchers have highlighted these limitations. They propose new lines of research focused on the use of rigorous statistics and new technologies for the development of methods for estimating age. Our main objective in this study is to contribute to the strengthening of these novel ideas, for which we show the existing empirical evidence about the inadequacy of some age estimation methods in calculating, expressing and interpreting the errors obtained. With this aim, a total of 500 simulations have been performed, in which hypothetical research teams develop and validate methods for age estimation. The data employed in this study was obtained from the Centers for Disease Control and Prevention (CDC) Growth Charts: United States released in 2000. The charts relate age with height, weight and head circumference of US male children. Five learning algorithms have been employed as age estimators. We have performed three experiments in which the following aspects have been analyzed: frequency with which "negative" results can be obtained in the validation studies; which are the most appropriate criteria to compare and select the age estimation methods; and what analysis should be employed to carry out the validation studies. The results show possible errors in the interpretation of validation studies as a consequence of the confusion of statistical

A. Valsecchi
Andalusian Research Institute in Data Science and Computational Intelligence, Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain. E-mail: andreavalsecchi@ugr.es

J. Irurita Olivares
Laboratory of Anthropology, Department of Legal Medicine, Toxicology and Physical Anthropology, University of Granada, Granada 18012, Spain. E-mail: javierirurita@gmail.com

P. Mesejo
Andalusian Research Institute in Data Science and Computational Intelligence, Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain. E-mail: pmesejo@decsai.ugr.es

concepts. To conclude, we made a proposal of "good practices" for the correct calculation, expression and interpretation of the error when validating age estimation methods in forensic anthropology.

Keywords age estimation · regression problems · methodological review · validation studies

1 Introduction

Among the admissibility criteria for expert testimony there is the reliable assessment of the error rate associated with the forensic methods being used. This is one of the main requirements of the Daubert standard, which was proposed in 1993 in the United States as a guide to assist judicial bodies in assessing the admissibility of testimonies provided by scientists [1]. Currently, the Daubert standard has been adopted by many countries as part of the requirements for the admissibility of the expert evidence [2] and its approaches are widely accepted by the scientific community [3,4]. However, some authors suggest that the Daubert standard has not had a real impact in Forensic Anthropology (FA) contexts, since it has been limited to a statement of "good intentions" with little practical impact on the preparation of expert reports [5]. One possible explanation may be that, although current methods include indicators for describing the assumed error, this term is often poorly defined or misinterpreted [1].

Another aspect to keep in mind is that FA is one of the disciplines with a greater number of methods for its application in human identification [1]. This is because it must be adapted to many specific circumstances depending, for example, on the age group, the preservation state or possible taphonomic alterations. As a consequence, there is often no clear consensus on what methods should be applied in each circumstance [6], since little emphasis has been placed on defining the criteria to be used to select the most appropriate method. In relation to this, Buckberry states that forensic anthropologists tend to use only the methods they learned during their academic training, or those designed by themselves in the case of researchers, rather than using more robust scientific-based criteria [7]. This shows a lack of continuous training that limits the updating of concepts and the use of a more advanced methodology.

Validation studies are the main tool to provide the necessary criteria to decide which method is most appropriate, since they evaluate its effectiveness and applicability. When the FA methodology is tested in different samples, in many cases significant differences are observed between the actual and the estimated age of the individuals. In these cases, researchers often modify or adapt the original methods to obtain better specific results for their study populations [8,9]. As several authors propose [10,11,12], these errors may be misinterpreted as population differences or deficiencies of the original method. Probably, these differences can be explained (at least in part) as the normal variability of the process or as differences in structure and distribution of the study sample, so adapting or redesigning new methods could be unjustified. In other words, errors in validation studies do not imply population differences, especially when the original method does not describe sufficiently well the variability of the population under study. If one compares the training error of a method with the validation error of the same

method in a different population, the results will not allow to conclude that there are population differences.

On the other hand, technological advances allow us to design increasingly complex and robust methods. This allows researchers to create more tight-fitting models for their study samples, with the aim of reducing the error and achieving more accurate results. In contrast to this, some authors suggest that the increase in the complexity of the method may lead to an increase in error in the validation studies, and the increase in precision usually means a decrease in accuracy and vice versa [7, 10, 13]. Therefore, many of these new methods may not represent an improvement in the current methodology used in FA, despite the promising results obtained in their original study samples.

For one reason or another, there seems to be a trend in FA research marked by euphoria for designing new methods, increasingly complex and specific to each population, motivated by the unfavorable results obtained in the validation studies [1, 13]. We argue that this tendency is the direct result of incorrect procedures being used in computing, reporting and comparing the prediction error. Namely, we point to two widespread issues:

Lack of testing over unseen data Whenever an age estimation method is derived using a set of data, the prediction ability of the method should be tested on different, unseen data. It is a well-known and extremely common issue that any estimation method will perform much better when tested on the very same data used for its creation (the so-called *training* data). This tendency is called *overfitting*; it occurs because the derivation of the estimation method picks up some subtle pattern which is only present in the training data, rather than being part of the real phenomenon under investigation. Due to the possibility of overfitting, the error obtained over the training data (i.e. the training error) is useless in assessing the prediction ability of the estimation method. To measure the latter, the researcher should reserve part of the available data and create a *testing* dataset (over which a testing error will be calculated).
Unfortunately, a large number of studies do not include a separated testing procedure and simply report the training error. The new method is thus believed to be much more accurate than it actually is [14], and therefore it is likely to spring replications and validation studies from other researchers. Those studies are very likely to fail, as the error they are comparing with is unrealistically low.

