Near infrared spectral fingerprinting: A tool against origin-related fraud in the sector of processed anchovies

Near infrared spectral fingerprinting: a tool against origin-related fraud in the sector of processed anchovies

Maria Olga VARRÀ, Sergio GHIDINI, Adriana IANIERI, Emanuela ZANARDI[*]

Department of Food and Drug, University of Parma, Strada del Taglio 10, 43126 - Parma, Italy

E-mails: mariaolga.varra@studenti.unipr.it (M. O. Varrà); sergio.ghidini@unipr.it (S. Ghidini); adriana.ianieri@unipr.it (A. Ianieri); emanuela.zanardi@unipr.it (E. Zanardi).

*CORRESPONDING AUTHOR:

Emanuela Zanardi

Department of Food and Drug, University of Parma

Strada del Taglio 10, 43126 - Parma (Italy)

Phone: +39.0521.902.760

Email: emanuela.zanardi@unipr.it

ABSTRACT

In the present study near-infrared (NIR) spectroscopy was used to assess the geographical traceability of salted ripened anchovies, whose raw product originated from fishing areas of Morocco, Spain, Tunisia, and Croatia. Two different products were tested: semi-finished and finished salted anchovies. The development and optimization of combined discrimination models based on orthogonal partial least square-discriminant analysis successfully led to the identification of the geographical origin of both anchovy datasets with >98% sensitivity, >99% specificity, and >99% accuracy on average. While NIR absorption bands related to proteins and degradation compounds highly characterized samples from Morocco and those of unsaturated lipids and derivatives globally contradistinguished anchovies from Tunisia, absorptions of both protein and lipid compounds were responsible for the discrimination of samples from Croatia and Spain. The proposed method is particularly helpful to guarantee the authenticity of salted ripened anchovies and, thus, to deter commercial frauds throughout the fish value chain and ensure traceability along the whole food chain.

Abbreviations

Area under the Receiver Operating Characteristic Curves, AUROC; Central Mediterranean Sea, CM; cross-validation, CV; analysis of variance testing of cross-validation predictive residuals, CV-ANOVA; Croatia, CR; Morocco, MO; multiple scatter correction, MSC; near infrared spectroscopy, NIRS; orthogonal partial last square-discriminant analysis, OPLS-DA; principal component analysis, PCA; principal component, PC; root mean square error from cross-validation, RMSECV; root mean square error of estimation, RMSEE; root mean square error of prediction, RMSEP; second derivative, 2SD; Savitzky-Golay smoothing, SG; Spain, SP; Tunisia, TU; variable influence on projection, VIP;

Keywords: food authenticity; traceability; fishery products; rapid methods; geographical origin; spectroscopy; chemometrics.

1. Introduction

Salted anchovy is obtained through the processing of the fresh anchovy, consisting of salting of the beheaded and partially gutted raw fish, followed by fermentation until the degree of ripeness required is reached. The salt-cured Mediterranean-style anchovy (Martin, Carter, Flick, & Davis, 2000) is obtained from the species *Engraulis encrasicholus* L., which naturally populates the Eastern North and Central Atlantic, Mediterranean, Western Black, and Azov seas coasts (FAO, 2010). Spain, Italy, Greece, Tunisia, France, and Morocco have a strong tradition in producing and consuming salted anchovies to the point of representing the main large-scale producers as well the leading exporters of the product to the whole Mediterranean area (EUMOFA, 2018).

From a commercial point of view, the supply chain of anchovy is overly complex and diversified. The salt-cured product obtained by fermentation can be considered as an intermediate product intended for further minimal or heavy processing. The anchovy semi-preserves meant for the direct sales are simply obtained by salt-packaging of the whole fermented dressed anchovy. Nevertheless, many other transformed products can be found in the marketplace, as filleted salted anchovy preserved in oil, anchovy paste, and several gastronomic preparations.

Based on both the production process adopted from the fish industry and the desired organoleptic characteristics of the final product, the duration of the ripening process may extend from a minimum of 2-3 months to an average time of 12-15 months. During the ripening process complex chemical, biochemical and physicochemical changes occur in the fish flesh, determining the conversion of the major compounds and the development of final aroma and sensory traits. The magnitude and the speed of these reactions are the results of the specific processing conditions (e.g. temperature, pH, salt, water activity) plus the original biological characteristics of the fresh fish. Fat, protein, water content, endogenous enzymatic pattern, and microbiota

composition of the raw fish are in fact fully reflected in the overall qualitative attributes of the salted ripened anchovy (Hernandez-Herrero, Roig-Sagues, Lopez-Sabater, Rodriguez-Jerez, & Mora-Ventura, 2002). In turn, the whole composition of the raw fish tissues is influenced by several factors, among which environment characteristics (e.g. temperature and salinity of the water) are determinant (Romotowska, Karlsdóttir, Gudjónsdóttir, Kristinsson, & Arason, 2016).

Although the indication of the geographical origin of semi-preserved pre-packed fish products is not mandatory under the current European legislation (European Parliament and Council of the European Union, 2013), the provision of increasingly detailed information to the consumer through the label is growing. This is due to a number of reasons, including the interest towards sustainable fisheries and the correlation between origin and perceived quality and tradition of the product. On the other hand, the fish supply chain is subjected to such a complex sector regulation that the monitoring of traceability and labelling of processed fish through all the stages of production, processing, and distribution is equally complex (Regulation (EU) No 1379/2013). These factors together have all meant that adequate and quick tools to verify the origin authenticity of raw material, semi-finished, and final products are becoming of utmost importance.

The use of NIR spectroscopy so as to find information related to the geographical origin is an elaborated and often unsuccessful process, since geographical provenance is the sum of a huge amount of different intrinsic or extrinsic factors which are difficult to be properly identified (Ghidini, Varrà, & Zanardi, 2019). Despite this, few but significant applications of NIR spectroscopy aimed at discriminating fish according to different countries or fishing areas of origin are currently available in literature (Liu et al., 2015; Guo et al., 2018; Ghidini et al., 2019). In these cases, the success in addressing the origin authenticity issues was likely due to the use of chemometrics to handle complex spectral data. In addition, these works converged upon the conclusion that the characteristics of the aquatic environment (latitude, water temperature, water exchange, salinity) reflects on the whole final product in term of flesh composition. Protein patterns, but, above all fatty acid profiles, represented the most affected fish constituents to these variations and, as such, their high detectability by NIR spectroscopy made them useful indicators of geographical origin. Other targeted investigations underlined still more significantly the close relationship between fatty acid profiles

and the geographical provenance of fish and seafood (Ricardo et. al, 2015; Zhang, Liu, Li, & Zhao, 2017). Besides, some proteins were shown to be expressed as different allozymes in relation to different provenances (Mork & Giæver, 1999; Drengstig Fevolden, Galand, & Aschan, 2000) or overexpressed is response to environmental stressing conditions (El Sheikha & Montet, 2016).

