



# UNIVERSITÀ DI PARMA

## ARCHIVIO DELLA RICERCA

University of Parma Research Repository

Classification of transformed anchovy products based on the use of element patterns and decision trees to assess traceability and country of origin labelling

This is the peer reviewed version of the following article:

*Original*

Classification of transformed anchovy products based on the use of element patterns and decision trees to assess traceability and country of origin labelling / Varra, M. O.; Husakova, L.; Patocka, J.; Ghidini, S.; Zanardi, E.. - In: FOOD CHEMISTRY. - ISSN 0308-8146. - 360:(2021), pp. 129790-129790-129800. [10.1016/j.foodchem.2021.129790]

*Availability:*

This version is available at: 11381/2893558 since: 2021-06-03T08:51:23Z

*Publisher:*

Elsevier Ltd

*Published*

DOI:10.1016/j.foodchem.2021.129790

*Terms of use:*

Anyone can freely access the full text of works made available as "Open Access". Works made available

*Publisher copyright*

note finali coverpage

(Article begins on next page)

1 **Classification of transformed anchovy products based on the use of**  
2 **element patterns and decision trees to assess traceability and country**  
3 **of origin labelling**

4

5 *Maria Olga VARRÀ<sup>a</sup>, Lenka HUSÁKOVÁ<sup>b\*</sup>, Jan PATOČKA<sup>b</sup>, Sergio GHIDINI<sup>a</sup>,*

6

*Emanuela ZANARDI<sup>a\*</sup>*

7

8 *<sup>a</sup>Department of Food and Drug, University of Parma, Parma, Via del Taglio, 10, Parma 43126, Italy*

9 E-mail addresses: emanuela.zanardi@unipr.it (E. Zanardi); sergio.ghidini@unipr.it (S. Ghidini);

10 mariaolga.varra@studenti.unipr.it (M. O. Varrà).

11 *<sup>b</sup>Department of Analytical Chemistry, Faculty of Chemical Technology, University of Pardubice,*12 *Studentská 573 HB/D, Pardubice, CZ-532 10, Czech Republic*

13 E-mail address: jan.patočka@upce.cz (J. Patočka)

14

15 **\*CORRESPONDING AUTHORS:**16 **Lenka Husáková** (L. Husáková)

17 Department of Analytical Chemistry, Faculty of Chemical Technology, University of Pardubice

18 Studentská 573 HB/D, Pardubice, CZ-532 10 (Czech Republic)

19 Tel. +420 466 037 029; Fax: +420 466 037 068; E-mail address: lenka.husakova@upce.cz

20

21 **Emanuela Zanardi** (E. Zanardi)

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

23 Tel. +39.0521.902.760; E-mail address: emanuela.zanardi@unipr.it

24

25 **ABSTRACT**

26 Quadrupole inductively coupled plasma mass spectrometry (Q-ICP-MS) and direct mercury analysis were used  
27 to determine the elemental composition of 180 transformed (salt-ripened) anchovies from three different  
28 fishing areas before and after packaging. To this purpose, four decision trees-based algorithms, corresponding  
29 to C5.0, classification and regression trees (CART), chi-square automatic interaction detection (CHAID), and  
30 quick unbiased efficient statistical tree (QUEST) were applied to the elemental datasets to find the most  
31 accurate data mining procedure to achieve the ultimate goal of fish origin prediction. Classification rules  
32 generated by the trained CHAID model optimally identified unlabelled testing bulk anchovies (93.9% F-score)  
33 by using just 6 out of 52 elements (As, K, P, Cd, Li, and Sr). The finished packaged product was better modelled  
34 by the QUEST algorithm which recognised the origin of anchovies with F-score of 97.7%, considering the  
35 information carried out by 5 elements (B, As, K, Cd, and Pd). Results obtained suggested that the traceability  
36 system in the fishery sector may be supported by simplified machine learning techniques applied to a limited  
37 but effective number of inorganic predictors of origin.

38

39 **Abbreviations**

40 Certified reference materials, CRMs; classification and regression trees, CART; chi-square automatic  
41 interaction detection, CHAID; hierarchical cluster analysis, HCA; high energy Helium mode, HE He;  
42 inductively coupled plasma mass spectrometry, ICP-MS; inductively coupled plasma optical emission  
43 spectrometry ICP-OES; internal standard, ISTD; kinetic energy discrimination, KED; method detection limit,  
44 MDL; method limit of quantification, MLOQ; quick unbiased efficient statistical tree, QUEST.

45

46 **Keywords:** *Engraulis encrasicolus*; fish products; decision trees; geographical origin; data mining; ICP-MS.