Flawed comparison procedure When comparing the performance of an age estimation method on different datasets, an appropriate statistical test should be used. Indeed, the aim of the comparison is to assess the performance of the method over the two populations from which the datasets have been drawn, thus it is an example of statistical inference. Instead, many studies simply compare the two mean error values and assume that the same result would apply to the two populations.

Validation studies are essential to justify the admissibility of the methods as expert evidence, however, what measure should we take when the results of these studies are unfavorable? The result of the validation studies should not be a binary (correct or incorrect) answer, but should instead provide information on the precision and accuracy of the method [1].

Unfortunately, the calculation and interpretation of the error are not always properly addressed in validation studies; Konigsberg [15] argues that this may be due because anthropologists often work away from statisticians. In order to arrive at a consensus on the criteria for selecting the most appropriate method, it is imperative that we first approach a consensus around validation studies, from an interdisciplinary perspective, concerning both the calculation and the interpretation of the assumed error.

In this paper, in addition to the theoretical explanation of why the procedures commonly used in age estimation studies are often flawed, we provide an experimental proof-of-concept of those issues. This is performed by running an extensive computer simulation (using actual age estimation data) and quantifying how often such practices can lead to wrong conclusions.

2 Background

2.1 Regression Problems

In most cases, designing a method for forensic age estimation is an example of a more general task called *regression*. In regression, the aim is to learn the relationship between a certain continuous *outcome* variable (e.g. age) and other *features* (e.g. height, weight, etc.) from a set of examples (*training data*), which records the expected outcome and feature measurements for a set of objects (e.g. people). Using these data, we build a prediction *model* (e.g. a formula or a table), which should be able to accurately predict the outcome of new, unseen objects. To create a model from the data, we employ a *learning algorithm*, which describes how to tailor the model for the training data (or, in other words, to learn from the training data). For instance, the algorithm may specify how to calculate the model's parameters.[1]

Regression lies in the overlap between different fields of science, in particular statistics and machine learning, and the terminology varies accordingly (Table 1). In this paper we use the more modern terminology of machine learning. In what follows, Y denotes the outcome variable, while $X_i = (X_{1,i}, \ldots, X_{m,i})$ is the i-th vector of m input features. Our training sample (training set or, simply, dataset) S contains n examples of the form

$$(Y_i, X_{1,i}, \ \ldots, \ X_{m,i}) \quad \text{for } i = 1, \ldots, n$$

A model is a function f that provides a *predicted outcome* $\hat{Y}$ given the features X, i.e. $\hat{Y} = f(X)$. The difference between predicted and actual outcome is called error or *residual* ϵ

$$\epsilon_i = Y_i - \hat{Y}_i = Y_i - f(X_i)$$

[1] As a model is the result of applying a learning algorithm, the two terms sometimes used interchangeably. However, whenever more than one training sample is considered, it is important to remark that the same algorithm would produce different models, depending on which dataset is actually used for training.

Table 1 Alternative terminology used in regression depending on the field of study.

outcome	response, dependent variable, output
features	predictors, independent variables, inputs
learning algorithm	learner, regression method
pattern, example, instance	observation, data point

To measure the overall prediction error, the *mean squared error* (MSE) is often used[2]

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\epsilon_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2$$

The error term can be due to biological variability, noise or measuring errors in the data, but it could also mean that the model is not suited for the task. The last element is the learning algorithm $\mathcal{L}$, which is what produces the model M from the training data S, i.e. $\mathcal{L}(S) = M$. It is important to notice that this M corresponds to the aforementioned function f.

Consider the example of *linear regression*, which is a simple but powerful and popular method. With this learning algorithm, the outcome is modeled as a linear combination of the features, i.e.

$$\hat{Y} = b_0 + b_1 X_1 + \ldots + b_m X_m$$

The coefficients $b_0, \ldots, b_m$ are the parameters of the model, which are calculated in order to minimize the MSE over the training sample. The computation of those parameters is the very core of the learning algorithm, while the resulting formula, say $\hat{Y} = 3.4 + 7 \cdot X_1 - 2.1 \cdot X_2$ is an actual model.

2.2 Overfitting and model validation

Just like the coefficients of the linear regression model are chosen to result in the smallest possible MSE value, every regression model is fitted to the specific set of training data. However, the model should be able to make reliable predictions on new, unseen data, not just the training data. The term *overfitting* describes the situation where a model is excessively tailored for the training data, so it even models the random errors and the noise occurring in the dataset, instead of the underlying relationship between outcome and features. Figure 1 shows an example of such scenario.