Origin traceability explored in NIR-based studies was limited to raw unprocessed fish products, the characteristics of which were not altered or modified by processing and whose composition was more closely related to the provenance. Nonetheless, in several other fermented foodstuffs such as wine, cheese, tea, and soy sauce, the geographical origin of the raw material employed was correctly discriminated by NIR spectroscopy (Liu et al., 2008; Pillonel, Schaller, Picque, Cattenoz, & Bosset, 2005; Ren et al., 2013; Iizuka & Aishima, 1997). Similar applications concerning fermented fish products are not available since the technique has been exploited only to study the proximate composition and to establish optimal parameters for fermentation (Huang et al., 2001; Huang et al., 2003; Svensson, Nielsen, & Bro, 2004).

On the light of these considerations, the present work was aimed to guarantee the geographical traceability of salted ripened anchovies throughout some stages of the production chain by using a fingerprinting approach based on NIR spectroscopy and chemometrics. For this reason, both semi-finished products and finished semi-preserved products originating from different geographical origins, i.e. Morocco, Spain, Tunisia, and Croatia were considered.

2. Materials and Methods

2.1. Sampling of anchovy specimens

Two sets of products of different provenances and obtained from the industrial scale processing of European anchovy (*Engraulis encrasicolus* L.) were considered in the present study: i) semi-finished salted anchovies (intended for further packaging or processing, i.e. filleting and oil preservation); ii) finished salted anchovies packaged in glass jars (intended for direct marketing).

The salted, ripened products were industrially manufactured according to the traditional process following a standard procedure. Briefly, the raw anchovies were firstly immersed into a saturated brine solution,

121 manually beheaded, partially eviscerated, and then rinsed again with brine solution. Fish was then placed
122 into barrels and layered alternatively with sodium chloride. The salt-curing process in barrels was carried out
123 under pressure and at controlled room temperature (18 °C), for different months. After ripening, fish from
124 each provenance and from different batches was randomly taken from barrels and directly vacuum-
125 packaged into plastic bags to obtain the semi-finished product, or it was further processed by addition of salt
126 and packaged into glass jars (net weight=280 g) to obtain the finished product.
127 The total number of anchovy specimens to be considered for the construction of statistical models aimed at
128 discriminating fish according to different origin was chosen on the basis of a recent systematic revision of
129 the literature dealing with the use of vibration spectroscopy to assess fish authenticity (Ghidini, Varrà, &
130 Zanardi, 2019), according to which an average number of 20 specimens per provenance was reported as
131 suitable for achieving more than satisfactory results in terms of robustness of the final discriminant model.
132 The set of semi-finished anchovies included 350 samples coming from 4 different geographical areas,
133 namely Spain (SP, Cantabrian Sea, FAO fishing area 27.8), Morocco (MO, Eastern Central Atlantic, FAO
134 fishing area 34.1), Croatia (CR, upper Central Mediterranean Sea, FAO fishing area 37.2.1), and Tunisia
135 (TU, lower Central Mediterranean Sea, FAO fishing area 37.2.2). The same geographical provenances, with
136 the exception of the Moroccan fishery zone, were also investigated for the set of finished anchovies (250
137 samples from different glass jars), i.e. SP (FAO fishing area 27.8), CR (FAO fishing area 37.2.1), and TU
138 (FAO fishing area 37.2.2). Detailed information on the sampling and characteristics of the transformed fish
139 used in the present study is reported in Tab. 1.
140 Although the experimental design was initially conceived in such a way that the same number and the same
141 provenances of both semi-finished and finished anchovy fish were included in the final datasets, during the
142 experimental work some finished samples from Morocco were no longer available and, therefore, they were
143 not provided by the fish canning company. Moreover, the choice to develop discriminant models for
144 anchovies from Morocco, Spain, Tunisia, and Croatia but not from Italy was justified by the fact that the
145 volumes of catch of anchovy fish are larger along the Moroccan, Spanish, Tunisian, and Croatian areas than
146 in Italy. As a consequence, the Italian industrial transformation of local anchovy fish takes place on a smaller

147 scale compared to other countries, leading to the presence on the market of very diversified products that are
148 less suitable for standardisation. Moreover, Italian transformed products are specifically addressed to local
149 and niche markets.

150 *2.2. Sample preparation*

151 Plastic bags and glass jars containing the anchovies were stored at refrigerated temperature (4 ± 2 °C). At the
152 time of analysis each individual fish was removed from the package and patted dry with filter paper to
153 remove the excess salt. The skin, viscera, fins, and main bone were carefully removed, and the resulting two
154 fillets were manually and finely chopped by knife until a doughty consistence was reached and non-
155 homogeneous fragment of fish muscles were no longer visible. To reduce inter-individual variability, two
156 minced fillets of the same specimens were then merged and mixed with the two minced fillets obtained from
157 another specimen to create a unique, final sub-sample intended for NIR spectra acquisition.
158 The workflow summarising the experimental procedure adopted in the present work, from samples provision
159 to spectra statistical elaboration, is illustrated in Fig. 1.

160 *2.3. NIR spectroscopy*

161 Round quartz sample holders (35 mm diameter) were filled with approx. 8 g of minced sample and scanned
162 in diffuse reflectance mode by using a NIRFlex® N-500 FT-NIR benchtop spectrometer (Büchi
163 Labortechnik AG, Flawil, Switzerland) over the wavelength range of 1000–2500 nm and at a spectral
164 interval of 1 nm (32 scans/spectrum). At regular intervals of acquisition (every 20 samples) the instrument
165 was calibrated using a proper external calibration standard intended for the establishment of the accuracy of
166 the wavelength scale and consisting of the rare earth oxide Standard Reference Material® 1920a of the
167 National Institute of Standards and Technology. To obtain more accurate results, four replicates of the NIR
168 spectrum were collected for each sample by rotating 90° the sample holder three times. The four spectra
169 were further averaged.

170 *2.4. Multivariate data analysis*

171 Mean reflectance (R) NIR spectra were converted into absorbance (A) spectra (A = log 1/R). The range
172 1000–1199 nm was excluded from the subsequent chemometric elaboration, since it was characterised by a
173 low signal-noise ratio. Therefore, the final data matrix of semi-finished salted anchovies included a total of
174 1301 variables (NIR absorbance values in the 1200-2500 nm range) and 350 spectra, while data matrix of
175 finished salted anchovies included 1301 variables and 250 spectra.
176 Chemometrics analysis was performed by the software SIMCA-P 14.1 (Umetrics, Umeå, Sweden). Raw
177 spectral data were firstly elaborated by means of principal component analysis (PCA) to look at the data
178 structure, verify whether the only application of the unsupervised data modelling was suitable to achieve a
179 separation of sample groups, and identify strong outlier samples through the Hotelling's T^2 test (at 95%
180 confidence interval).
181 Multi-class orthogonal partial least square-discriminant analysis (OPLS-DA) was instead applied to create
182 combined models for discriminating and predicting the geographical origin of the two data matrices. OPLS-
183 DA is a qualitative regression-based supervised discriminant analysis allowing to maximize the separation
184 among classes through the partitioning of the total variation within data into a related predictive variation
185 (enclosing information related to the discrimination purpose) and an orthogonal (non-related) variation
186 (Trygg & Wold, 2002). OPLS-DA is a modification of the classical partial least square discriminant analysis
187 (PLS-DA) which, despite also being aimed at maximising class separation, does not allow to slit the total
188 variation into predictive and non-predictive variations. Therefore, PLS-DA may result in misinterpretation of
189 the results (Bylesjö, Rantalainen, Cloarec, Nicholson, Holmes, & Trygg, 2007).
190 Prior to perform OPLS-DA, the reduction of the undesirable large baseline shifts, as well as the separation of
191 the typical combination spectral bands, were attempted by applying the following mathematical pre-
192 treatments to the raw spectra: multiplicative scatter correction (MSC), second derivative (2SD, second
193 polynomial order, 15 point), and Savitzky-Golay smoothing (SG, 15 points).
194 While PCA was performed on the whole data, the OPLS-DA calibration models were computed on the
195 training set including 80% of randomly selected and representative original spectra (280 spectra for semi-
196 finished anchovies and 200 spectra for finished anchovies). The 20% of the spectra left (70 spectra for semi-

