Journal of Statistical Theory and Practice (2023) 17:44
https://doi.org/10.1007/s42519-023-00341-8

ORIGINAL ARTICLE

Regularization Methods for High-Dimensional Data as a Tool for Seafood Traceability

Clara Yokochi[1] · **Regina Bispo**[1,2] · **Fernando Ricardo**[3] · **Ricardo Calado**[3]

Accepted: 13 August 2023 / Published online: 30 August 2023

Abstract

Seafood traceability, needed to regulate food safety, control fisheries, combat fraud, and prevent jeopardizing public health from harvesting in polluted locations, depends heavily on the prediction of the geographic origin of seafood. When the available datasets to study traceability are high-dimensional, standard classic statistical models fail. Under these circumstances, proper alternative methods are needed to predict accurately the geographic origin of seafood. In this study, we propose an analytical approach combining the use of regularization methods and resampling techniques to overcome the high-dimensionality problem. In particular, we analyze comparatively the *Ridge regression*, LASSO and *Elastic net* penalty-based approaches. These methods were applied to predict the origin of the saltwater clam *Ruditapes philippinarum*, a non-indigenous and commercially very relevant marine bivalve species that occurs commonly in European estuaries. Further, the resampling method of *Monte Carlo Cross-Validation* was implemented to overcome challenges related to the small sample size. The results of the three methods were compared. For fully reproducibility,

This article is part of the special issue "Ecological Statistics" guest edited by Tiago Marques, Charlotte Jones-Todd, Ben Stevenson, Theo Michelot, and Ben Swallow.

✉ Regina Bispo
r.bispo@fct.unl.pt

Clara Yokochi
c.sampaio@campus.fct.unl.pt

Fernando Ricardo
fafr@ua.pt

Ricardo Calado
rjcalado@ua.pt

1 NOVAMATH Center for Mathematics and Applications, NOVA School of Science and Technology, Universidade NOVA de Lisboa, Caparica, Portugal

2 Department of Mathematics, NOVA School of Science and Technology, Universidade NOVA de Lisboa, Caparica, Portugal

3 ECOMARE, CESAM - Centre for Environmental and Marine Studies, Department of Biology, University of Aveiro, Santiago University Campus, Aveiro, Portugal

an R Markdown file and the used dataset are provided. We conclude highlighting the insights that this methodology may bring to model a multi-categorical response based on high-dimensional dataset, with highly correlated explanatory variables, and combat the mislabeling of geographic origin of seafood.

Keywords Elastic net · LASSO · Regularization · Ridge regression · Traceability

1 Introduction

As a result of the growing globalization of seafood markets, traceability has become paramount to safeguard food safety. With consumers being increasingly aware of the potential hazards that contaminated seafood may cause to their health, ensuring that the geographic origin of seafood is not mislabeled is a first step to fight fraudulent practices that aim to cover-up illegal fishing [1–3]. Features associated with the traceability of seafood place of origin have been reviewed by Leal et al. [2]. Among these, biochemical and biogeochemical fingerprints have been pointed out as of interest when aiming to trace the geographic origin of seafood, including tens of features.

Data analytic techniques typically applied for detecting fish and seafood authenticity and adulteration were recently reviewed by Kotsanopoulos et al. [4]. These authors identified the approaches that have been adopted so far, which includes the use of Multivariate Analysis of Variance, Principal Component Analysis, Canonical Correlation Analysis, Multidimensional Scaling and some analytical variants of these methods (for a complete description, see [4]). In their paper, the authors out some difficulties in using the described methods including the normality, homoscedasticity and/or linearity strict assumptions. Generalized linear models are mentioned as a strategy to accommodate variance heterogeneity and/or non-normality. However, there is no reference in how to deal with high-dimensional datasets which are very likely to arise in these studies as the number of species fingerprints is commonly high (increasing the number of variables) and the number of captures tends to be small to minimize invasive sampling (decreasing the sample size). Further, dimensional reduction by variable selection is seldom addressed in the field literature, although, in this context, finding which features are most relevant for predicting the outcome is essential, as attaining information on each of these features is commonly highly time and cost demanding. In addition not all fingerprints necessarily have to be used when assembling models to predict the geographic origin of a seafood species. In fact, if not properly selected, some elements may even act as confounding variables between origins [5].

The analysis of high-dimensional datasets may be challenging, specially with a multi-categorical outcome, as typically the model will be included at least one parameter per category. In addition, as the number of covariates increases relatively to the sample size, problems with the convergence of parameter estimates arise and the usual maximum likelihood estimates do not exist [6]. Regularization penalty-based models, including *Ridge regression*, LASSO (*Least Absolute Shrinkage and Selection Operator*) and *Elastic net*, solve these limitations. Despite not being new these techniques are relatively less used by statisticians which traditionally tend to adopt more classic approaches.

In this paper we propose a methodological path for the problem of predicting a multi-categorical response variable, based on regularized models. This approach is suggested to allow a sparse representation of the link between predictors and response, solving, at one go, multicollinearity and feature selection issues. In addition, we propose the resampling method of *Monte Carlo Cross-Validation* to study the models performance and overcome the difficulties related to the small sample size.

To exemplify the use of the proposed methodology, these methods were applied to predict the origin of the Manila clam (*Ruditapes philippinarum*), a non-indigenous marine bivalve species that commonly occurs in European estuaries. This species was chosen because, when illegally harvested from places where captures have been forbidden (due to public health issues), it can pose a threat to consumers if traded [7]. A good example of this scenario is that of the illegal harvesting of Manila clams from the Tagus estuary in Portugal. The capture of this highly priced bivalve from this specific ecosystem has been prohibited by national authorities due to public health issues [5]. Nonetheless, this prohibition has not deterred poachers to illegally capture these bivalves and supply them to organized crime networks that make them available to consumers. This criminal practice is made possible through the mislabeling of the place of origin of Manila clams, taking advantage of the loopholes of current traceability protocols currently incapable of safeguarding the labeling of seafood geographic origin from fraud. Thus, we assembled regularization-based methods to oppose the fraudulent practice of mislabeling the place of harvesting of seafood.