Overfitting generally occurs when a model is excessively complex, for instance because it has too many parameters with respect to the size of dataset. A common example is polynomial regression using a polynomial of higher degree than the

[2] Other popular measures of the overall prediction error is the *mean absolute error* (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\epsilon_i| = \frac{1}{n}\sum_{i=1}^{n}|Y_i - f(X_i)|$$

Often, median, maximum and minimum errors, either squared or absolute, are also reported. For simplicity, here we only use the MSE.

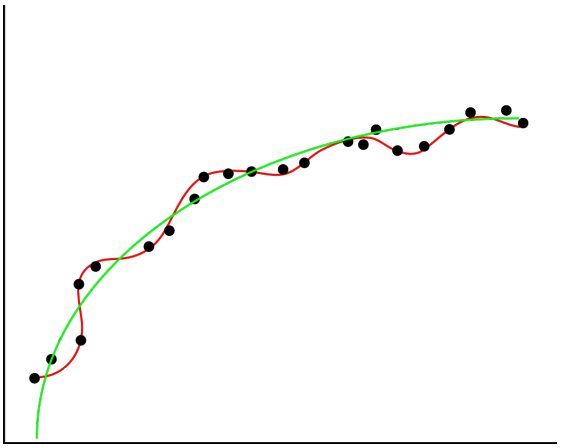

Fig. 1 Two interpolation lines. While the red line best follows the training data, it is too dependent on the specific points in the training set, thus it is likely to perform poorly on unseen data. The green line, instead, seem to capture the underlying pattern. The underfitting case would be represented by a straight line that due to an excessive bias does not fit data sufficiently well.

number of available data points in the training set. While essentially unavoidable, a lot of research has been devoted to reduce overfitting [16,17,18,19].

Because of overfitting, the model's error over the training data is not at all a reliable measure of the quality of the model for prediction. This is a key point: to test the prediction capability, additional, unseen data should be used, creating what is called *testing dataset* (or simply test set). It is a very common mistake in age estimation to simply report the error obtained on training data and assume that the model would be as accurate on new data. In fact, the difference can be extremely large.

When new data can not be easily gathered, one solution is to split the available data between training and testing datasets. Researchers are faced with a trade-off in choosing what percentage of data to use for training. More training data is likely to result in a better model, while more testing data makes the assessment more reliable. The technique called *cross-validation* (CV) provides an excellent solution to this issue [20]. Consider any learning algorithm $\mathcal{L}$. In cross-validation, the data S is randomly split into k disjoint subsets $S_1, \ldots, S_k$ of the same size, called *folds*. The process is performed in k rounds, one for each fold. In the first round, we build a model M_1 using the data $S \setminus S_1 = S_2 \cup S_3 \cup \ldots \cup S_k$ for training. Then, M_1 is tested over S_1, which has not been used for training, resulting in a certain testing error value E_1. In the second round, the model M_2 is trained on $S \setminus S_2$ and tested on S_2, resulting in an error measure E_2. In general, for $i = 1, \ldots, k$, we have $M_i = \mathcal{L}(S \setminus S_i)$ and $E_i = \mathrm{MSE}(M_i(S_i))$. Once the k round is over, the errors of each round are averaged, resulting in an average error

$$\overline{E}_{\mathrm{cv}} = \frac{1}{k} \sum_{i=1}^{k} E_i$$

Cross-validation stands out as a practical and thorough way of assessing the performance of a learning algorithm. As the whole learning process is repeated

multiple times, the variability of the performance can be measured and taken into account in comparing different learning algorithms.

The number of folds determines the number of models being trained and tested. More folds mean more thorough testing but also larger running times. Setting k to 10 is a common choice. Cross-validation can also be repeated multiple times, changing the random split of the data into folds. The most comprehensive form of CV, called *leave-one-out*, is performed by having each fold containing a single example (datum), so $k = |S|$. This can be prohibitively time consuming for large datasets and, in fact, this validation strategy is mainly applied when training data are scarce.

Note that, while this matter is traditionally called *model validation*, the subject of the procedure is not a model. Rather, what we are actually measuring is the ability of the learning algorithm to produce good predictive models regardless of the actual training sample being used.

2.3 Model selection

The term *model selection* refers to the task of choosing the best learning algorithm among those available for a regression task. As in the previous section, despite the traditional terminology, the subject is not really a model but a learning algorithm. Suppose that we want to compare two learning algorithms $\mathcal{L}_A$ and $\mathcal{L}_B$ over certain regression data S. Using k-fold CV on each algorithm, we obtain the prediction errors corresponding to the k rounds of the CV procedure, i.e. $E^A_{\text{cv}} = E^A_1, \ldots, E^A_k$ and $E^B_{\text{cv}} = E^B_1, \ldots, E^B_k$.