finished anchovies; 50 spectra for finished anchovies) was completely independent and was used as test set to externally validate the OPLS calibration models. Considering that no general rules concerning data splitting ratio are available (Westad & Marini, 2015) and given that the number of total samples to be investigated in the present work was quite high, the 80/20 partitioning of data, according to the so-called Pareto principle (Massart, Vandeginste, Buydens, Jong, Lewi, & SmeyersVerbeke, 1997), was initially investigated. Since the results achieved were more than satisfactory in terms of prediction, therefore other possible data splitting ratios were not tested. Moreover, the 80/20 splitting ratio was already reported by other authors to provide a robust and stable model for the discrimination of the geographical origin of fish dataset composed of approximately the same number of samples (Guo et al., 2017).

2.4.1. Validation of the chemometric models

An internal 10-fold cross validation (CV) was employed to establish the correct number of principal component (PCs) for PCA and predictive the orthogonal components for OPLS-DA, based on the lowest significant values of root mean square error from CV (RMSECV) achieved. The overall quality of the resulting OPLS-DA models was estimated by the following statistical indicators: $R^2X_{(cum)}$ (total amount of variation of spectra, i.e. X-matrix, collected); $Q^2_{(cum)}$ (predictive variation of X-matrix collected); $R^2Y_{(cum)}$ (total amount of variation of OPLS classifier models strictly related to class membership of samples, i.e. Y-matrix); p-values from the analysis of variance of the cross-validation residuals (CV-ANOVA). Finally, the high risk of overfitting and overestimation of the OPLS-DA models was averted by permuting the class labels of samples 200 times in CV. Models were considered valid and robust when the resulting values of R^2 and Q^2 Y-intercepts were lower than 0.4 and 0.05, respectively (Van der Voet, 1994).

As for the external validation procedure of the OPLS classifiers, sensitivity, specificity, and overall accuracy parameters were calculated for each of the groups in which the new samples of the test were classified. Sensitivity corresponded to the proportion of true positive samples, specificity to the proportion of true negative samples, and accuracy to the proportion of true positive and negative samples. To sum up how the combination of different values of sensitivity and specificity impacted the model discrimination

performances, Areas under the Receiver Operating Characteristic curves (AUROC) were also calculated for each predicted group. Sensitivity, specificity, accuracy, and AUROC values equal to 1, suggested perfect outcomes in prediction of the discriminant model (Forina, Armanino, Leardi, & Drava, 1991). For further details about the equations to calculate of these parameters, readers are referred to Fawcett (2006).

Finally, the variable influence on projection (VIP) analysis for the OPLS predictive components was applied to better interpret the discrimination models and pinpoint the most relevant NIR wavelengths for the discrimination of the geographical origins of the anchovies (VIP indexes ≥ 1).

2.4.2. Identification of the NIR signature features of each geographical provenance

Since in multi-class OPLS-DA the VIP analysis does not provide information about the spectral regions bearing the separating information for single classes, subsidiary pairwise OPLS-DA models were built (i.e. by discriminating the single classes by two, consecutively). The NIR bands with the highest power of discrimination between two classes were therefore identified via the resulting S-line plots, by looking at the extent of the predictive loadings and at the associated absolute values of the correlation coefficients (r).

3. Results and Discussion

3.1. Overview of spectral characteristics and sample natural distribution by PCA

Raw NIR absorbance spectra, as such, were regarded as unfit and misleading for statistical elaboration (data not shown). The corrected averaged spectra (MSC, 2SD, SG) of the semi-finished and finished anchovies coming from different fishing areas are reported in Fig. 2. No apparent differences in terms of spectral patterns and shapes were present in the spectra, although some slight changes in peak intensity were observed in the 2020–2090 nm region (whose absorption peak are linked to protein, urea, oil, and -OH group), 2180–2200 nm region (mainly associated to protein absorption), and 2400–2440 nm region (absorption of CH of aryl functional group) of the semi-finished anchovies (Fig. 2A) (Shenk, et al., 2001; Workman & Weyer, 2012). Similarly, the 2000–2050 (protein, urea) and 2400–2450 (CH of aryl functional group) nm regions of the NIR spectra of finished anchovies showed some variations (Fig. 2B). The most

prominent absorption bands (negative peaks) of both anchovy products were observed as follows: around 1360 nm (hydrocarbons), 1730 nm (hydrocarbons/alcohol), 2010–2100 nm (protein, urea, oil, and -OH), 2180 nm (proteins), 2350 nm (aryl functional group), and 2390–2460 nm (aryl functional group) (Shenk, et al., 2001; Workman & Weyer, 2012).

From the application of PCA, a total of 13 PCs explaining 88.5% of total variation of the original raw spectra were calculated in CV for the semi-finished anchovy dataset ($R^2X_{(cum)}$ =0.885, $Q^2_{(cum)}$=0.733). Similarly, 15 PCs explaining 83.5% of variation were extracted in CV for the finished anchovy dataset ($R^2X_{(cum)}$ =0.835, $Q^2_{(cum)}$=0.657). Despite numerous, all the extracted components were found to carry enough spectral information to be retained in CV during the computation of the analysis, but the first two PCs calculated from the semi-finished and finished anchovy datasets, enclosed 72% and 64% of the total variations, respectively, thus underling the redundance of the other PCs. This phenomenon is likely the consequence of a large original number of spectral variables (1501 absorption values) to be compressed by PCA on the one hand, and of the natural multicollinearity proper of NIR absorption bands on the other.

As expected, no evident clusters of samples were visualized in the score plots of the first two PCs of both anchovy datasets, although for Moroccan anchovy of the semi-finished anchovy dataset and for Tunisian anchovies of the finished anchovy dataset, a distribution along the negative axis of the PC2 and a trend in separating from the other groups was observed (Suppl. Figg. S1A, S2A). Spanish and Croatian semi-finished and finished samples appeared to be closer each other, but Croatian semi-finished and finished samples were found to be more widely scattered across the plot area compared to the other group, thus suggesting a greater diversification in terms of fish flesh composition.

During this preliminary data elaboration stage, some samples fell outside the 95% confidence interval (ellipse) of the PCA score plot, but they were not found to be strong outliers according to the Hotelling's T^2 test (Suppl. Figg. S1C, S2C).