47

## 48 **1. Introduction**

49 Foodstuff free-trade between nations all over the world, together with increasing diversification into food-  
50 related products, recently made the development of easy, rapid, cheap, and robust tools to assess traceability  
51 of foodstuffs to become a hot topic in the scientific community as well as in an industrial context.

52 The fishery sector is particularly prone to fraudulent practices but, on the other hand, it is insufficiently  
53 protected. The high complexity of the fish supply chain, the high number of stakeholders involved, and the  
54 fast perishability of fish, are a few of the many factors hampering the fight against fraud, which, in turn, reflect  
55 negatively on producers, transformers and final consumers from both economical and sanitary point of view  
56 (FAO, 2018; European Parliament Resolution 2013/2091(INI), 2014).

57 The perception of quality fresh or processed fish and seafood products by consumers is the sum of several  
58 different objective and subjective factors and it directly influences the global economic and market values of  
59 the product. At present, mislabelling or misrepresenting the origin of fish products keep getting encouraged by  
60 the so-called country-of-origin effect, according to which the consumers increasingly tend to associate high  
61 quality fish products with specific production areas because of specific sensorial characteristics, ethical or  
62 ecological motivations.

63 In this context, processed fish products deriving from the industrial transformation of the highly valuable  
64 European anchovy (*Engraulis encrasicolus*, L. 1758) are frequently subjected to fraud (Velasco, Aldrey,  
65 Pérez-Martín, & Sotelo, 2016).

66 European anchovy is a small pelagic fish that is mainly fished in the Mediterranean Sea and Black Sea, as well  
67 as in Eastern Central Africa (alongside the Moroccan coasts) and in Northeast Atlantic, especially in the  
68 Cantabrian Sea (FAO, 2020). In addition to the direct consumption as fresh fish, the product is frequently  
69 found in the European marketplace in the form of transformed, brine-fermented anchovy or filleted and canned  
70 (preserved in oil) anchovy (FAO, 2020).

71 The traditional anchovy transformation process by brine-ripening finds a long tradition in southern Europe.  
72 The fish, typically caught by purse seines, is quickly transported to the fish canning industry where it is  
73 beheaded, partially eviscerated and put into ripening containers (barrels), alternating layers of fish and salt.  
74 A pressure is then applied on the top layer to facilitate the progressive elimination of water. The fish is ripened  
75 until the desired degree of maturation is reached (from 3 up to 11-12 month on average) to then be moved from

76 the barrels. From that point on, the bulk ripened anchovies can be preserved and packaged in salt to be  
77 commercialised or further processed to obtain different products and preparations, for example by filleting and  
78 packaging-in-oil.

79 During the ripening, several chemical and physical modifications occur, including lipolysis, lipid oxidation,  
80 and proteolysis (Hernandez-Herrero, Roig-Sagués, López-Sabater, Rodriguez-Jerez, & Mora-Ventura, 1999;  
81 Czerner, Agustinelli, Guccione, & Yeannes, 2015). These modifications are of fundamental importance to  
82 prolong the shelf life and reduce the microbiological-associated risks and, at the same time, they influence the  
83 final organoleptic characteristics of the products (Besteiro, Rodríguez, Tilve-Jar, & Pascual López, 2000).

84 Salted anchovy from the Cantabrian Sea (Northern Spain) is worldwide appreciated as a high-quality product  
85 thanks to the sensorial characteristics of the raw fish, the strong link with the territory, and the long artisanal  
86 tradition behind its manufacturing (Laso et al., 2017). Taking into consideration the Cantabrian anchovy  
87 overall reputation and its high commercial value, it is therefore assumed to be object of fraud by substitution  
88 with fish from other sources. Therefore, developing methods that aim at providing concrete protection to the  
89 product is a matter of the utmost importance.

90 Up to now, the scientific research dealing with the identification of fish and seafood origin has been mainly  
91 focused on raw untransformed fish and seafood and has made use of different approaches. Among these,  
92 approaches based on the use of the inorganic components, such as stable isotopes (Carrera & Gallardo, 2017),  
93 mineral, trace- and/or ultra-trace elements (Smith & Watts, 2009), and a combination of stable isotopes and  
94 trace elements (Li, Han, Dong, & Boyd, 2019; Varrà, Ghidini, Zanardi, Badiani, & Ianieri, 2019) have been  
95 demonstrated to be successful strategies since offering several advantages depending on the reflection of  
96 seawater overall compositions on fish flesh.