One might think that he can just compute the average of the two sets of errors E^A_{cv} and E^B_{cv} and be done with it. However, that would tell which algorithm has the lowest average error with respect to the folds $S_1, \ldots, S_k$. Instead, what we want to know is which algorithm has the lowest error when tested on *any* set of data from S, and not just the actual sets used in the CV process. In this sense, the errors $E^A_1, \ldots, E^A_k$ are a *sample* of the *population* of prediction errors associated with the algorithm A. As a consequence, the comparison between learning algorithms is actually an instance of *statistical inference*, i.e. when somebody wants to conclude something about populations based on samples. It is indeed analogous to comparing the mean height of two populations based on two samples. Following the recommendation of [21] and many others, we suggest the use of non-parametric statistical testing procedure such as the Wilcoxon's test [22].

3 Materials and methods

We used data from the 2000 CDC Growth Charts [23]. These growth charts consist of a series of percentile curves that illustrate the distribution of selected body measurements in male U.S. children. The growth charts were developed by the National Center for Health Statistics (NCHS) as a clinical tool for health professionals to determine if the growth of a child is adequate. The data come in the form of five charts relating the age of a child with either his head circumference, weight or length. The 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the

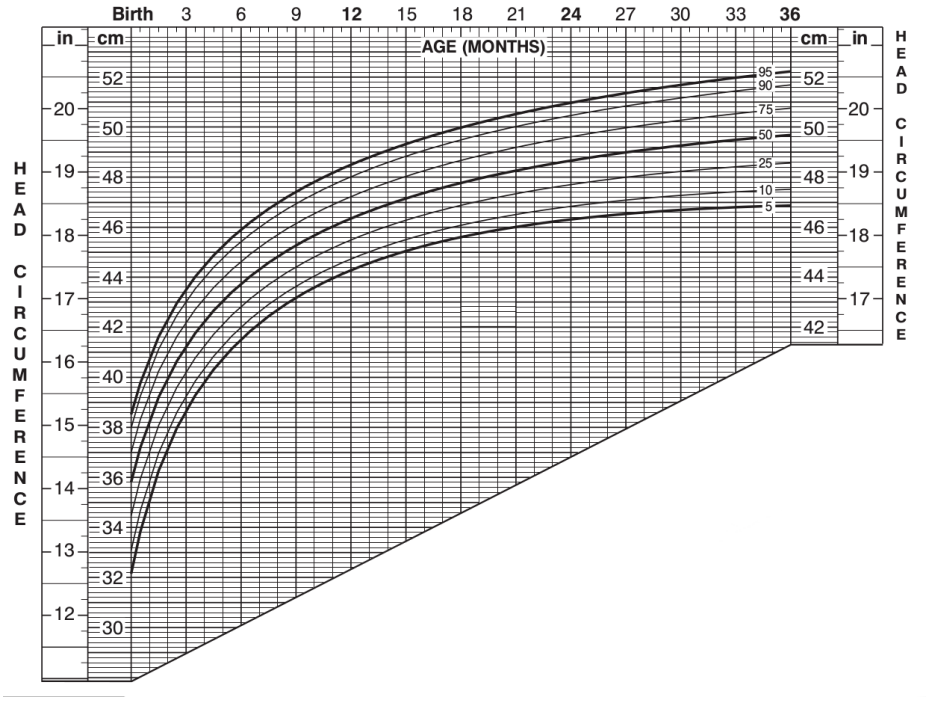

Fig. 2 The grow chart corresponding to case study I, detailing the distribution of head circumference of child ages 0–36 months. Taken from [23].

corresponding statistical distribution are reported. This provides detailed information about each distribution, allowing us to accurately simulate drawing samples of individuals from the population covered by the study. All the data samples used in this paper are obtained by sampling from the statistical distributions described in the charts.

The detail of the five case studies are provided in Table 2. Note that on each case study, the age is estimated from a single feature, either head circumference, weight or length. The grow chart of case study I is shown in Figure 2.

The experiments are performed considering 5 widely used regression algorithms: linear regression (LR), classical calibration (CC)[24], a 6th degree polynomial regression (Poly6), locally weighted scatterplot smoothing (LOESS) [25], and random forest (RF) [26]. The LOESS algorithm used a span value of 0.75 and a degree of 2, the default values.

Considering multiple algorithms ensures that our findings do not depend on a particular feature of a specific regression technique. Besides, we wanted to encompass a variety of methods: linear models (LR, CC), polynomial models (Poly6), non-parametric methods (LOESS) and even computationally intensive approaches like ensemble-based methods (RF). It is very important to remark that our aim is not to find the best possible regression method for age estimation; instead, we want to establish guidelines about how to calculate, express and interpret the error when validating age estimation methods in forensic anthropology.

Table 2 Detailed information over the case studies.

case study	feature	age range
I	head circumference	0–36 months
II	length	0–36 months
III	weight	0–36 months
IV	length	24–240 months
V	weight	24–240 months

In what follows, three experiments are described. Each experiment simulates the actions of a research team developing a method for age estimation. This involves collecting a sample of data, training different models and testing them. By performing the simulation hundreds of times, we can show how much the results can vary due to the randomness of the sampling.