The structure of the paper is as follows. Section 2 summarizes the regularization approaches and the model validation strategies. In Sect. 3 we describe the dataset, present and discuss the results, comparing the three regularization methods regarding the models fitting and their predictive performance. Finally, in Sect. 4 we summarize the main conclusions and comment on how the proposed method may be further used and contribute toward seafood traceability.

2 Methodology

In this section we summarize the proposed methodological path. All the procedures were coded using `R` software [8]. Penalty-based regularization approaches were implemented using the `glmnet` package [9], with the assistance of the packages `ensr` [10] and `glmnetUtils` [11]. *K-fold Cross-Validation* approach was ran at each iteration of the *Monte Carlo Cross-Validation* using `glmnet` package [9]. Simultaneous penalization and mixing parameters optimization was implemented using `caret` package [12]. Finally, ROC curves were constructed resorting to packages `ROCR` [13] and `PRROC` [14].

Furthermore, to avoid difficulties in reproducing the proposed methodology and ensure transparency we have created a dedicated GitHub repository (https://github.com/rb1970/RM4Traceability) in which we made available an R Markdown file and the used data to assist in replicating the methodology.

2.1 Regularization

Regularization is generally obtained by maximizing the penalized log-likelihood of the model. Different types of penalties result in different methods. In this section we briefly describe the approaches known as *Ridge regression*, LASSO and *Elastic net*.

2.1.1 Ridge regression

Ridge regression [15] minimizes the residual sum of squares plus the sum of the squared beta coefficients affected by a parameter, λ ($\lambda \geq 0$), that regulates the strength of the penalization, i.e., the model coefficients β are estimated by

$$\text{argmin}_{\beta}\left(\sum_{i=1}^{n}(\text{y}_i - \beta_0 - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right) \tag{1}$$

given the data $\{(\mathbf{x}_i, y_i)\}$ $(i = 1, ..., n)$ with Y being the response variable and $\mathbf{x}$ the vector of predictors. This method regulates the dimensionality by shrinking the coefficients of the less relevant features toward 0, but never actually removes features from the model. It can achieve good performance measures, but given that it does not perform feature selection, it is commonly seen as less convenient when dealing with high-dimensional datasets [16].

2.1.2 LASSO

LASSO [17] also imposes a penalization to the residual sum of squares but in this case given by the sum of the absolute values of the coefficients, i.e.,

$$\text{argmin}_{\beta}\left(\sum_{i=1}^{n}(\text{y}_i - \beta_0 - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right). \tag{2}$$

This method has proven to be a flexible tool that performs variable selection allowing to shrink some coefficients to zero, for some suitably chosen value of λ. Despite its popularity, it is generally inefficiency in dealing with groups of highly correlated predictors. In such circumstances, for each group of correlated variables, LASSO tends to arbitrarily select one predictor into the model, discarding the others [18]. Another drawback of the LASSO method is not being able to select more variables than the sample size. In addition, some studies show that LASSO aggressivity might eliminate interesting features and be outperformed by *Ridge regression* [19].

2.1.3 Elastic net

Elastic net penalty [20] combines LASSO and *Ridge regression* approaches through the penalty term [21]

$$\lambda\left[\frac{1-\alpha}{2}\sum_{j=1}^{p}\beta_j^2+\alpha\sum_{j=1}^{p}|\beta_j|\right]. \tag{3}$$

where the parameter α ($\alpha \geq 0$) controls the mixture between the two types of penalties. As described, the parameter λ is responsible for controlling the strength of the penality. As this non-negative parameter increases the regularization effect is strengthened, and the model develops into keeping the coefficients small instead of minimizing the loss function. The α parameter controls the proximity between *Ridge regression*, *Elastic net* and LASSO penalties. A value of $\alpha = 0$ corresponds to *Ridge regression* and $\alpha = 1$ implements LASSO approach. The *Elastic net* method is suited to deal with highly correlated predictors [18] and, at the same time, perform feature selection. Further, it can retain more than n explanatory variables [19].

2.2 Resampling Techniques

The *K-fold Cross-Validation* method is one of the most common approaches to optimize the penalization and mixing parameters. Generically, it selects optimal values as the ones that lead to the model with the smallest error [22, 23]. When $K = 1$, this procedure consists in one simply partitioning of the data into two complementary sets: training and testing, fitting the model with the training set and evaluating the predictive quality of the model on the testing set. For more statistically meaningful results, multiple rounds ($K > 1$) of cross-validation can be executed. In this case the validation results for each iteration can be combined to postulate about the overall predictive performance of the model.

However, this validation method can only be performed within the span of K iterations, making it dependent on the size of the dataset. *Monte Carlo Cross-Validation* overcomes this limitation. This method also combines the results of several iterations to postulate about the overall performance of a model but, unlike *K-fold Cross-Validation* which, at each iteration, divides the original dataset into K complementary partitions and creates a random split of the dataset into training and testing sets [24]. For this reason, some observations may never be selected to a testing set, whereas others may be selected more than once throughout the different iterations of this process. Hence, contrarily to *K-fold Cross-Validation*, it is not dependent on sample size.

In this study, optimal hyperparameters, λ and α, were estimated through *fivefold Cross-Validation* ran for each regularization method at each iteration of the *Monte Carlo Cross-Validation*. For *Ridge* and LASSO approaches, the optimal λ value was chosen as the one that culminated in the model with the smallest percentage of misclassification error, while selecting the least number of predictors into the model (for LASSO). For *Elastic net* we estimated the optimal mixing and penalization parameters using *fivefold Cross-Validation*, selecting the parameters that originated the model with the lowest error.

To validate the results, we implemented a 1000 iterations *Monte Carlo Cross-Validation* method, constructing 1000 testing and training sets at random, and hence fitting 1000 distinct models for each regularization approach. We ran and tested each model separately and, ultimately, for each regularization approach, combined the results of the 1000 fitted models. At each iteration, the dataset was partitioned into training and testing sets with a respective ratio of 80%/20%.

2.3 Model Validation

To perform a model validation, we analyzed cross-entropy, confusion matrices and ROC (*receiver operating characteristic*) curves.

2.3.1 Cross-entropy

Cross-entropy (also named *Log-Loss*, LL) measures how close a model is to predicting the correct class with probability 1. It is defined as

$$LL = -\frac{1}{n}\sum_{i=1}^{n}\sum_{q=0}^{J} a_{iq}\log(p_{iq}), \tag{4}$$

where a_{iq} $(i = 1, \ldots, n;\ q = 0, \ldots, J)$ is a binary indicator of whether or not label q is the correct classification for the instance i, and p_{iq} is the predicted probability of assigning label q to instance i.