97 Tracing back to the origin of processed or highly processed products is considerably difficult because of the  
98 manipulation and the addition of several compounds during preparation procedures. The use of salt during  
99 anchovy manufacturing may represent the most critical point since it can potentially mask the natural elemental  
100 content of fish. Nevertheless, the multiple identification of elements using techniques such as inductively  
101 coupled plasma-optical emission spectroscopy (ICP-OES) and inductively coupled plasma-mass spectrometry  
102 (ICP-MS) have been already successfully applied to identify the origin of transformed food products such  
103 processed tomato products (Lo Feudo, Naccarato, Sindona, & Tagarelli, 2010; Fragni, Trifirò, & Nucci, 2015),

104 fruit juices (Turra et al., 2017), wines (da Costa, Ximenez, Rodrigues, Barbosa, & Barbosa, 2020), dried beef  
105 (Franke et al., 2008), hams (Epova et al., 2018), and different types of cheese (Suhaj & Kore, 2008; Moreno-  
106 Rojas, Cámara-Martos, Sánchez-Segarra, & Amaro-López, 2012; Magdas et al., 2019). One application  
107 dealing with the use of multi-elemental analysis to authenticate seafood products is also available and it  
108 concerns the identification of caviar from different origins, which, as anchovy, is a salted product (Rodushkin  
109 et al., 2007).

110 The success of most of these applications was anyway strictly dependent on the support provided by  
111 chemometrics and machine learning methods for the identification of those elemental patterns echoing the  
112 original environment.

113 In this study, four decision tree algorithms corresponding to C5.0, classification and regression trees (CART),  
114 chi-square automatic interaction detection (CHAID), and quick unbiased efficient statistical tree (QUEST)  
115 were applied to elemental content of transformed anchovy products to predict the country of origin of the raw  
116 materials and support traceability of the products before and after packaging. The reasons behind the selection  
117 of decision trees as the basis method of this study are since decision tree models are easily understandable and  
118 interpretable, quick to build and, in general, low training times are associated with their use. Moreover, these  
119 techniques have high prediction accuracy in many fields so that makes them preferable and trustable choices  
120 for this kind of task.

## 121 **2. Materials and Methods**

### 122 *2.1. Chemicals, standards, and reference materials*

123 All the aqueous solutions employed for analyses were prepared using ultrapure water ( $0.05 \mu\text{S cm}^{-1}$ ) obtained  
124 by the Milli-Q<sup>®</sup> water purification system (Millipore, Bedford, USA).

125 For microwave digestion, hydrogen peroxide ( $\text{H}_2\text{O}_2$ ,  $\geq 30\%$  w/w) for ultra-trace analysis (Fluka Chemie AG,  
126 Buchs, Switzerland) and sub-boiled nitric acid prepared from nitric acid (65%, w/w, Selectipur quality, Lach-  
127 Ner, Neratovice, Czech Republic) by means of the sub-boiling distillation apparatus Distillacid<sup>™</sup> BSB-939-  
128 IR (Berghof, Eningen, Germany) were used.

129 The working calibration solutions for ICP-MS analysis were prepared daily using the multi-element stock  
130 solutions “A”, “B1”, “B2”, and “C”. Stock solution “A” ( $10 \text{ mg L}^{-1}$  of Li, B, Al, V, Cr, Fe, Mn, Ni, Cu, Zn,

131 Co, As, Se, Rb, Sr, Zr, Mo, Ru, Pd, Cd, Sn, Sb, Cs, Ba, Hf, Re, Pt, Tl, Pb, Bi, and Th) was prepared from the  
132 Supelco ICP multi-element standard solution IV (Merck, Darmstadt, Germany) (containing 1 g L<sup>-1</sup> of Li, B,  
133 Al, Cr, Fe, Mn, Ni, Cu, Zn, Co, Sr, Cd, Ba, Tl, Pb, and Bi) and single element standards (V, As, Se, Rb, Zr,  
134 Mo, Ru, Pd, Sn, Sb, Cs, Hf, Re, Pt, and Th) of concentration 1 ± 0.002 g L<sup>-1</sup> (Analytika Ltd., Prague, Czech  
135 Republic or SCP Science, Montreal, Canada).

136 Stock solutions “B1” (1 mg L<sup>-1</sup> of La, Ce, Pr, Nd, and U) and “B2” (0.20 mg L<sup>-1</sup> of Y, Tb, Ho, Yb, Sm, Eu,  
137 Gd, Er, Tm, Lu, and Dy) were prepared from the stock solution of rare earth elements Astatol mix “M008”  
138 (Analytika Ltd., Prague, Czech Republic). Stock solution “C” (50 mg L<sup>-1</sup> of Na, Mg, P, K, Ca, Mn, Cu, and  
139 Zn) was prepared from single element standards of 1 g L<sup>-1</sup> (Analytika Ltd., Prague, Czech Republic).