3.1 Experiment one: testing the fallibility of validation studies

In the first experiment, we simulate the validation of an existing age estimation method using a different data collection. Say that research team A has collected a data sample S_A and, by applying some learning algorithm, obtained an age estimation method (i.e. a model) M_A. Suppose that the model is assessed using the training error $\mathrm{MSE}(M_A, S_A)$, which is an incorrect but unfortunately common practice. Team B wants to validate the method M_A on their own sample S_B, which may or may not have been collected from the same population. Therefore, the latter team tests M_A over S_B and computes the error value $\mathrm{MSE}(M_A, S_B)$. If the two error values are substantially different, the validation is considered to be failed.

A common misconception is that, if the validation fails, this must be due to differences between the populations from which the samples were drawn. The underlying assumption is that if data from the same population had been used, the two error values would have been very similar. In this experiment we test this idea. For each case study, we simulate the training of a model M_A and its validation on a different sample from the same population. The simulation has been repeated 500 times, using $500 \cdot 2 = 1000$ different samples of 100 elements. Then, we measured the percentage of the simulations in which the performance of M_A deteriorated when applied to S_B, considering a 20% or more increase in MSE value as a significant deterioration.

3.2 Experiment two: testing the inadequacy of using the training error as validation error

The second experiment aims to show how the predictive ability of a learning algorithm cannot be assessed using the training error. Instead, the algorithm must be tested on new, unseen data, used for testing purposes only.

For each algorithm we simulated the complete training and testing process 500 times over completely independent data samples, obtaining 500 training MSE values and the corresponding 500 testing MSE values.

3.3 Experiment three: using cross-validation to obtain a more realistic estimation of the validation error

In this third experiment, we show that using cross-validation, the MSE values are much closer to the testing performance compared to the training error. As in the previous experiment, for each algorithm we simulated the complete training, testing and cross-validation process 500 times. Training and cross-validation used the same data samples, while testing was performed over completely different data.

4 Results and discussion

4.1 Experiment one

Table 3 reports the rate of failed validations. For all algorithms and case studies, the failure rate is higher than 19%. That means that even in the best scenario, almost one time out of five the second team could have wrongly concluded that there are population differences between their collection and the other team's. This percentage gets much higher depending on the algorithm, with RF scoring 100% and poly6 scoring above 37%.

This clearly shows that a difference between training and validation errors is not necessarily a sign of the samples belonging to different populations. Nevertheless, many papers reach this conclusion. As Jayaraman et al. [27] points out, differences among populations should not be rashly concluded due to those observed in the validation studies. These authors emphasize the importance of other factors such as the distribution of age and sex of the sample, differences due to social status, ethnic (non-regional) differences, different times of the samples, etc. However, researchers usually do not include as possible explanation for their results the variability of the process being studied. In many cases, the methods used in FA have been designed from samples unrepresentative of the actual underlying population. One of the reasons for this is that these methods have been designed for being applied in unusual circumstances. Therefore, the samples from which they come may be difficult to obtain. Some examples may be the analysis of the sternal end of the fourth rib [28], degenerative processes in the pubic symphysis [29], transparency of the dental root [30] or the great majority of methods that analyze the skeleton of children [14,31,32,33]. The use of unsuitable study samples does not necessarily imply bad methods, as these may be used in exceptional circumstances. However, it is especially important in these cases that validation studies are carried out properly. Separating a fraction of the sample to test the method is not a good option when reduced samples are available; for this reason, as observed in experiment three, CV will be the most appropriate option in these cases. Furthermore, as Konigsberg argues [15], when only small samples are available, it is more important to analyze them collectively to obtain larger samples than to look for population differences.

Table 3 Portion of experiments in which the validation resulted in a $\geq 20\%$ increment in MSE value compared to training.

case study	learning algorithm	failed validations
I	LOESS	33.0%
	LR	19.0%
	Poly6	45.0%
	RF	100.0%
	CC	27.0%
II	LOESS	33.0%
	LR	23.0%
	Poly6	37.0%
	RF	100.0%
	CC	26.0%
III	LOESS	37.0%
	LR	23.0%
	Poly6	43.0%
	RF	100.0%
	CC	26.0%
IV	LOESS	32.0%
	LR	27.0%
	Poly6	44.0%
	RF	100.0%
	CC	28.0%
V	LOESS	34.0%
	LR	21.0%
	Poly6	56.0%
	RF	100.0%
	CC	27.0%

4.2 Experiment two

Table 4 reports mean and standard deviation of the results obtained in experiment two. Consider the mean error, for LR the difference is very small, with the testing error being just slightly larger than the training error. In the case of RF, the testing error can be as large as four times the training one. This is a clear sign of overfitting, which is not surprising due to the large complexity of the models produced by RF, especially with data having a single feature. LOESS training error values are just moderately smaller than the testing ones. Poly6 and CC are also showing overfitting, while CC is generally the worst performing algorithm. When considering the standard deviation of the MSE, training values are consistently smaller than testing ones. This means that the variability experienced during a validation is larger than that of the training process, which can explain why some many validation studies are (incorrectly) considered negative. Note that all those differences have been successfully tested for statistical significance at 99% confidence level, using Wilcoxon signed rank test and adjusting for multiple comparisons using Holm's method [34].