Confusion matrices

Confusion matrices contrast models prediction with the known class affiliation. Each entry of these matrices indicates the number of predictions made by the model, and whether the classes were classified correctly or incorrectly. Based on these tables several measures can be computed to evaluate models performance, including *Accuracy*, *Precision*, *Recall*, *Micro-F1*, *Macro-F1* and *Weighted F1* indicators (definitions in "Appendix B").

ROC curves

ROC curves allow to evaluate the discriminatory ability of a binary classification model by mapping *True positive rate vs*. *False positive rate* (definitions in "Appendix B"). An uninformative diagnosis tool is represented by the line with unit slope. ROC curve is typically described using the *Area Under Curve* (AUC) index, defined by the curve integral between 0 and 1. A perfect diagnosis tool has an AUC $= 1$ and a diagonal line, corresponding to an uninformative tool, has an AUC $= 0.5$. Model A is said to be a better diagnosis tool than model B if the respective AUC statistics are order such that $\text{AUC}_A \geq \text{AUC}_B$.

3 Application

In this section we present a case study. We start by describing the data (Sect. 3.1), followed by discussing the models fitting (Sect. 3.2) and models performance (Sect. 3.3) results.

3.1 Data

Ruditapes philippinarum, a saltwater clam, is a marine bivalve that is commercially harvested for human consumption, being one of the most important bivalve species grown in aquaculture worldwide [25–28]. The place of origin of bivalve species is expected to be traceable by features such as their biochemical and geochemical fingerprints [5, 7, 29].

The available dataset contained detailed information on the bio- and geochemical compositions of the adductor muscle and the shell (respectively) of *Ruditapes philippinarum*, including 30 clam samples and 44 composition features. The data, collected in the summer of 2018, included the geographic origin of the clams (response variable), a three class variable with the levels: `RV` (*Ria de Vigo*; 8° 43′ 9.59″ W, 42° 15′ 38.44″ N), `RAv` (*Ria de Aveiro*; 8° 41′ 18.93″ W, 40° 46′ 6.95″ N) and `TE` (*Tagus Estuary*; 9o 0′ 58.66″ W, 38° 45′ 16.55″ N). The 44 predictors were all continuous variables including 26 concerning the quantification of the fatty acids of the adductor mussel of the clams (biochemical fingerprints, Table 3) and the remaining 18 concerning the quantification of the chemical elements on the composition of their shell (geochemical fingerprints, Table 4).

As the performance of the regularization methods depends on the data characteristics, namely on the existence of groups of correlated data, we present the main descriptive statistics of the dataset in "Appendix A". Two traits are worth to highlight: (1) magnitudes and standard deviations are quite different across variables, so data were standardized (Table 5), and (2) the data contain several groups of variables highly correlated (Fig. 5).

3.2 Models Fitting

Model fitting for each of the regularization methods was performed considering the optimal penalization and mixing estimated parameters selected at each iteration. In this section, we comment on the hyperparameters estimates and coefficients shrinkage obtained across the 1000 models.

3.2.1 Hyperparameters Estimation

Table 1 summarizes the obtained penalization parameter ranges, and Fig. 1 depicts the parameters observed distributions. *Ridge regression* consistently attained much higher penalization values than the other methods (Fig. 1A), which indicates a stronger shrinkage effect. As this method does not perform variable selection, severe shrinkage is the means to force the distinction between the least and the most important predictors.

Table 1 Models fitting summary

	Ridge regression	LASSO	*Elastic net*
λ (log scale)	1.451–6.056	−5.366 to −1.092	−3.9203 to −0.4746
Number of coefficients	135	6–20	6–74
Residuals	0.2–0.8	0.0001–0.711	0.003–0.688

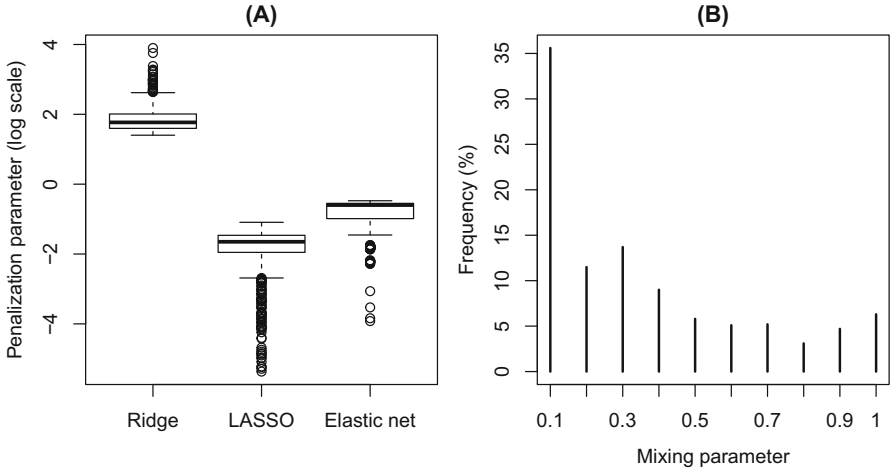

Fig. 1 **A** Penalization and **B** mixing parameters observed distributions across the 1000 models

LASSO and *Elastic net* penalization values were closer to each other with *Elastic net* presenting, as expected, a compromised result between the other two approaches. *Elastic net* mixing parameter distribution shows clear mode at 0.1 (Fig. 1B). This means that, in most cases, this model prioritized shrinking the coefficients toward zero without eliminating them entirely. Thus, these models tend to retain relatively more variables favoring less sparse solutions (note that the maximum number of coefficients included in *Elastic net* models was 74, Table 1).