140 A 1 g L<sup>-1</sup> stock solution of Rh (SCP Science, Montreal, Canada) was used to prepare the internal standard  
141 solution (ISTD) at concentration of 200 µg L<sup>-1</sup>.

142 A 10 g L<sup>-1</sup> stock solution prepared from urea (TraceSelect quality, Fluka Chemie AG, Buchs, Switzerland) was  
143 used to prepare carbon reference solutions.

144 The element quantification accuracy was evaluated using the following certified reference materials (CRMs):  
145 NIST SRM 1577 Bovine Liver (National Institute of Science and Technology, NIST, Gaithersburg,  
146 MD, USA); NIST SRM 1566 Oyster Tissue (NIST, Gaithersburg, MD, USA); BCR<sup>®</sup> certified reference  
147 material (CRM)184 Bovine muscle (Institute for Reference Materials and Measurements, IRMM, Geel,  
148 Belgium); BCR<sup>®</sup> 185 Bovine Liver (IRMM, Geel, Belgium); CRM NCS ZC73015 Milk Powder (National  
149 Research Centre for Certified Reference Materials, NRCRM, Beijing, China); P-WBF CRM 12-2-04 Essential  
150 and Toxic Elements in Wheat Bread Flour (pb-anal, Kosice, Slovakia); CRM12-2-03 P-Alfalfa Essential and  
151 toxic elements in Lucerne (pb-anal, Kosice, Slovakia); SMU CRM 12-02-01 Bovine liver (pb-anal, Kosice,  
152 Slovakia).

## 153 2.2. Anchovy sampling and processing

154 Salt-ripened anchovies as bulk product (semi-finished, non-packaged) and as packaged (finished) product were  
155 obtained from the processing of European anchovy (*Engraulis encrasicolus* L.) and provided by the same fish  
156 preserves company.

157 A total of 90 bulk specimens were randomly collected from different ripening barrels' batches after maturation  
158 and suddenly vacuum-packaged into plastic bags. Similarly, a total of 90 finished specimens were obtained  
159 after salt-packaging of bulk anchovies and provided packaged into glass jars.

160 Both types of products were prepared from salting process (using not iodised sea salt of the same origin) of  
161 raw fish caught in the following geographical areas: Cantabrian Sea (Spain, FAO fishing area 27.8, n=30),  
162 upper Central Mediterranean Sea (Croatia, FAO fishing area 37.2.1, n=30) and lower Central Mediterranean  
163 Sea, (Tunisia, FAO fishing area 37.2.2, n=30).

164 Detailed information on the sampling and characteristics of the transformed fish used in the present study is  
165 reported in Table S1 (Supplementary Materials).

166 Before analysis, each individual fish was carefully cleansed with filter paper to remove external salt and  
167 manually peeled, eviscerated, and deboned, and finally minced with a ceramic knife. After that, samples were  
168 individually stored into glass vials and frozen at -20 °C.

#### 169 *2.2.1 Lyophilisation process*

170 Around 3.5 g of each trimmed fish sample were transferred into 5mL lyophilisation vials (borosilicate glass  
171 Vacule® equipped with 3-leg stopper, Wheaton, USA) wherein the material was dried. Before the freeze-  
172 drying process, the samples were deep-frozen at -80 °C for 24 hours to provide a necessary conditioning for  
173 low temperature drying. CoolSafe 4-15 L benchtop freeze dryer (LaboGene, Lyngø, Denmark) was employed  
174 for the lyophilisation of samples, with the CoolSafe condenser working temperature held at -110°C and a total  
175 chamber pressure of 3 hPa. The freeze-dried samples were subsequently homogenised directly inside the glass  
176 vials using a plastic rod to obtain a fine powder.

#### 177 *2.2.2 Microwave-assisted acid digestion*

178 For subsequent ICP-MS analysis, 0.1 g of freeze-dried samples or CRMs were weighted (in triplicate) into a  
179 10 mL perfluoroalkoxy (PFA) tube and 4 mL of 16% HNO<sub>3</sub> (65%, w/w HNO<sub>3</sub>, 1:3 diluted) and 1 mL of 30%  
180 H<sub>2</sub>O<sub>2</sub> were added. Three PFA tubes were placed into DAC-100S polytetrafluoroethylene vessels (Berghof,  
181 Eningen, Germany) previously filled with 25 mL of HNO<sub>3</sub> (16%, v/v), by ensuring that the level of liquid in  
182 the outer polytetrafluoroethylene vessel was higher than those in the PFA tubes. This way, the vapor pressures  
183 were compensated and the evaporation of the solution from the PFA tubes was avoided (Husáková et al., 2015).