These results support our thesis: a comparison based on the training error would have incorrectly chosen RF as the best algorithm in all case studies. Nevertheless, RF ranks at the bottom (third, fourth or fifth place) when considering the testing MSE. On the other hand, LOESS performs consistently better that all other algorithms in testing (thus it should be considered the 'winner' of this

Table 4 Difference between average training and testing errors.

case study	learning algorithm	MSE mean		MSE sd		p-value
		training	test	training	test	
I	LOESS	31.1	33.7	5.2	6.1	4.5E-10
	LR	39.4	40.7	5.2	5.4	1.8E-03
	Poly6	31.4	142.9	5.2	845.1	1.7E-27
	RF	10.6	43.0	2.0	7.9	3.2E-82
	CC	59.6	63.2	11.2	17.3	1.2E-03
II	LOESS	8.7	9.4	1.6	1.7	1.9E-08
	LR	12.2	12.7	1.7	1.8	6.9E-04
	Poly6	8.5	9.9	1.6	4.2	5.3E-15
	RF	3.0	11.9	0.6	2.2	3.2E-82
	CC	13.6	14.4	2.0	2.6	6.5E-08
III	LOESS	20.9	23.2	3.6	4.3	3.6E-17
	LR	24.7	25.7	3.4	3.7	1.2E-04
	Poly6	20.9	27.6	3.7	40.4	4.2E-27
	RF	7.1	29.6	1.4	5.6	3.2E-82
	CC	31.3	33.0	5.5	6.2	1.5E-07
IV	LOESS	256.0	278.1	40.9	50.1	2.2E-10
	LR	284.2	291.7	47.4	52.0	3.3E-02
	Poly6	247.4	306.3	41.4	230.5	1.1E-27
	RF	88.7	345.7	17.0	63.8	3.2E-82
	CC	306.4	317.9	55.0	55.1	1.8E-03
V	LOESS	235.1	256.2	40.6	46.0	3.4E-12
	LR	379.9	397.2	59.5	63.6	1.1E-03
	Poly6	228.0	420.1	40.8	1899.5	3.4E-42
	RF	82.8	331.0	16.5	62.7	3.2E-82
	CC	431.4	446.4	79.2	81.7	5.3E-03

particular comparison) but it ranks in the middle when considering the training error.

4.3 Experiment three

The resulting MSE values are reported in Table 5. For all algorithms and case studies, the testing error is always closer to CV error than training error. In most cases, the difference between CV and testing error is very small, while that of training and testing is large. Figure 3 clearly shows this pattern. LR and CC are the least affected, while the effect is much stronger on Poly6 and RF. Overall, the experiment shows how performing CV is a sound and efficient way to assess the testing error without requiring additional data.

5 Conclusion

This paper aims to highlight the limitations that exist in many studies where the training set is used indistinctly to train and validate the proposed models. From this point of view, the regression methods used in this article, as well as the prediction error metric, are no more than mere tools to empirically show this phenomenon: how the validation of a regression method through the same data

Table 5 Difference between average training, cross-validation and testing errors.

case study	learning algorithm	mean MSE training	mean MSE cross-validation	mean MSE test
I	LOESS	31.1	34.0	33.7
	LR	39.4	40.8	40.7
	Poly6	31.4	158.7	142.9
	RF	10.6	43.2	43.0
	CC	59.6	64.3	63.2
II	LOESS	8.7	9.5	9.4
	LR	12.2	12.7	12.7
	Poly6	8.5	10.5	9.9
	RF	3.0	11.9	11.9
	CC	13.6	14.3	14.4
III	LOESS	20.9	22.6	23.2
	LR	24.7	25.3	25.7
	Poly6	20.9	24.9	27.6
	RF	7.1	28.9	29.6
	CC	31.3	32.9	33.0
IV	LOESS	256.0	278.5	278.1
	LR	284.2	294.0	291.7
	Poly6	247.4	295.0	306.3
	RF	88.7	349.1	345.7
	CC	306.4	320.1	317.9
V	LOESS	235.1	256.8	256.2
	LR	379.9	395.4	397.2
	Poly6	228.0	390.3	420.1
	RF	82.8	325.9	331.0
	CC	431.4	452.3	446.4