By looking at the negative loadings of the PC2 of the semi-finished anchovy dataset (Suppl. Fig. S1B), the largest peaks pointing downwards were found to be around 1460, 1510, 1820, 1980, and 1990, all previously attributed to amides, proteins and urea groups absorption (Workman & Weyer, 2012). Since the influence

exerted by the large negative peaks in the 2140–2440 nm lipid-associated region was much lower, the protein fraction was considered the most significant in contradistinguish Moroccan specimens. By contrast, a stronger contribution of lipids to the slight separation of the specimens coming from Tunisia in the finished anchovy dataset was evident. Peaks associated with the absorption of lipids, hydrocarbon chains and urea located at 1390 and 1680 nm (hydrocarbons), 2030 and 2070 nm (urea), 2140 nm (lipids), 2350 nm (aromatic compounds), 2420 nm, and 2440 nm (CH aryl group) were found to be the most important loadings of the negative PC2 (Suppl. Fig. S2B). The two peaks located at 1530 and 2060 nm were associated to the absorption of CH group of alkynes/NH group of secondary amines and $CONH_2$ combination of amide and proteins (Workman & Weyer, 2012).

Nevertheless, preliminary results from PCA suggested that most of the total variability in the spectral pattern was not specifically related to the fishing geographical area and stated the view that a large amount of orthogonal variability hid the discrimination. This orthogonal variability is likely the sum of many other factors influencing spectral shapes and intensities which were previously verified to be dependent on feeding habits, age, muscular activity, competition, etc. (Guo et al., 2018). For these reasons, the predictive variability within the data was selectively extracted by OPLS methods and processed to obtain accurate classification outcomes.

3.2. Combined discrimination of geographical origins using OPLS-DA algorithm

Two prediction models based on multi-class OPLS-DA were built for the two sample datasets (training sets) to fit all the geographical provenances at once.

The application of OPLS-DA to the training set of the semi-finished anchovies led to the distinction of samples into four main clusters in the resulting score plot (Suppl. Fig. S3A), each corresponding to one of the four different fishing areas under investigation. A clear-cut distribution of CR and MO specimens in one specific quadrant of the score plot was not observed, but they were perfectly discriminated along the t[1], while anchovies from TU and SP groups reflected along the t[2].

296 Despite the fitting ability of 74% ($R^2X_{(cum)}$=0.742), it was found that only 24% of spectral variability was
297 predictive and enclosed by 3 components; the remaining 50% was not correlated with the groups and was
298 collected by 6 orthogonal components. Nevertheless, the high values of explained predictive variability
299 ($R^2Y_{(cum)}$=0.878) and predictive power ($Q^2_{(cum)}$=0.812), as well as results from the permutation test and CV-
300 ANOVA (p<0.05) proved the validity of the calibration model (Tab. 2).
301 Results from the recognition of the unknown samples of the test set by the fitted calibration model are
302 reported in Fig. 3 and Tab. 3. MO, SP, and TU anchovies were 100% assigned to the correct classes.
303 Although the natural biological diversity within specimens of the same group, only one CR sample was
304 wrongly classified as coming from SP, as it fell in the other class-regions (Fig. 3D). This misclassified
305 sample was responsible for the lower levels of accuracy of both SP and CR classifiers (accuracy=0.98) and
306 the lower level of specificity of 0.98 of the SP classifier compared to the other groups. By consequence, SP
307 group also showed the lowest AUROC value of 0.69. Furthermore, RMSEP values in prediction (Tab. 3)
308 were found to be remarkably similar to RMSECV values in calibration (Tab. 2), thus reconfirming the
309 validity of the fitted model as an estimator of the class membership of new anchovy specimens.
310 As for the main spectral differences responsible for the discrimination of samples in the score plot, the
311 absorption bands characterized by the higher VIP indexes for predictive components were located in the
312 2020–2120 nm region (with the maximum peak at 2030 nm, VIP=1.75), followed by the 2280–2330 and the
313 2360–2440 regions, all indicative of lipid absorption (Suppl. Fig. S3B). Other influential variables were
314 found at 1320–1400 and 1650–1685 nm, whose presence also disclosed the contribution of lipids to the
315 discrimination of samples by geographical origin. The same outcomes were previously discussed for PCA
316 applied to semi-finished products (see Sec. 3.1), pointing out a matching between the maximum variance
317 extracted by PCA and the maximum separating variance extracted by OPLS-DA, which corresponded
318 mainly to amide different profiles among groups.
319 As for the finished anchovy dataset, the discriminant OPLS calibration model was fitted in CV with 2
320 predictive and 6 orthogonal components, enclosing 19% and 50% of the total spectral variation, respectively.
321 The overall predictability reached 80% ($Q^2_{(cum)}$=0.805) (see Tab. 2), but the amount of variation exclusively

correlated with SP, CR and TU provenances was slightly higher ($R^2Y_{(cum)}$= 0.902) compared to the former model based on semi-finished products. Similarly, RMSECV values for the three classes were found to be lower compared to RMSECV values for SP, CR and TU semi-finished products.

A symmetrical separation of SP and TU samples along the t[1] was achieved, whit the first ones locating in the lower left part of the OPLS-DA score plot and the second ones in the lower right part (Suppl. Fig. S4A). At the same time, CR samples were tightly clustered along the positive axis of the t[2], straddling the positive and the negative axes of the t[1]. Despite the proximity of the fishing area of CR anchovies (FAO 37.2.1) to the fishing area of TU anchovies (FAO 37.2.2), no overlapping samples were observed in the plot.

The ability of the model to perfectly recognise 100% of the finished anchovies was also confirmed by the optimal statistics resulting from the external validation procedure (Tab. 3). The highest attainable values of sensitivity, specificity, accuracy, and AUROC were, in fact, reached for SP, CR and TU classes, without any sample of the test set being misclassified in the wrong geographical group, as it can be observed from the Y-predicted score plots reported in Fig. 4.

Similarly to what previously reported for the discrimination of the semi-finished anchovies, the VIP plot for the predictive components showed that the most important NIR peaks influencing sample discrimination were associated to the absorption of nitrogenous compounds, since located in the 2020–2120 nm (with the highest VIP value of 2.48 at 2055 nm), in the 2280–2330, and in the 2360–2440 nm regions (Suppl. Fig S4B). Other prominent but lower peaks corresponding to lipid absorption were observed at 1375 nm (VIP=2.05), 1665 nm (VIP=1.77), and 1750 nm (VIP= 1.61). Hence, in contrast to what was observed in the loading plot of the PCA (Sec. 3.1), the influence exerted by lipids was negligible, and likewise the OPLS model for semi-finished product discussed above, the differentiation of the geographical origin of the finished anchovies was mainly guided by variation associated to proteins and derivative molecules.

The lack of research related to the use of NIR spectroscopy to discriminate the geographical origin of fish and, in particular, to the use of VIP parameters to identify the most influential NIR absorption bands, hindered an in-depth and reliable comparison of the VIP indexes results obtained. Despite this, only one study performed on raw fish reported the use of VIP parameters but, in this case, the discrimination of fish

origin was mainly driven by the influence exerted by lipid absorption bands located in the 1620-1720 nm region (Ghidini et al., 2019). These conflicting results are likely to be the consequence of the different chemical composition and moisture content of raw fish compared to salted ripened anchovy.