184 Samples were decomposed using a Berghof Speedwave™ MWS-3+ microwave digestion system (Berghof,  
185 Germany) with the maximum total output of the microwave generator (1450 W) via the following multistep  
186 program: step 1, 20 min at 180 °C (ramp 5 min); step 2, 20 min at 220 °C (ramp 5 min); steps 3, 5 min at  
187 100 °C (ramp 5 min).

188 The clear digested samples were diluted with deionised water up to 25 mL and the residual carbon content  
189 quantified at  $5.58 \pm 0.12\%$  by ICP-OES, following the method previously reported by Husáková et al. (2011).

### 190 2.3. Mercury analysis

191 Total Hg content was determined directly on lyophilised solid samples or CRMs using a single-purpose atomic  
192 absorption spectrometer AMA 254 (Altec Ltd., Prague, Czech Republic).

193 Analytical operation conditions as follows: sample mass, 50 mg; drying step, 60 s at 120 °C; decomposition  
194 step, 150 s at 750 °C; Hg release step, 45 s at 900 °C; reading step, 60 s monitoring the 253.6 nm absorbance  
195 peak. The flow rate of oxygen (99.5%) carrier gas was  $170 \text{ mL min}^{-1}$ .

### 196 2.4. ICP-MS analysis

197 Element quantification in samples was performed by using an Agilent 7900 quadrupole ICP-MS apparatus  
198 (Agilent Technologies, Inc., Santa Clara, CA, USA) equipped with a quartz concentric nebulizer MicroMist  
199 ( $400 \mu\text{L min}^{-1}$ ), the Peltier-cooled (2 °C) Scott quartz spray chamber, quartz torch with 2.5 mm internal  
200 diameter injector, standard sampling and skimmer nickel cones with orifices of 1 and 0.45 mm, and an octopole  
201 collision/reaction cell for interference removal using kinetic energy discrimination. A low-pulsation, 10-roller  
202 peristaltic pump with three separate channels was employed to precisely deliver both samples and ISTD. For  
203 the simultaneous ISTD aspiration and mixing with the sample the connecting tubing, connectors, and the “Y”  
204 piece (included into internal standard kit) were employed. Analytical conditions were enhanced before starting  
205 sample measurement by using the multi-elemental tuning solution (Agilent Technologies, Inc., Santa Clara,  
206 CA, USA) containing  $1 \mu\text{g L}^{-1}$  of Ce, Co, Li, Mg, Tl and Y, to obtain the highest possible sensitivity for  
207 elements of low, middle, and high m/z. Using the typical operating conditions summarised in Table S2  
208 (Supplementary Materials), a sensitivity of 6000 counts  $\text{s}^{-1}$  per  $\mu\text{g L}^{-1}$  and a resolution of 0.64 amu peak width

209 (full width at half maximum intensity) were achieved for  ${}^7\text{Li}^+$ . The same parameters were 50000 counts  $\text{s}^{-1}$  per  
210  $\mu\text{g L}^{-1}$  and 0.62 for  ${}^{89}\text{Y}^+$ , and 30000 counts  $\text{s}^{-1}$  per  $\mu\text{g L}^{-1}$  and 0.60 for  ${}^{205}\text{Tl}^+$ .

211 While the quantification of certain elements was performed without pressurising the collision cell (i.e., “no-  
212 gas” mode), “Helium” mode (He) and “High Energy Helium” mode (HE He) were instead used for the  
213 quantification of problematic elements mostly suffering from polyatomic interferences.