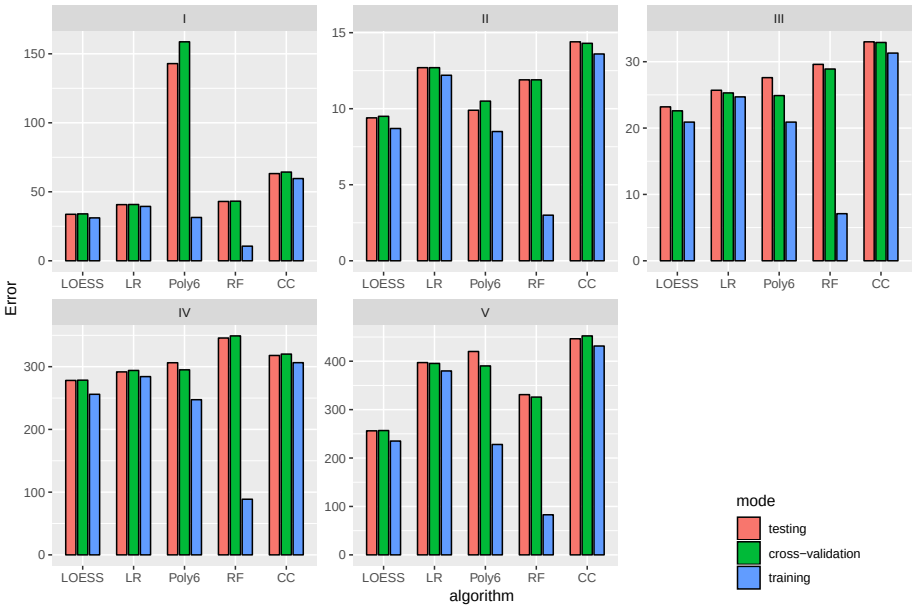

Fig. 3 Difference between average training, cross-validation and testing errors.

set with which it was trained, produces regressors whose performance is overestimated. In opposition, cross-validation represents a rigorous validation technique for assessing how the results of a statistical analysis will generalize to an unseen data set. In particular, we have, first, shown that a difference between training and validation errors is not necessarily a sign of the samples belonging to different populations. Second, we have tested the inadequacy of using the training error as validation error. And, finally, we have shown how cross-validation provides a realistic estimation of the validation error.

There is a clear consensus in the scientific community regarding the need to discuss aspects related to the methodological review in forensic anthropology [1, 6,7,15]. In order to contribute to this demand, we present as conclusions of this study a proposal of good practices for the validation of age estimation methods:

- An adequate calculation and interpretation of the error must be the priority when designing and validating methods for age estimation. It is necessary to use appropriate validation studies and to put emphasis on getting adequate samples.
- All methods must report the testing error and not the training error. This testing error must be employed when making estimates and comparing age estimation methods, and not the training error.
- Differences between groups cannot be assessed by simply comparing mean error values. The use of statistical tests is recommended to evaluate the existence of statistically significant differences. In particular, the use of nonparametric tests (such as the Wilcoxon signed-rank test) is recommended, to avoid a decrease in the reliability of the conclusions obtained when using parametric statistical tests whose assumptions (homoscedasticity, normality) are not met.
- Complex or very specific methods, apparently good because they present a reduced training error, can yield poor results in validation studies. For this reason, we must always use the testing error as an indication of the reliability of a method.
- When only small samples are available, cross-validation has traditionally proven to be an effective protocol of experimental validation. In case the size of the dataset to train the regression model is extremely reduced, the use of the leave-one-out validation protocol is a recommendable alternative.

Acknowledgment

This work has been supported by the Spanish Ministerio de Economía y Competitividad (MINECO) under the NEWSOCO project (ref. TIN2015-67661-P) and the Andalusian Dept. of Innovación, Ciencia y Empresa under the project TIC2011-7745, including European Regional Development Funds (ERDF-FEDER). Pablo Mesejo is funded by the European Commission H2020-MSCA-IF-2016 through the Skeleton-ID Marie Curie Individual Fellowship [grant number 746592].

References

1. Angi M. Christensen and Christian M. Crowder. Evidentiary standards for forensic anthropology*. Journal of Forensic Sciences, 54(6):1211–1216, 2009.