3.3. Pairwise OPLS-DA and identification of class-related spectral hallmarks

The previously discussed multi-class models can be considered affordable and very advantageous in prevailing practice, since more than one provenance can be predicted at a time. Therefore, in the present section the whole point of applying OPLS-DA to couples of single classes was not to build classifiers to be used for the discrimination of the geographical origin, but rather to increase the understanding on how and how much single spectral variables contributed to the characterisation of each geographical provenance of both the semi-finished and the finished product.

3.3.1. Characterisation of samples from Morocco

By analysing the influence of NIR absorption bands to the separation of the semi-finished anchovies, two important loading values at 2060 and 2093 ($0.82 \leq r \geq 0.95$), linked to the absorption of proteins, peptides, and amino acids, plus one peak at 2426 nm (unassigned), always distinguished MO samples from the other three geographical groups (Figg. 5A, 5B and 5C). These loadings were specifically assigned to combination bands of $CONH_2$ of amide A and amide I from polypeptides, combination bands of O–H groups, and stretching/bending vibration of the CH_2 groups of the side chains of amino acids, respectively (Shenk, et al., 2001; Wang, Sowa, Ahmed, & Mantsch, 1994). The 1960–1990 nm NIR region (N–H combination bands of aromatic amines) (Workman & Weyer, 2012) and the peak at 2270 (N–H and C=O groups of peptide backbone referred to as the β-sheet structure) (Workman & Weyer, 2012) also showed their slight influence for the discrimination of MO from SP anchovies. In accordance with the results reported in the present study, the spectral variations in amide I absorption regions underlining modifications of the secondary structure of proteins were already reported in fish as being influenced by the ripening process of fish (Bocker, Kohler, Aursand, & Ofstad, 2008). Along with this, the possibility of effectively monitoring changes of the second

structure of proteins by using NIR spectroscopy was previously demonstrated also for different matrices. As an example, increasing NIR absorption peaks intensities at 2184 nm, 2259 nm, and 2276-2280 nm and decreasing NIR absorption peaks intensities at 2167 nm, 2209 nm, and 2264 nm were attributed to protein structural alterations of a qualitative nature (Bruun, Søndergaard, & Jacobsen, 2007). Moreover, peaks at 2172 nm and 2289 nm were attributed to α-helix structure, peaks at 2205 nm, 2264 nm, and 2313 nm were related to β-sheet structure, and the peak 2265 nm was found to be typical of the unordered protein structure (Robert, Devaux, Mouhous, & Dufour, 1999).

Finally, only one NIR peak assigned to the absorption of CH_3 groups of hydrocarbons (1705 nm) proved to be an effective discriminant variable for the separation of MO from CR anchovies, even though its lower level of correlation (r=0.52) (Fig. 5C).

Despite the well-known dependence of the lipidic composition on the geography of the fishing area, the study of protein and peptide variations has been demonstrated to be a useful tool to track the geographical origin of food, even if poorly applied to food of animal origin (Kumari et al., 2018; Wang et al., 2009). Modifications of the amino acid sequences of selected groups of proteins were already observed in raw hake specimens from American or African origin (Carrera & Gallardo, 2017), and the whole protein profile of different species of unprocessed shrimps showed an important correlation with their geographical provenance (Salla & Murray, 2013). Moreover, the induction of the expression of specific proteins in response to environmental pollutants has been demonstrated (Shepard & Bradley, 2000; Rodriguez-Ortega, Grosvik, Rodriguez-Ariza, Goksoyr, & Lopez-Barea, 2003). Therefore, proteomic could be further explored to identify those protein patterns more susceptible to change allow for the indirect identification of geographical origin of fish.

As for ripened products, in certain cases the relation between the peptide profiles and the country of origin was masked by fermentation process (Kumari et al., 2018), but some amino acids and degradation products of amine nature (trimethylamine, choline, and phosphocholine) were reported by other authors as useful markers of provenance of salted and dried mullet roe (Locci, Piras, Mereu, Cesare Marincola, & Scano, 2011).

Protein and polypeptide of the transformed anchovy products investigated in the present study also underwent several chemical and structural modification during the brining and the ripening processes, which made the final composition different, but reflecting the original composition of the raw products. Therefore, variations of the absorption of N–H, O–H and C–H groups assigned to proteins, indirectly supposes they mainly refer to the complex pattern of degradation compounds deriving from proteolysis, i.e. peptides, amino acids, peptide nucleotides and their decomposition fragments.

3.3.2. Characterisation of samples from Spain and Croatia

Contrary to MO samples, NIR wavelengths contributing the characterisation of SP and CR anchovies (both as semi-finished and as finished products) were highly variable based on the geographical counterpart to be compared in the two-class OPLS-DA models. For instance, SP semi-finished specimens stood out from MO specimens due to the impact of the 2030–2060 nm and the 2286–2420 nm regions, with the peaks at 2030 (C=0 stretching of urea or N-H combination bands from primary amides) (Workman & Weyer, 2012), 2055 (symmetric N–H stretching and amide I combination bands of proteins), and 2416 nm (C–H of aromatic molecules) showing the highest correlation coefficient ($r \geq 0.75$) (Fig. 5A). The peak found at 2084 nm was previously attributed to the absorption of substances such as methyl oleate/linoleate hydro-peroxides (Takamura, Hyakumoto, Endo, Matoba, & Nishiike, 1995).

Correlations of specific spectral bands with the perfect separation of SP from CR semi-finished samples ($r>0.65$) were instead observed as follows (Fig. 5E): 2290 ($CONH_2$, specifically due to the α-helix peptide structure), 2363 (CH_2 methylene group of aliphatic hydrocarbons), 2440 nm (C–H of aromatic hydrocarbons) (Workman & Weyer, 2012). Similarly, both proteins and lipids were responsible for the discrimination of SP from CR finished anchovies (Fig. 6C) since the SP specimens showed highly correlated peaks at 2290 nm ($r=0.67$) and at 2363 nm ($r=0.58$). The peak at 2090 (typical of O-H groups) was less relevant ($r=0.55$) for the discrimination of CR from TU semi-finished samples (Fig. 5F) and it was completely uncorrelated in the case of the finished anchovies (Fig. 6B), for which the most significant loadings corresponded to the absorption bands at 2065, 2426, and 2440 nm.

Proteins and derivatives discriminated CR from MO semi-finished samples (peaks at 2030, 2055, and 2416 nm) (Fig. 5C). Nevertheless, lipids and derivatives were more influential (r>0.6) in discriminating CR from SP semi-finished products (peaks at 1900, 2310, and 2318 nm, Fig. 5E) and CR from SP finished products (peak at 2380, Fig. 6C).

Therefore, a more balanced contribution of molecules of both protein and lipid origin justified the perfect discrimination of SP and CR anchovies from the other classes.