3.2.2 Coefficients Shrinkage and Variable Selection

Without dropping any of the 44 predictors, *Ridge regression* required 135 ($44 \times 3 + 3$) coefficients at each iteration (Table 1). LASSO selected a very small number of variables to predict each class ranging from 6 to 20, regardless of the low penalty values (Table 1). In fact, 80% of the LASSO models included less than 10 coefficients. This aggressive feature selection was most likely due to how this approach handles groups of correlated variables, as multicollinearity was a marked trait in the dataset (see Fig. 5) and LASSO tends to keep just one variable representing each group which limited the number of coefficients included in the models. *Elastic net* models kept a number of coefficients between 6 and 74 (Table 1), with nearly 35% of the models

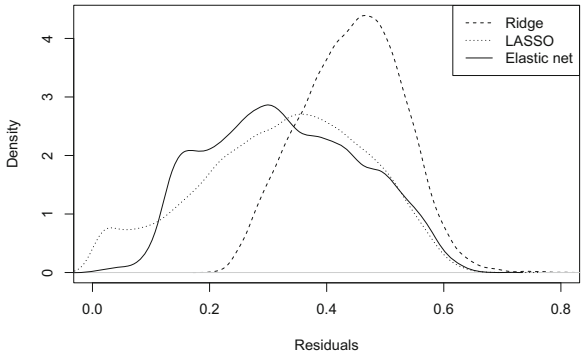

Fig. 2 Residuals densities obtained from using the three regularization methods across the 1000 models

having between 60 and 70 coefficients. In summary, LASSO approach originated the most parsimonious models, followed by *Elastic net* and *Ridge regression* methods.

3.2.3 Residual Analysis

Table 1 summarizes the obtained residuals ranges for the three approaches. Figure 2 depicts the respective estimated densities. LASSO and *Elastic net* methods present quite similar ranges and estimated residual densities. These curves display a flattened shape, with values relatively scattered, ranging from 0.0001 to approximately 0.7. In contrast, the density of the residuals from the *Ridge regression* has a completely different shape and central tendency. This curve is more narrow and is centered around higher residual values, ranging from approximately 0.2 to 0.8. Thus, *Ridge regression* has the highest residuals deviance. LASSO and *Elastic net* display similar adjustments.

3.3 Models Performance

In this section we discuss models performance when applied to the test datasets, through the analysis of cross-entropy values, confusion matrices and ROC curves. Table 2 summarizes the performance measures for the three studied regularization methods.

Cross-entropy
Based on the predicted probabilities of affiliation, we computed the cross-entropy values for each iteration of *Monte Carlo Cross-Validation* and each regularization method (Table 2, Fig. 3). In general, three approaches performed similarly. However, the results show that the cross-entropy values varied considerably more across LASSO models than among *Ridge regression* or *Elastic net* models, suggesting that LASSO performance could be considered less stable.

Confusion matrices
Metrics based on confusion matrices (definitions in "Appendix B") are summarized in Table 2 for each regularization approach. Note that *Ridge regression* misclassification rates reach values up to 12%. *Elastic net* achieves a maximum value around 6%, and LASSO shows a maximum value of only 3%. All the other performance metrics

Table 2 Models performance summary

	Ridge regression	LASSO	*Elastic net*
Cross-entropy	0.418–1.665	0.007–5.727	0.076–1.922
Misclassification rate (%)	0.1–12.2	0.3–2.9	0.2–5.9
Accuracy	0.879–0.999	0.971–0.997	0.941–0.998
Precision	0.802–0.997	0.965–0.999	0.884–0.998
Recall	0.798–1.000	0.957–0.993	0.884–0.996
Specificity	0.898–0.998	0.975–0.999	0.939–0.999
F1	0.818–0.998	0.956–0.996	0.913–0.997
Micro-F1	0.878	0.971	0.941
Macro-F1	0.879	0.971	0.941
Weighted F1	0.878	0.971	0.941
AUC	0.86–0.99	0.96–0.99	0.89–0.99

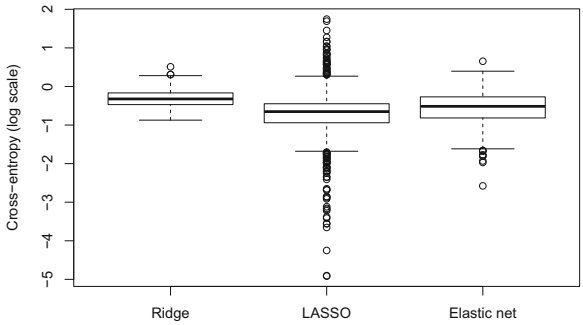

Fig. 3 Cross-entropy (log scale) distributions obtained from using the three regularization approaches across the 1000 models

indicate consistently a lower performance for *Ridge regression*. The other two methods show very similar results. However, LASSO never attains as low performances as the ones reached by *Elastic net* models. These results suggest that LASSO models present the best overall performance, followed closely by *Elastic net* and lastly by *Ridge regression*.

ROC curves

ROC curves are depicted in Fig. 4. The predictive ability was remarkably good for all classes regardless of the regularization approach (Fig. 4). *Ridge regression* showed the poorest results with AUC values between approximately 0.86 and 0.99 (Table 2). The LASSO method performed better than *Ridge regression* with AUC values between 0.96 and 0.9990 (Table 2). Overall, LASSO also performed better than *Elastic net* with AUC values ranging from 0.89 to 0.99 (Table 2). In summary, based on ROC curves, LASSO performed better than *Elastic net*, followed by *Ridge regression*.