214 The acquisition mode for elements as follows:  ${}^7\text{Li}$ ,  ${}^{11}\text{B}$ ,  ${}^{24}\text{Mg}$ ,  ${}^{66}\text{Zn}$ ,  ${}^{85}\text{Rb}$ ,  ${}^{88}\text{Sr}$ ,  ${}^{89}\text{Y}$ ,  ${}^{90}\text{Zr}$ ,  ${}^{95}\text{Mo}$ ,  ${}^{101}\text{Ru}$ ,  ${}^{103}\text{Rh}$ ,  
215  ${}^{105}\text{Pd}$ ,  ${}^{111}\text{Cd}$ ,  ${}^{118}\text{Sn}$ ,  ${}^{121}\text{Sb}$ ,  ${}^{133}\text{Cs}$ ,  ${}^{138}\text{Ba}$ ,  ${}^{139}\text{La}$ ,  ${}^{140}\text{Ce}$ ,  ${}^{141}\text{Pr}$ ,  ${}^{146}\text{Nd}$ ,  ${}^{147}\text{Sm}$ ,  ${}^{153}\text{Eu}$ ,  ${}^{157}\text{Gd}$ ,  ${}^{159}\text{Tb}$ ,  ${}^{163}\text{Dy}$ ,  ${}^{165}\text{Ho}$ ,  ${}^{166}\text{Er}$ ,  
216  ${}^{172}\text{Yb}$ ,  ${}^{175}\text{Lu}$ ,  ${}^{178}\text{Hf}$ ,  ${}^{185}\text{Re}$ ,  ${}^{195}\text{Pt}$ ,  ${}^{205}\text{Tl}$ ,  ${}^{206+207+208}\text{Pb}$ ,  ${}^{209}\text{Bi}$ ,  ${}^{232}\text{Th}$ ,  ${}^{238}\text{U}$ , all by “No gas mode”;  ${}^{23}\text{Na}$ ,  ${}^{27}\text{Al}$ ,  ${}^{39}\text{K}$ ,  
217  ${}^{44}\text{Ca}$ ,  ${}^{51}\text{V}$ ,  ${}^{52}\text{Cr}$ ,  ${}^{55}\text{Mn}$ ,  ${}^{56}\text{Fe}$ ,  ${}^{59}\text{Co}$ ,  ${}^{60}\text{Ni}$ ,  ${}^{63}\text{Cu}$ ,  ${}^{103}\text{Rh}$ , all by “He mode”;  ${}^{31}\text{P}$ ,  ${}^{75}\text{As}$ ,  ${}^{78}\text{Se}$ ,  ${}^{103}\text{Rh}$  all by “HE He  
218 mode”.

219 The calibration curves for the quantification of 51 elements ( $R^2 > 0.999$ ) resulted from the acquisition of  
220 working calibration solutions prepared from multi-element stock solutions “A”, “B1”, “B2”, and “C” described  
221 in Section 2.1. The concentration of elements for calibration were as follows: blank, 1, 5, 10, 50, 100  $\mu\text{g L}^{-1}$  of  
222 Li, Be, B, Al, V, Cr, Fe, Mn, Ni, Cu, Zn, Co, Ga, Ge, As, Se, Rb, Sr, Zr, Mo, Ru, Cd, In, Sn, Sb, Te, Cs, Ba,  
223 Hf, Ta, Re, Pt, Tl, Pb, Bi, Th; 0.1, 0.5, 1, 5, 10  $\mu\text{g L}^{-1}$  of La, Ce, Pr, Nd, U; 0.02, 0.1, 0.2, 1, 2  $\mu\text{g L}^{-1}$  of Y, Sm,  
224 Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu; 0.5, 1, 5, 10  $\text{mg L}^{-1}$  of Na, Mg, P, K, Ca, Mn, Cu, Zn.

225 Samples were measured in triplicate. Standards and blanks were analysed after a single initial calibration.  
226 Continuing calibration blanks and calibration verification standards were automatically run after every 25  
227 samples. To compensate possible instrumental drift and matrix effects, a 200  $\mu\text{g L}^{-1}$  Rh ISTD was  
228 simultaneously aspirated and mixed with samples. The percent recovery of the ISTD responses for the entire 12-  
229 hour sequence normalised to the calibration blanks shows that there was no gradual loss of sensitivity over  
230 time, even when running high matrix samples. Long term stability measured by comparing ISTD responses  
231 from the beginning of the sequence to the end was better than 5 %.

## 232 2.5. Analytical validation

233 The accuracy of the methodology was checked by analysing the five certified reference materials listed in  
234 Section 2.1 (NIST 1566 Oyster Tissue, BCR CRM 184 Bovine muscle, BCR CRM 185 Bovine Liver, CRM  
235 12-2-01 Bovine Liver, NIST SRM 1577 Bovine Liver) intended for the evaluation of analytical methods and

236 instruments and used for the determination of the mass fraction values of selected elements in marine tissue,  
237 foods, or similar materials. In addition, CRM 12-2-03 Essential and toxic elements in Lucerne, NCS ZC 73015  
238 Milk Powder, and CRM 12-2-04 Wheat bread flour, were analysed to assess accuracy in determining  
239 lanthanides and actinides. The high level of agreement between target and found values demonstrated trueness  
240 of data obtained (Supplementary Materials, Table S3).