2. Stefano De Luca, Fernando Navarro, and Roberto Cameriere. The expert witness and its admissibility as evidence in the Spanish legal system. Revista Electrónica de Ciencia Penal y Criminología, (15-19):19:1–19:14, 2013. [Online; http://criminet.ugr.es/recpc/15/recpc15-19.pdf].
3. Daniel Franklin. Forensic age estimation in human skeletal remains: Current concepts and future directions. Legal Medicine, 12(1):1 – 7, 2010.
4. Nicholas Márquez-Grant. An overview of age estimation in forensic anthropology: perspectives and practical considerations. Annals of Human Biology, 42(4):308–322, 2015.
5. Kate M. Lesciotto. The impact of daubert on the admissibility of forensic anthropology expert testimony. Journal of Forensic Sciences, 60(3):549–555, 2015.
6. E. Cunha, E. Baccino, L. Martrille, F. Ramsthaler, J. Prieto, Y. Schuliar, N. Lynnerup, and C. Cattaneo. The problem of aging human remains and living individuals: A review. Forensic Science International, 193(1):1 – 13, 2009.
7. Jo Buckberry. The (mis)use of adult age estimates in osteology. Annals of Human Biology, 42(4):323–331, 2015.
8. Erin H. Kimmerle and Richard L. Jantz. Variation as evidence: Introduction to a symposium on international human identification. Journal of Forensic Sciences, 53(3):521–523, 2008.
9. Erin H. Kimmerle, Richard L. Jantz, Lyle W. Konigsberg, and Jose Pablo Baraybar. Skeletal estimation and identification in american and east european populations*. Journal of Forensic Sciences, 53(3):524–532, 2008.
10. Lyle W. Konigsberg, Nicholas P. Herrmann, Daniel J. Wescott, and Erin H. Kimmerle. Estimation and evidence in forensic anthropology: Age-at-death. Journal of Forensic Sciences, 53(3):541–557, 2008.
11. H.M. Liversidge. Interpreting group differences using demirjian's dental maturity method. Forensic Science International, 201(1):95 – 101, 2010.
12. Helen M. Liversidge. Controversies in age estimation from developing teeth. Annals of Human Biology, 42(4):397–406, 2015.
13. Douglas H. Ubelaker. Issues in the global applications of methodology in forensic anthropology*. Journal of Forensic Sciences, 53(3):606–607, 2008.
14. Louise Corron, François Marchal, Silvana Condemi, and Pascal Adalian. A critical review of sub-adult age estimation in biological anthropology: Do methods comply with published recommendations? Forensic science international, 2018.
15. Lyle W. Konigsberg. Multivariate cumulative probit for age estimation using ordinal categorical data. Annals of Human Biology, 42(4):368–378, 2015.
16. Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.
17. Lutz Prechelt. Early Stopping-But When? In Neural Networks: Tricks of the Trade, This Book is an Outgrowth of a 1996 NIPS Workshop, pages 55–69, 1998.
18. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J. Mach. Learn. Res., 15(1):1929–1958, 2014.
19. Leo Breiman. Bagging predictors. Machine learning, 24(2):123–140, 1996.
20. Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai, volume 14, pages 1137–1145, 1995.
21. Janez Demšar. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7:1–30, 2006.
22. Frank Wilcoxon. Individual comparisons by ranking methods. Biometrics bulletin, 1(6):80–83, 1945.
23. Robert J. Kuczmarski, Cynthia L. Ogden, Shumei S. Guo, Laurence M. Grummer-Strawn, Katherine M. Flegal, Zuguo Mei, Rong Wei, Lester R. Curtin, Alex F. Roche, and Clifford L. Johnson. 2000 CDC Growth Charts for the United States: methods and development. Vital and health statistics. Series 11, Data from the national health survey, (246):1–190, 2002.
24. R. G. Aykroyd, D. Lucy, A. M. Pollard, and T. Solheim. Technical note: Regression analysis in adult age estimation. American Journal of Physical Anthropology, 104(2):259–265, 1997.
25. William S. Cleveland and Susan J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. Journal of the American Statistical Association, 83(403):596–610, 1988.

26. Tin Kam Ho. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8):832–844, 1998.
27. Jayakumar Jayaraman, Hai Ming Wong, Nigel M King, and Graham J Roberts. The french–canadian data set of demirjian for dental age estimation: a systematic review and meta-analysis. Journal of forensic and legal medicine, 20(5):373–381, 2013.
28. M Yaşar Işcan, Susan R Loth, and Ronald K Wright. Metamorphosis at the sternal rib end: a new method to estimate age at death in white males. American journal of physical anthropology, 65(2):147–156, 1984.
29. Sheilagh Brooks and Judy M Suchey. Skeletal age determination based on the os pubis: a comparison of the acsádi-nemeskéri and suchey-brooks methods. Human evolution, 5(3):227–238, 1990.
30. H Lamendin, E Baccino, JF Humbert, JC Tavernier, RM Nossintchouk, and A Zerilli. A simple technique for age estimation in adult corpses: the two criteria dental method. Journal of Forensic Science, 37(5):1373–1379, 1992.
31. Hugo FV Cardoso, John M Vandergugten, and Louise T Humphrey. A ge estimation of immature human skeletal remains from the metaphyseal and epiphyseal widths of the long bones in the post-natal period. American journal of physical anthropology, 162(1):19–35, 2017.
32. Javier Irurita Olivares and Inmaculada Alemán Aguilera. Proposal of new regression formulae for the estimation of age in infant skeletal remains from the metric study of the pars basilaris. International journal of legal medicine, 131(3):781–788, 2017.
33. Javier Irurita Olivares, Inmaculada Alemán Aguilera, Joan Viciano Badal, Stefano De Luca, and Miguel Cecilio Botella López. Evaluation of the maximum length of deciduous teeth for estimation of the age of infants and young children: proposal of new regression formulas. International journal of legal medicine, 128(2):345–352, 2014.
34. Sture Holm. A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics, pages 65–70, 1979.