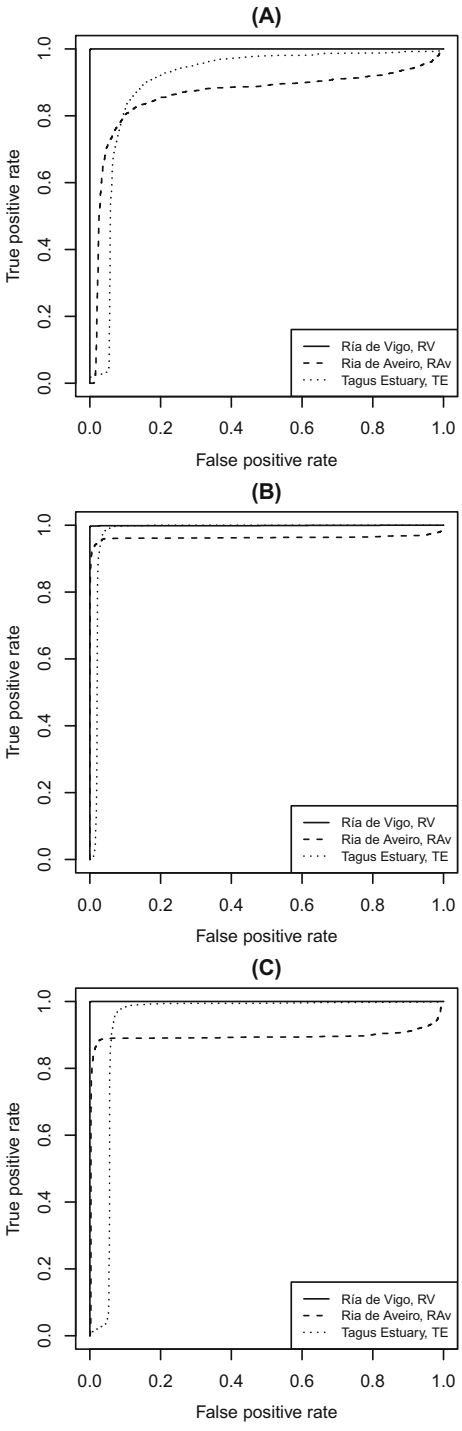

Fig. 4 ROC *curves* for **A** *Ridge regression*, **B** LASSO and **C** *Elastic net* approaches

4 Conclusion

In this paper we propose a methodological path for the problem of predicting a multi-categorical response variable based on high-dimensional data. This methodology uses regularized models to deal with multicollinearity and allow variable selection. Further, we propose the resampling method of *Monte Carlo cross-validation* to overcome the difficulties related to the small sample size while analyzing models performance. Regularization methods included *Ridge regression*, LASSO and *Elastic net* penalty-based approaches. The dimensionality reduction ability and the predictive quality of the models were studied by applying complementary methods of model validation, including the analysis of cross-entropy values, confusion matrices and ROC curves.

Overall, the *Ridge regression* was the poorest performing method. Besides being incapable of removing variables from the models, it also systematically presented the worst performance metrics. LASSO and Elastic net methods clearly outperformed *Ridge regression*. LASSO and *Elastic net* performed equivalently being hard to emphasize one as better, since both outperformed the other in different aspects. In fact, although the presence of severe multicollinearity favored LASSO method in terms of feature selection, this method showed the least stable cross-entropy values attaining, in some cases, very high prediction losses. Nonetheless, we stress that the number of features selected by LASSO was much smaller than the number selected by the *Elastic net* and LASSO showed consistently high prediction quality. We have applied this methodological path to the problem of predicting the geographical origin of Manila clams, under a scenario of severe small sample size and highly correlated predictors. This example has shown that it is possible to select predictors without compromising the models prediction performance. Our application also demonstrated the adequacy of the proposed method in dealing with high-dimensional data to study traceability. We believe and hope that our study may represent a valuable contribution to fight against illegal, unreported, and unregulated (IUU) fishing, one of the major threats to achieve United Nations Sustainable Development Goal 14—Life Below Water [30].

Although the establishment of this methodology was initially motivated by this specific problem, its application potential goes evidently beyond this particular context. Our study presents results that may be used to guide the modelling process of a multi-categorical response based on high-dimensional dataset with highly correlated explanatory variables, which is relevant in diverse contexts. Further, by providing the code and data in an open-source manner, we aim to eliminate barriers that often hinder the reproducibility of research. Thus, researchers and practitioners can easily access the code, gaining a deeper understanding of the steps involved in implementing the methodology. They can also use the provided data to validate and compare their own results, fostering a collaborative and iterative approach. With this open sharing of resources not only we intend to promote transparency but also empower others to build upon our work and contribute to the collective knowledge in the field.

Funding Open access funding provided by FCT|FCCN (b-on). This work is funded by national funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects UIDB/00297/2020 and UIDP/00297/2020 (Center for Mathematics and Applications).

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix A Explanatory Variables Description

See Tables 3, 4, 5 and Fig. 5.

Table 3 Biochemical fingerprints

	Common name	Fatty acid
1	Myristic acid	FA14:0
2	Pentadecanoic acid	FA15:0
3	Palmitic acid	FA16:0
4	Palmitoleic acid	FA16:1n-7
5	(Z)-7-Hexadecenoic acid	FA16:1n-9
6	Margaric acid	FA17:0
7	Stearic acid	FA18:0
8	Cis-vaccenic acid	FA18:1n-7
9	Oleic acid	FA18:1n-9
10	Linoleic acid	FA18:2n-6
11	Alpha-linolenate acid	FA18:3n-3
12	Stearidonic acid	FA18:4n-3
13	Paulinic acid	FA20:1n-7
14	Gondoic acid	FA20:1n-9-11
15	Eicosadienoic acid	FA20:2n-6
16	Dihomo-gamma-linolenic acid	FA20:3n-6
17	Eicosatetraenoic acid	FA20:4n-3
18	Arachidonic acid	FA20:4n-6
19	Eicosapentaenoic acid	FA20:5n-3
20	Docosatrienoic acid	FA22:2n-6
21	Docosadienoic acid	FA22:2n-9
22	Docosahexaenoic acid	FA22:3n-6
23	Docosatetraenoic acid	FA22:4n-6
24	Docosapentaenoic acid	FA22:5n-3
25	Adrenic acid	FA22:5n-6
26	Docosahexanoic acid	FA22:6n-3

Table 4 Geochemical fingerprints

	Common name	Symbol
1	Sodium	Na
2	Zinc	Zn
3	Magnesium	Mg
4	Strontium	Sr
5	Aluminum	Al
6	Yttrium	Y
7	Phosphorus	P
8	Bariu	Ba m
9	Manganese	Mn
10	Lanthanum	La
11	Iron	Fe
12	Cerium	Ce
13	Cobalt	Co
14	Neodymium	Nd
15	Nickel	Ni
16	Gadolinium	Gd
17	Copper	Cu
18	Uranium	U