241 Intra-day and inter-day precisions were calculated to assess the overall precision of the method and were  
242 determined by analysing single CRMs three times during the same day and during three different days over  
243 one month, respectively. The method was found to be precise enough due to the percent relative standard  
244 deviations (RDS %) of intra-day and inter-day precision which were mostly below 14 % (Supplementary  
245 Materials, Table S3).

246 The 3 sigma method detection limits (MDL) and the method limits of quantification (MLOQ) reported in Table  
247 S4 (Supplementary Materials), were experimentally determined through the measurement of ten replicates of  
248 a blank sample and by calculating triple and tenfold standard deviations divided by the slope of the calibration  
249 curve, respectively. In all cases, detection limits were found significantly below the typical requirements for  
250 this analysis so that selected elements could be determined at the background level. Table S4 (Supplementary  
251 Materials) also summarises relative sensitivities of Q-ICP-MS for analysis of individual elements with the use  
252 of Rh ISTD.

## 253 2.6. Data elaboration

254 Statistics was applied to elemental concentrations referring to anchovy dry matter (d.m.) and carried out using  
255 IBM SPSS software (v. 23.0, SPSS Inc., Chicago, IL, USA), SIMCA software (v. 16.0.2, Sartorius  
256 Stedim Data Analytics AB, Sweden), NCSS 2020 software (v. 20.0.3, NCSS LLC., Kaysville, UT, USA), and  
257 IBM SPSS Modeler software (v. 18.2, SPSS Inc., Chicago, IL, USA).

258 Univariate data analysis, consisting of nonparametric Mann-Whitney test ( $p \leq 0.05$ ) and Kruskal–Wallis test  
259 plus Dunn’s post hoc test ( $p \leq 0.05$ ) was applied to the whole data of both the set of bulk anchovies (30 samples  
260  $\times$  3 replicates  $\times$  3 provenances) and the set of packaged anchovies (30 samples  $\times$  3 replicates  $\times$  3 provenances)  
261 to investigate any significant difference between groups of samples. Nonparametric tests were chosen instead

262 of classical parametric ones because most of the data presented non normal distribution and heteroscedasticity,  
263 as verified though the application of Shapiro-Wilk's and Levene's tests, respectively ( $p \leq 0.05$ ).

264 As a classical unsupervised chemometric method, hierarchical cluster analysis (HCA) using factorial  
265 coordinates of principal component analysis was applied to the two anchovy datasets in order to scout data  
266 structures and identify group of samples by the similarity of their variables (Drab & Daszykowski, 2014).

267 To create classification models with good validity and consistency, the holdout technique (stratified randomly  
268 sampling) was adopted when machine learning was applied. The bulk and the packaged anchovy datasets were  
269 organised into three different data matrices each in a 70:15:15 partition ratio to create training, validation, and  
270 testing and sets, respectively. The training sets were used to estimate the models, the validation set to test and  
271 select the best models and the testing sets to confirm the reliability of the selected models.

272 Different learning algorithms relying on the principle of decision trees and which do not require assumptions  
273 about data distribution were chosen to learn how the data of the training set were classifiable according to  
274 origin of samples and to create proper prediction models. These were C5.0, CART, CHAID, and QUEST.

275 Briefly, decision trees work on the division/classification of samples driven by the values of the variables under  
276 examination, by creating subsets of samples which are progressively split across the structure of the tree,  
277 independently of the distribution of the predictor. Therefore, decision trees are nonparametric machine learning  
278 algorithms specifically seeking for data partitioning into response-homogeneous zones (Kotsiantis, 2013;  
279 Barbosa, Nacano, Freitas, Batista, & Barbosa, 2014).

280 The output of the method is a set of concatenated classification rules in the form of a decision tree composed  
281 by nodes (identifying the features that need to be sorted) and branches (identifying the values assumed by  
282 nodes) (Han, Kamber, & Pei, 2011). CART, C5.0, CHAID, and QUEST represent different methods by which  
283 the architecture of the decision trees can be built up and differ each other mainly for the segmentation rules  
284 applied and the tree optimisation method. The C5.0 algorithm is based on binary splitting and works by  
285 choosing progressively the instances that allow to gain the maximum partitioning information and stopping  
286 via the pruning rule, i.e., by removing from the splits which do not add significant information. CART  
287 algorithm is also based on binary splitting but differs from C5.0 essentially for a different stopping rule in the  
288 creation of the tree, consisting of the evaluation of the purity of the node, i.e. the maximum degree of  
289 homogeneity between categories. The CHAID algorithm is a multiway splitting system based on Chi-square

290 statistics to decide for tree ramifications and is based on the measures of the impurity of the nodes. Finally, the  
291 QUEST is a binary-split algorithm which, in the case of continuous variables, uses an ANOVA F-Test to create  
292 tree nodes. Further information can be retrieved from Han, Kamber, & Pei (2011) and from Rokach & Maimon  
293 (2008).