3.3.3. Characterisation of samples from Tunisia

The lipidic fraction has been reported in literature as the most susceptible to variation induced by environmental conditions, especially in oily fish such as anchovies. In the raw fish, unsaturated and polyunsaturated fatty acids such as oleic, eicosenoic, eicosapentaenoic, and docosahexaenoic acids and n-3/n-6 ratio were found to be the most influent lipidic molecules for the distinction of the geographical origin of anchovies (Öksuz, Özyilmaz, & Turan, 2009, Diraman & Dibeklioglu, 2009, Standal et al., 2012). Increases of the n-3 polyunsaturated fatty acids were found to be strongly linked to higher latitudes and, thus, to the water temperatures characterizing specific marine environment (Colombo, Wacker, Parrish, Kainz, & Arts, 2016).

In the ripened products, however, lipolysis and oxidation processes are responsible for the modification of the lipid fraction composition, as well as for the formation of several degradation compounds (Anggo, Ma, Swastawati, & Rianingsih, 2015). Therefore, the lipid composition of the raw fish changes considerably after the ripening process, but the signature of the original composition of the fresh fish remains in the pattern of new degradation metabolites.

In the present study, the impact of lipidic molecules to the discrimination of the semi-finished product coming from TU was the most pronounced and it was retained also after the processing to the finished products. The highest correlation values in TU vs. SP and TU vs. CR models for both semi-finished (Figg. 5D, 5F) and finished anchovies (Figg. 6A, 6B), corresponded to the 1694, 1950, 2144, 2310, 2380, and 2416 nm absorption peaks ($r \geq 0.65$). These bands were assigned to CH_3 second overtone of aliphatic hydrocarbons,

C=O stretching of acids and ester, C-H stretching/bending of lipids, and C–H of aromatic molecules deriving from lipids (Workman & Weyer, 2012). The peaks found at 2144 and 2310 nm, were previously reported to correspond to the C=C and C–H stretch combination tone of *cis* unsaturated fatty acids (Cozzolino, Murray, Chree, & Scaife, 2005), thus suggesting that the unsaturated lipidic fraction of fish may be strongly influenced by the environmental conditions.

4. Conclusions

In the present work, information related to the geographical origin of salt ripened anchovies was found in specific regions of the 1200-2500 nm NIR spectra of both the semi-finished and the finished product. This information was successfully extracted by means of multi-class OPLS-DA and the discriminant models, whose sensibility, specificity, and accuracy values were optimal, worked well thanks the overall differences in lipid and protein pattern among classes. A more in-depth evaluation of the anchovy spectra fingerprints was further carried out, with a view to bringing to light the distinctive compositional marks related to each single provenance. Proteins, peptides, amino acids, and proteolytic compounds signatures distinguished anchovies from Morocco, while the lipid signature and, presumably, the unsaturated lipid fraction closely characterised anchovies from Tunisia. At the same time, anchovies from Spain and Croatia were separated from the other groups, owing to the equal contribution of protein and lipid compounds. Although a more accurate characterisation of compounds involved in the discrimination of the anchovy origins by using auxiliary techniques would be helpful for the overall understanding of the NIR spectral changes observed, the present spectral fingerprinting strategy is particularly advantageous from the perspectives of rapidity, non-destructiveness, and cost-effectiveness in comparison to other well-known analytical approaches such as those based on chromatography and mass spectrometry. In addition, the proposed method, especially if transferred to handheld and portable NIR spectroscopy instrumentations, may be particularly suited for routine surveillance of transformed fish products, whose manufacturing industry is known to be characterized by extreme fast production rhythms. The authentication of anchovies is fundamental to ensure

the traceability of the product along the food chain, the loss of which may have detrimental impact on the safety due to the possible exposure to not identified hazards and related risks.

Acknowledgements

The research was carried out within the project "Development of rapid tools for anchovies authentication" funded by Rizzoli Emanuelli S.p.a. (Parma, Italy). The financial support and valuable cooperation of Rizzoli Emanuelli are gratefully acknowledged.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at

Declaration of interest: none.

References

Anggo, A. D., Ma, W. F., Swastawati, F., & Rianingsih, L. (2015). Changes of amino and fatty acids in anchovy (Stolephorus sp) fermented fish paste with different fermentation periods. *Procedia Environmental Sciences*, *23*, 58–63. https://doi.org/10.1016/j.proenv.2015.01.009

Bocker, U., Kohler, A., Aursand, I. G., & Ofstad, R. (2008). Effects of Brine Salting with Regard to Raw Material Variation of Atlantic Salmon (Salmo salar) Muscle. *Journal of Agricultural and Food Chemistry*, *56*, 5129–5137. https://doi.org/10.1021/jf703678z

Bruun, S. W., Søndergaard, I., & Jacobsen, S. (2007). Analysis of protein structures and interactions in complex food by near-infrared spectroscopy. 1. Gluten powder. *Journal of Agricultural and Food Chemistry*, 55(18), 7234-7243. https://doi.org/10.1021/jf063680j

Bylesjö, M., Rantalainen, M., Cloarec, O., Nicholson, J.K., Holmes, E., & Trygg, J. (2007) OPLS Discriminant Analysis, Combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics*, 20, 341–351. https://doi.org/10.1002/cem.1006

Carrera, M., & Gallardo, J. M. (2017). Determination of the Geographical Origin of All Commercial Hake Species by Stable Isotope Ratio (SIR) Analysis. *Journal of Agricultural and Food Chemistry*, *65*(5), 1070–1077. https://doi.org/10.1021/acs.jafc.6b04972

Colombo, S. M., Wacker, A., Parrish, C. C., Kainz, M. J., & Arts, M. T. (2016). A fundamental dichotomy in long-chain polyunsaturated fatty acid abundance between and within marine and terrestrial ecosystems. *Environmental Reviews*, *25*(2), 163-174. https://doi.org/10.1139/er-2016-0062

Cozzolino, D., Murray, I., Chree, A., & Scaife, J. R. (2005). Multivariate determination of free fatty acids and moisture in fish oils by partial least-squares regression and near-infrared spectroscopy. *LWT - Food Science and Technology*, *38*(8), 821–828. https://doi.org/10.1016/j.lwt.2004.10.007

Diraman, H., & Dibeklioglu, H. (2009). Chemometric characterization and classification of selected freshwater and marine fishes from turkey based on their fatty acid profiles. *JAOCS, Journal of the American Oil Chemists' Society*, *86*(3), 235–246. https://doi.org/10.1007/s11746-008-1338-3

Drengstig, A., Fevolden, S. E., Galand, P. E. and Aschan, M. M. (2000). Population structure of the deep-sea