Table 5 Fatty acids (FA) and elements to calcium concentration ratios per site

Variables	*Ria de Vigo*	*Ria de Aveiro*	*Tagus Estuary*
FA14:0	0.009 (1.82×10^{-3})	0.005 (1.58×10^{-3})	0.004 (7.47×10^{-4})
FA15:0	0.004 (5.73×10^{-4})	0.005 (1.25×10^{-3})	0.005 (8.66×10^{-4})
FA16:0	0.139 (5.87×10^{-3})	0.123 (4.45×10^{-2})	0.129 (6.63×10^{-3})
FA16:1n-9	0.001 (2.48×10^{-4})	0.003 (4.62×10^{-3})	0.002 (4.40×10^{-4})
FA16:1n-7	0.041 (9.96×10^{-3})	0.017 (5.12×10^{-3})	0.023 (4.31×10^{-3})
FA17:0	0.01 (1.35×10^{-3})	0.014 (9.69×10^{-4})	0.013 (1.37×10^{-3})
FA18:0	0.082 (9.37×10^{-3})	0.098 (1.08×10^{-2})	0.078 (7.58×10^{-3})
FA18:1n-9	0.037 (3.09×10^{-3})	0.045 (3.62×10^{-3})	0.048 (9.62×10^{-3})
FA18:1n-7	0.026 (2.96×10^{-3})	0.016 (2.12×10^{-3})	0.019 (1.58×10^{-3})
FA18:2n-6	0.002 (4.67×10^{-4})	0.002 (6.34×10^{-4})	0.003 (6.13×10^{-4})

Table 5 continued

Variables	*Ria de Vigo*	*Ria de Aveiro*	*Tagus Estuary*
FA18:3n-3	0.005 (8.95×10^{-4})	0.004 (1.01×10^{-3})	0.009 (1.95×10^{-3})
FA18:4n-3	0.009 (1.83×10^{-3})	0.011 (2.11×10^{-3})	0.015 (1.98×10^{-3})
FA20:1n-9-11	0.041 (3.21×10^{-3})	0.044 (8.10×10^{-3})	0.048 (3.35×10^{-3})
FA20:1n-7	0.032 (2.80×10^{-3})	0.024 (2.30×10^{-3})	0.028 (2.39×10^{-3})
FA20:2n-6	0.017 (1.52×10^{-3})	0.015 (2.87×10^{-3})	0.018 (3.61×10^{-3})
FA20:3n-6	0.002 (4.87×10^{-4})	0.001 (4.35×10^{-4})	0.002 (4.13×10^{-4})
FA20:4n-6	0.03 (4.05×10^{-3})	0.036 (5.38×10^{-3})	0.037 (3.32×10^{-3})
FA20:4n-3	0.009 (1.55×10^{-3})	0.005 (1.49×10^{-3})	0.008 (1.51×10^{-3})
FA20:5n-3	0.173 (9.88×10^{-3})	0.094 (1.54×10^{-2})	0.116 (7.49×10^{-3})
FA22:2n-9	0.009 (8.87×10^{-4})	0.014 (2.90×10^{-3})	0.012 (1.16×10^{-3})
FA22:2n-6	0.03 (4.30×10^{-3})	0.031 (9.58×10^{-3})	0.032 (3.78×10^{-3})
FA22:3n-6	0.015 (1.69×10^{-3})	0.008 (9.27×10^{-4})	0.012 (1.39×10^{-3})
FA22:4n-6	0.019 (2.16×10^{-3})	0.014 (2.56×10^{-3})	0.015 (2.28×10^{-3})
FA22:5n-6	0.01 (1.90×10^{-3})	0.015 (1.55×10^{-3})	0.022 (6.78×10^{-3})
FA22:5n-3	0.059 (3.53×10^{-3})	0.04 (4.50×10^{-3})	0.044 (4.83×10^{-3})
FA22:6n-3	0.189 (9.80×10^{-3})	0.316 (3.18×10^{-2})	0.26 (1.31×10^{-2})
Na	21.886 (1.62)	24.195 (5.28×10^{-1})	24.369 (6.06×10^{-1})
Mg	0.485 (4.98×10^{-2})	0.596 (6.52×10^{-2})	0.709 (1.57×10^{-1})
Al	0.041 (2.65×10^{-2})	0.141 (7.81×10^{-2})	0.109 (5.24×10^{-2})
P	0.617 (2.01×10^{-1})	0.447 (1.04×10^{-1})	0.525 (1.20×10^{-1})
Mn	0.02 (3.47×10^{-2})	0.004 (2.45×10^{-3})	0.046 (3.11×10^{-2})
Fe	0.161 (1.29×10^{-1})	0.223 (8.94×10^{-2})	0.45 (3.29×10^{-1})
Co	0.004 (1.44×10^{-4})	0.004 (6.30×10^{-5})	0.004 (6.22×10^{-4})
Ni	0.005 (4.52×10^{-4})	0.005 (1.72×10^{-3})	0.004 (2.61×10^{-4})
Cu	2.74×10^{-4} (1.74×10^{-4})	9.06×10^{-4} (5.68×10^{-4})	0.008 (2.22×10^{-2})
Zn	4.82×10^{-4} (1.75×10^{-5})	0.002 (3.55×10^{-3})	0.016 (4.31×10^{-2})
Sr	1.763 (6.56×10^{-2})	1.562 (8.16×10^{-2})	1.44 (8.15×10^{-2})
Y	4.14×10^{-4} (7.17×10^{-5})	2.86×10^{-4} (3.76×10^{-5})	2.58×10^{-4} (3.93×10^{-5})
Ba	0.003 (1.57×10^{-3})	0.004 (1.35×10^{-3})	0.003 (7.03×10^{-4})
La	4.95×10^{-5} (3.18×10^{-5})	3.85×10^{-5} (4.49×10^{-5})	3.29×10^{-5} (2.15×10^{-5})
Ce	1.13×10^{-4} (6.14×10^{-5})	1.08×10^{-4} (9.29×10^{-5})	8.86×10^{-5} (4.99×10^{-5})
Nd	3.15×10^{-5} (2.20×10^{-5})	3.37×10^{-5} (3.79×10^{-5})	3.06×10^{-5} (1.95×10^{-5})
Gd	1.01×10^{-5} (8.01×10^{-6})	8.88×10^{-6} (1.05×10^{-5})	6.97×10^{-6} (6.47×10^{-6})
U	5.89×10^{-8} (3.15×10^{-8})	2.51×10^{-8} (1.64×10^{-8})	1.27×10^{-8} (5.53×10^{-9})