525 shrimp (*Pandalus borealis*) in the north-east Atlantic based on allozyme variation. *Aquatic Living*
526 *Resources, 13*(2), 121–128. https://doi.org/10.1016/S0990-7440(00)00142-X
527 El Sheikha, A. F., & Montet, D. (2016) How to Determine the Geographical Origin of Seafood? *Critical*
528 *Reviews in Food Science and Nutrition, 56* (2), 306-317.
529 https://doi.org/10.1080/10408398.2012.745478
530 EUMOFA (European Market Observatory for Fisheries and Aquaculture products) (2018). Processed
531 anchovies in Italy. Price structure in the supply chain. https://doi.org/10.2771/181369
532 FAO (Food and Agriculture Organization of the United Nations) (2020). FAO Species Fact Sheets, Engraulis
533 encrasicolus (Linnaeus, 1758). http://www.fao.org/fishery/species/2106/en. Accessed 12.09.20
534 Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.
535 https://doi.org/10.1016/j.patrec.2005.10.010
536 Forina, M., Armanino, C., Leardi, R., & Drava, G. (1991). A class-modelling technique based on potential
537 functions. *Journal of Chemometrics, 5*(5), 435–453. https://doi.org/10.1002/cem.1180050504
538 Ghidini, S., Varrà, M. O., Dall'Asta, C., Badiani, A., Ianieri, A., & Zanardi, E. (2019). Rapid authentication
539 of European sea bass (Dicentrarchus labrax L.) according to production method, farming system, and
540 geographical origin by near infrared spectroscopy coupled with chemometrics. *Food Chemistry, 280*,
541 321–327. https://doi.org/10.1016/j.foodchem.2018.12.075
542 Ghidini, S., Varrà, M. O., & Zanardi, E. (2019). Approaching Authenticity Issues in Fish and Seafood
543 Products by Qualitative Spectroscopy and Chemometrics. *Molecules, 24*(9), 1812.
544 https://doi.org/10.3390/molecules24091812
545 Guo, X., Cai, R., Wang, S., Tang, B., Li, Y., & Zhao, W. (2018). Non-destructive geographical traceability
546 of sea cucumber (Apostichopus japonicus) using near infrared spectroscopy combined with
547 chemometric methods. *Royal Society Open Science, 5*(1). https://doi.org/10.1098/rsos.170714
548 Hernandez-Herrero, M. M., Roig-Sagues, A. X., Lopez-Sabater, E. I., Rodriguez-Jerez, J. J., & Mora-
549 Ventura, M. T. (2002). Influence of raw fish quality on some physicochemical and microbial
550 characteristics as related to ripening of salted anchovies (Engraulis encrasicholus L). *Journal of Food*

Science, *67*(7), 2631–2640. https://doi.org/10.1111/j.1365-2621.2002.tb08790.x

Huang, Y. H., Cavinato, A. G. C., Mayes, D. M. M., Kangas, L. J. K., Bledsoe, G. E. B., & Rasco, B. A. R. (2003). Nondestructive Determination of Moisture and Sodium Chloride in Cured Atlantic Salmon (Salmo salar) (Teijin) Using Short-wavelength. *Journal of Food Science*, *68*(2), 482–486. https://doi.org/10.1111/j.1365-2621.2003.tb05698.x

Huang, Y., Rogers, T. M., Wenz, M. A., Cavinato, A. G., Mayes, D. M., Bledsoe, G. E., & Rasco, B. A. (2001). Detection of Sodium Chloride in Cured Salmon Roe by SW - NIR Spectroscopy. *Journal of Agricultural and Food Chemistry*, *46*(6), 4161–4167. https://doi.org/doi.org/10.1021/jf001177f

Iizuka, K., & Aishima, T. (1997). Soy Sauce Classification by Geographic Region Based on NIR Spectra and Chemometrics Pattern Recognition. *Journal of Food Science*, *62*(1), 101–105. https://doi.org/10.1111/j.1365-2621.1997.tb04377.x

Kumari, N., Grimbs, A., D'Souza, R. N., Verma, S. K., Corno, M., Kuhnert, N., & Ullrich, M. S. (2018). Origin and varietal based proteomic and peptidomic fingerprinting of Theobroma cacao in non-fermented and fermented cocoa beans. *Food Research International*, *111*(February), 137–147. https://doi.org/10.1016/j.foodres.2018.05.010

Liu, L., Cozzolino, D., Cynkar, W. U., Dambergs, R. G., Janik, L., O'Neill, B. K., Colby, C. B., & Gishen, M. (2008). Preliminary study on the application of visible-near infrared spectroscopy and chemometrics to classify Riesling wines from different countries. *Food Chemistry*, *106*(2), 781–786. https://doi.org/10.1016/j.foodchem.2007.06.015

Liu, Y., Ma, D. H, Wang, X. C, Liu, L. P, Fan, Y. X, & Cao, J. X. (2015). Prediction of chemical composition and geographical origin traceability of Chinese export tilapia fillets products by near infrared reflectance spectroscopy. *LWT - Food Science and Technology*, *60*(2), 1214–1218. https://doi.org/10.1016/j.lwt.2014.09.009

Locci, E., Piras, C., Mereu, S., Cesare Marincola, F., & Scano, P. (2011). 1H NMR metabolite fingerprint and pattern recognition of mullet (Mugil cephalus) bottarga. *Journal of Agricultural and Food Chemistry*, *57*(17), 9497–9505. https://doi.org/10.1021/jf2012979

577 Martin, R., Carter, E., Flick, J., & Davis, L. (2000). *Marine & Freshwater Products Handbook*. (R. Martin,
578 E. Carter, J. Flick, & L. Davis, Eds.). Boca Raton, FL., USA: CRC Press (Taylor & Francis group).
579 https://doi.org/https://doi.org/10.1201/9781482293975
580 Massart, D.L., Vandeginste, B.G.M., Buydens, L.M.C., Jong, S.D., Lewi, P.J., & SmeyersVerbeke, P.
581 (1997). *Handbook of Chemometrics and Qualimetrics: Part A*. Elsevier, Amsterdam,. ISBN 10:
582 0444897240
583 Mork, J. & Giæver, M. (1999). Genetic structure of cod along the coast of Norway: Results from isozyme
584 studies. *Sarsia*, *84*(2),157–168. https://doi.org/10.1080/00364827.1999.10420442
585 Öksuz, A., Özyilmaz, A., & Turan, C. (2009). Comparative study on fatty acid profiles of anchovy from
586 Black Sea and Mediterranean Sea (Engraulis encrasichoius L., 1758). *Asian Journal of Chemistry*,
587 *21*(4), 3081–3086.
588 Pillonel, L., Schaller, E., Picque, D., Cattenoz, T., & Bosset, J. (2005). The potential of combined infrared
589 and fluorescence spectroscopies as a method of determination of the geographic origin of Emmental
590 cheeses. *International Dairy Journal*, *15*, 287–298. https://doi.org/10.1016/j.idairyj.2004.07.005
591 Regulation (EU) No 1379/2013 of the European Parliament and of the Council of 11 December 2013 on the
592 common organisation of the markets in fishery and aquaculture products, amending Council
593 Regulations (EC) No 1184/2006 and (EC) No 1224/2009 and repealing Council Regulation (EC) No
594 104/2000. *Official Journal of the European Union*, L354, 1–21.
595 Ren, G., Wang, S., Ning, J., Xu, R., Wang, Y., Xing, Z., Wan, X., & Zhang, Z. (2013). Quantitative analysis
596 and geographical traceability of black tea using Fourier transform near-infrared spectroscopy (FT-
597 NIRS). *FRIN*, *53*(2), 822–826. https://doi.org/10.1016/j.foodres.2012.10.032
598 Ricardo, R., Pimentel, T., Moreira, A. S., Rey, F., Coimbra, M. A., Domingues, M. R., Domingues, P., Costa
599 Leal, M., & Calado, R. (2015). Potential use of fatty acid profiles of the adductor muscle of cockles
600 (Cerastoderma edule) for traceability of collection site. *Scientific Reports*, *5*, 11125.
601 https://doi.org/10.1038/srep11125
602 Robert, P., Devaux, M. F., Mouhous, N., & Dufour, E. (1999). Monitoring the secondary structure of

603 proteins by near-infrared spectroscopy. *Applied spectroscopy*, 53(2), 226-232.
604 https://doi.org/10.1366/0003702991946361
605 Rodríguez-Ortega, M. J., Grøsvik, B. E., Rodríguez-Ariza, A., Goksøyr, A., & López-Barea, J. (2003).
606 Changes in protein expression profiles in bivalve molluscs (Chamaelea gallina) exposed to four model
607 environmental pollutants. *Proteomics*, *3*(8), 1535-1543. https://doi.org/10.1002/pmic.200300491
608 Romotowska, P. E., Karlsdóttir, M. G., Gudjónsdóttir, M., Kristinsson, H. G., & Arason, S. (2016). Seasonal
609 and geographical variation in chemical composition and lipid stability of Atlantic mackerel (Scomber
610 scombrus) caught in Icelandic waters. *Journal of Food Composition and Analysis*, *49*, 9–18.
611 https://doi.org/10.1016/j.jfca.2016.03.005
612 Salla, V., & Murray, K. K. (2013). Matrix-assisted laser desorption ionization mass spectrometry for
613 identification of shrimp. *Analytica Chimica Acta*, *794*, 55–59.
614 https://doi.org/10.1016/j.aca.2013.07.014
615 Shenk, J. S., Workman, J. J., & Westerhaus, M. O. (2001). Application of NIR spectroscopy to agricultural
616 products. In: D. A. Burns & E. W. Ciurczak (Eds.), Handbook of near-infrared analysis 3rd Ed.,
617 P*ractical Spectroscopy Series*, *35*, 356–357, Marcel Dekker, Inc., New York.
618 Shepard, J. L. & Bradley, B. P. (2000). Protein expression signatures and lysosomal stability in Mytilus
619 edulis exposed to graded copper concentrations. *Marine Environmental Research*, *50*(1), 5457–463.
620 https://doi.org/10.1016/S0141-1136(00)00119-7
621 Standal, I. B., Rainuzzo, J., Axelson, D. E., Valdersnes, S., Julshamn, K., & Aursand, M. (2012).
622 Classification of geographical origin by PNN analysis of fatty acid data and level of contaminants in
623 oils from Peruvian anchovy. *JAOCS, Journal of the American Oil Chemists' Society*, *89*(7), 1173–1182.
624 https://doi.org/10.1007/s11746-012-2031-0
625 Svensson, V. T., Nielsen, H. H., & Bro, R. (2004). Determination of the protein content in brine from salted
626 herring using near-infrared spectroscopy. *LWT - Food Science and Technology*, *37*(7), 803–809.
627 https://doi.org/10.1016/j.lwt.2004.03.004
628 Takamura, H., Hyakumoto, N., Endo, N., Matoba, T., & Nishiike, T. (1995). Determination of lipid

oxidation in edible oils by near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, *3*(4), 219–225. https://doi.org/10.1255/jnirs.72

Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, 16(3), 119–128. https://doi.org/10.1002/cem.695

Van der Voet, H. (1994). Comparing the predictive accuracy of models using a simple randomioation test. *Chemometrics and Intelligent Laboratory Systems*, *25*(2), 313–323. https://doi.org/10.1016/0169-7439(94)85050-X

Wang, J., Sowa, M. G., Ahmed, M. K., & Mantsch, H. H. (1994). Photoacoustic near-infrared investigation of homo-polypeptides. *Journal of Physical Chemistry*, *98*(17), 4748–4755. https://doi.org/10.1021/j100068a043

Wang, J., Kliks, M. M., Qu, W., Jun, S., Shi, G., & Li, Q. X. (2009). Rapid determination of the geographical origin of honey based on protein fingerprinting and barcoding using MALDI TOF MS. *Journal of Agricultural and Food Chemistry*, *57*(21), 10081–10088. https://doi.org/10.1021/jf902286p

Westad, F., & Marini, F. (2015). Validation of chemometric models–a tutorial. *Analytica chimica acta*, 893, 14-24. https://doi.org/10.1016/j.aca.2015.06.056

Workman, Jr.J, & Weyer, L. (2012). *Practical Guide and Spectral Atlas for Interpretive Near-Infrared Spectroscopy*. *Journal of Chemical Information and Modeling* (2nd ed.,). Boca Raton, FL., USA: CRC Press (Taylor & Francis group). https://doi.org/10.1017/CBO9781107415324.004

Zhang, X., Liu, Y., Li, Y., & Zhao, X. (2017). Identification of the geographical origins of sea cucumber (Apostichopus japonicus) in northern China by using stable isotope ratios and fatty acid profiles. *Food Chemistry*, *218*, 269–276. https://doi.org/10.1016/j.foodchem.2016.08.083

Figure Captions

Fig. 1. Workflow showing the followed experimental procedure.

Fig. 2. Pre-treated NIR spectra (MSC-2SD-SG) of semi-finished (A) and finished (B) anchovies and the main differences in spectral patterns. MO= blue solid line; SP=red dashed line; TU=green dotted line; CR=black dash-dotted line.

Fig. 3. Values of Y variables (geographical provenances) for MO (**A**), SP (**B**), TU (**C**), and CR (**D**) test set samples of the semi-finished anchovy dataset predicted by OPLS-DA. Levels for classification in the different groups were set based on the nearest class.

Fig. 4. Values of Y variables (geographical provenances) for TU (**A**), CR (**B**), and SP (**C**) test set samples of the finished anchovy dataset predicted by OPLS-DA. Levels for classification in the different groups were set based on the nearest class.

Fig. 5. Score plots of the first predictive (t[1]) and orthogonal (to[1]) components (left) and their respective S-line plots (right) for the pairwise OPLS-DA models separating the geographical origins of the semi-finished anchovy samples. (**A**) MO vs. SP; (**B**) MO vs. TU; (**C**) MO vs. CR; (**D**) SP vs. TU; (**E**) SP vs. CR; (**F**) TU vs. CR. Peaks in the positive or in the negative direction of the S-plots are coloured according to the absolute value of correlation (p(corr), from green= low values; to red = high values) and are influent in discriminating samples distributing in the positive or in the negative direction of the OPLS predictive component, respectively.

Fig. 6. Score plots of the first predictive (t[1]) and orthogonal (to[1]) components (left) and their respective S-line plots (right) for the pairwise OPLS-DA models separating the geographical origins of the finished anchovy samples. (**A**) TU vs. SP; (**B**) TU vs. CR; (**C**) CR vs. SP. Peaks in the positive or in the negative direction of the S-plots are coloured according to the absolute value of correlation (p(corr), from green= low values; to red = high values) and are influent in discriminating samples distributing in the positive or in the negative direction of the OPLS predictive component, respectively.