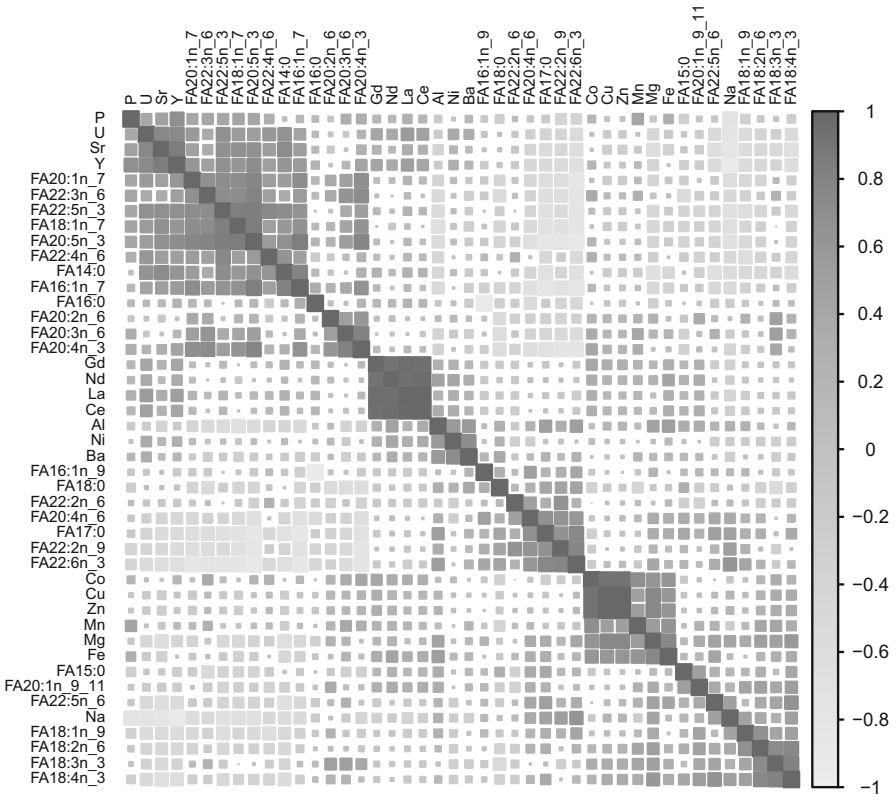

Fig. 5 Data correlation matrix

Appendix B Confusion Matrices Metrics

Table 6 displays a multi-class confusion matrix. Let TP, TN, FP and FN denote *Total Positive*, *Total Negative*, *False Positive* and *False Negative* counts for each individual class.

Table 7 showcases the formulas for the direct metrics of a 3-class classification problem.

Using these metrics, we can then calculate the performance measures for each class, simply applying formulas (5), (6), (7), (8), (9) and (10):

- **Accuracy:** Fraction of samples that were correctly classified by the model.

$$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{5}$$

- **Misclassification Rate:** Fraction of incorrect predictions.

$$\frac{\mathrm{FP}+\mathrm{FN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \quad \text{or} \quad 1-\mathrm{Accuracy} \tag{6}$$

Table 6 Confusion matrix for a 3-class classification problem

Predicted Class \ True Class	A	B	C	total
A	True A TP_A	False A $F_{A(B)}$	False A $F_{A(C)}$	T'_A
B	False B $F_{B(A)}$	True B TP_B	False B $F_{B(C)}$	T'_B
C	False C $F_{C(A)}$	False C $F_{C(B)}$	True C C	T'_C
total	T_A	T_B	T_C	

- **Precision:** Fraction of positive class predictions that were actually positive.

$$\frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

If, for a certain class, TP + FP = 0, then we consider Precision = 1, since the model did not actually fail to predict the class.

- **Recall/True Positive Rate/Sensitivity/Probability of Detection:** Fraction of all positive samples that the model correctly predicted as positive.

$$\frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

Once more, if, for a certain class, TP + FN = 0, then we consider Recall = 1, since the model did not actually fail to predict the class.

- **Specificity/True Negative Rate:** Fraction of all negative samples are correctly predicted as negative.

$$\frac{\text{TN}}{\text{FP} + \text{TN}} \tag{9}$$

In this case, if, for a certain class, FP + TN = 0, then we consider Specificity = 1, since the model did not actually fail to predict any negative samples.

Table 7 Formulas showcase for the metrics of a 3-class classification problem

	A	B	C
True Positive	TP_A	TP_B	TP_C
True Negative	$\mathrm{TN}_A = \mathrm{TP}_B + F_{B(C)} + F_{C(B)} + \mathrm{TP}_C$	$\mathrm{TN}_B = \mathrm{TP}_C + F_{A(C)} + F_{C(A)} + \mathrm{TP}_C$	$\mathrm{TN}_C = \mathrm{TP}_C + F_{A(B)} + F_{B(A)} + \mathrm{TP}_B$
False Positive	$\mathrm{FP}_A = F_{A(B)} + F_{A(C)}$	$\mathrm{FP}_B = F_{B(A)} + F_{B(C)}$	$\mathrm{FP}_C = F_{C(A)} + F_{C(B)}$
False Negative	$\mathrm{FN}_A = F_{B(B)} + F_{C(A)}$	$\mathrm{FN}_B = F_{A(B)} + F_{C(B)}$	$\mathrm{FN}_C = F_{A(C)} + F_{B(C)}$

- **F1-Score:** Merging Precision and Recall into a single measure, it is, mathematically, the harmonic mean of precision and recall.

$$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{10}$$

If $2\text{TP} + \text{FP} + \text{FN} = 0$ for a particular class, signifying no sample from that class was selected to the testing set, we consider F1-Score = 1, since the model did not actually fail to predict the class.

Additional performance measures can be computed for the overall model, instead of considering the performance for each individual class. Considering that:

- Total True Positives (*Total TP*): $\text{TP}_A + \text{TP}_B + \text{TP}_C$
- Total False Positives (*Total FP*): $\text{FP}_A + \text{FP}_B + \text{FP}_C$
- Total True Negatives (*Total TN*): $\text{TN}_A + \text{TN}_B + \text{TN}_C$
- Total False Negatives (*Total FN*): $\text{FN}_A + \text{FN}_B + \text{FN}_C$

These measures include:

- *Total Precision:*

$$\frac{\text{Total TP}}{\text{Total TP} + \text{Total FP}} \tag{11}$$

- *Total Recall:*

$$\frac{\text{Total TP}}{\text{Total TP} + \text{Total FN}} \tag{12}$$

- *Micro-F1:* Assesses the quality of multi-classification problems. Simply put, it measures the F1-score of the aggregated contributions of all classes.

$$2 \times \frac{\text{Total Precision} \times \text{Total Recall}}{\text{Total Precision} + \text{Total Recall}} \tag{13}$$

- *Macro-F1:* Calculates the F1-score metrics for each class individually and then takes unweighted mean of the measures.

$$\frac{\text{F1-Score}_A + \text{F1-Score}_B + \text{F1-Score}_C}{3} \tag{14}$$

where

$$\text{F1-Score}_k = \frac{2\text{TP}_k}{2\text{TP}_k + \text{FP}_k + \text{FN}_k}$$

- *Weighted F1:* It takes the weighted mean of the measures. The weights for each class are the total number of samples of that class.

$$\frac{TA \times \text{F1-Score}_A + TB \times \text{F1-Score}_B + TC \times \text{F1-Score}_C}{TA + TB + TB} \tag{15}$$

References

1. Astill J, Dara RA, Campbell M, Farber JM, Fraser ED, Sharif S, Yada RY (2019) Transparency in food supply chains: a review of enabling technology solutions. Trends Food Sci Technol 91:240–247
2. Leal MC, Pimentel T, Ricardo F, Rosa R, Calado R (2015) Seafood traceability: current needs, available tools, and biotechnological challenges for origin certification. Trends Biotechnol 33(6):331–336
3. Bennion M, Morrison L, Shelley R, Graham C (2021) Trace elemental fingerprinting of shells and soft tissues can identify the time of blue mussel (*Mytilus edulis*) harvesting. Food Control 121:107515
4. Kotsanopoulos K, Martsikalis PV, Gkafas GA, Exadactylos A (2022) The use of various statistical methods for authenticity and detection of adulteration in fish and seafood. Crit Rev Food Sci Nutr. https://doi.org/10.1080/10408398.2022.2117786
5. Mamede R, Ricardo F, Santos A, Díaz S, Santos SA, Bispo R, Domingues MRM, Calado R (2020) Revealing the illegal harvesting of Manila clams (*Ruditapes philippinarum*) using fatty acid profiles of the adductor muscle. Food Control 118:107368
6. Zahid FM, Tutz G (2013) Multinomial logit models with implicit variable selection. Adv Data Anal Classif 7(4):393–416
7. Ricardo F, Pimentel T, Maciel E, Moreira AS, Domingues MR, Calado R (2017) Fatty acid dynamics of the adductor muscle of live cockles (lit Cerastoderma edule) during their shelf-life and its relevance for traceability of geographic origin. Food Control 77:192–198
8. R Core Team (2021) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. R Foundation for Statistical Computing. https://www.R-project.org/
9. Simon N, Friedman J, Hastie T, Tibshirani R (2011) Regularization paths for Cox's proportional hazards model via coordinate descent. J Stat Softw 39(5):1–13
10. DeWitt P (2019) Ensr: elastic Net SearcheR. R package version 0.1.0. https://CRAN.R-project.org/package=ensr
11. Ooi H (2021) glmnetUtils: utilities for Glmnet. R package version 1.1.8. https://CRAN.R-project.org/package=glmnetUtils
12. Kuhn M (2020) Caret: classification and regression training. R package version 6.0-86. https://CRAN.R-project.org/package=caret
13. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. Bioinformatics 21(20):7881
14. Grau J, Grosse I, Keilwagen J (2015) PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. Bioinformatics 31(15):2595–2597
15. Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12(1):55–67
16. Aseervatham S, Antoniadis A, Gaussier É, Burlet M, Denneulin Y (2011) A sparse version of the Ridge logistic regression for large-scale text categorization. Pattern Recogn Lett 32(2):101–106
17. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Methodol) 58(1):267–288
18. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1
19. Ogutu JO, Schulz-Streeck T, Piepho H-P (2012) Genomic selection using regularized linear regression models: ridge regression, LASSO, Elastic net and their extensions. In: BMC proceedings, vol 6. Springer, pp 1–6
20. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Stat Methodol) 67(2):301–320
21. Hastie T, Qian J, Tay K (2016) An introduction to glmnet
22. Jung Y (2016) Efficient tuning parameter selection by cross-validated score in high dimensional models
23. Obuchi T, Kabashima Y (2016) Cross validation in LASSO and its acceleration. J Stat Mech Theory Exp 2016(5):053304
24. Xu Q-S, Liang Y-Z (2001) Monte Carlo cross validation. Chemom Intell Lab Syst 56(1):1–11
25. Mamede R, Ricardo F, Abreu MH, da Silva EF, Patinha C, Calado R (2021) Spatial variability of elemental fingerprints of sea lettuce (*Ulva* spp.) and its potential use to trace geographic origin. Algal Res 59:102451

26. Mamede R, Ricardo F, Gonçalves D, da Silva EF, Patinha C, Calado R (2021) Assessing the use of surrogate species for a more cost-effective traceability of geographic origin using elemental fingerprints of bivalve shells. Ecol Ind 130:108065
27. Bennion M, Morrison L, Brophy D, Carlsson J, Abrahantes JC, Graham CT (2019) Trace element fingerprinting of blue mussel (*Mytilus edulis*) shells and soft tissues successfully reveals harvesting locations. Sci Total Environ 685:50–58
28. Morrison L, Bennion M, Gill S, Graham CT (2019) Spatio-temporal trace element fingerprinting of king scallops (*Pecten maximus*) reveals harvesting period and location. Sci Total Environ 697:134121
29. Ricardo FAF (2017) Use of biogeochemical tools to trace the origin of bivalves-first steps towards origin certification. Ph.D. thesis, Universidade de Aveiro (Portugal)
30. FAO (2021) 14.4.1 Fish stocks sustainability | Sustainable Development Goals | Food and Agriculture Organization of the United Nations. http://www.fao.org/sustainable-development-goals/indicators/1441/en/. Online; accessed 6 Oct 2021