294 In the present work, CART was built using the Gini Impurity Index to determine the nodes impurity and select  
295 input variables. For CART, CHAID, and QUEST a maximum of five tree levels was set to avoid excessive  
296 splitting. Building settings for CHAID included the use of Pearson's Chi-square statistics and a Bonferroni  
297 adjustment to calculate the adjusted  $p$ -values. Significant level for splitting for both CHAID and QUEST was  
298 set at 95%.

299 Training models were compared each other by analysing classical metrics in multivariate classification  
300 methods, corresponding to accuracy (%), sensitivity (%), specificity (%), precision (%), and F-score (%)  
301 indexes, calculated from the unlabelled test set. Calculation were performed according to Cuadros-Rodríguez,  
302 Pérez-Castaño, & Ruiz-Samblás (2016).

### 303 **3. Results and discussion**

#### 304 *3.1. Initial data evaluation*

##### 305 3.1.1 Global elemental profiles of transformed anchovy products

306 The anchovy samples distributions according to the measured elemental concentrations which varied  
307 significantly in relation to the three investigated geographical provenances are shown Fig. 1. The beeswarm  
308 boxplots reported revealed that, according to Kruskal-Wallis and Dunn's multiple comparison test,  
309 concentrations of B, V, As, and Hg were different ( $p < 0.05$ ) among Cantabrian, Tunisian and Croatian bulk  
310 anchovies (Fig. 1A). Packaged anchovies, indeed, differed for the same element concentrations plus those of  
311 Li (Fig. 1B). Regardless the type of product, the highest amounts of B and V were found in Tunisian samples  
312 and those of As and Hg in Croatian samples (Fig. 1A, Fig. 1B).

313 Complete data matrices reporting the whole concentrations of the 52 elements measured in bulk and packaged  
314 anchovies can be found in Table S5 and Table S6 (Supplementary Materials), respectively. The most abundant  
315 element was found to be Na (whose median concentrations ranged from 142 mg kg<sup>-1</sup> in bulk Croatian

316 anchovies to 177 mg kg<sup>-1</sup> in Tunisian packaged anchovies) followed by P, K, Ca, and Mg. Bulk and packaged  
317 products from Cantabrian Sea differed from samples of Mediterranean origin (Croatian and Tunisian) because  
318 of significantly higher concentrations of P and K. At the same time, no differences in Na and Ca contents  
319 between Cantabrian and Croatian anchovies was encountered, which, instead, were both significantly lower  
320 compared to those found in samples originating from Tunisia (Table S5, Table S6, Supplementary Materials).  
321 Similarly, some minor, trace- and ultra-trace elements showed significant variations in relation to the  
322 provenance Besides, Cantabrian anchovies were characterised by higher Ni, Mo, Cs, and Tl concentrations  
323 and lower Al concentrations than those encountered in the two Mediterranean products (Supplementary  
324 Tables S5, S6).

325 Considering the peculiarity of the products investigated, no published works dealing with the same food  
326 matrices and purposes were found, except for similar canned products including, however, other ingredients  
327 (Ikem & Egiebor, 2005). For this reason, a more in-depth analysis of the elemental concentrations was not  
328 possible.

329 Beyond statistical differences related to the geographical origin, a high degree of variations of elemental  
330 concentrations within fish of the same provenance was however found. Considering equal characteristics of  
331 marine environment for each group, this variation is likely to be attributable to the natural biological diversity  
332 among individuals. Moreover, the technology behind the processing of anchovies does suggest that all  
333 manufacturing operations (including handling, treatment, production, and distribution) between the time of  
334 fishing and the end-product stage can impact the final elemental profile of anchovies. Therefore, potential  
335 markers of geographical origin need to go through a more-in-depth evaluation to be correctly identified,  
336 thereby preventing misinterpretations.

### 337 3.1.2 Pre-selection of elements as potential indicators of origin

338 The use of salt in the form of saturated brine during the processing and packaging of ripened anchovies is one  
339 of the main complicating factors which might limit the proper identification of markers of geographical origin  
340 since it inevitably modifies the natural element profile of the product. For instance, the use of a brine for  
341 fermentation purposes was reported to decrease the total amount of certain trace elements in transformed fish

342 roe because of partitioning phenomena between the solid and the liquid phases occurring during fermentation  
343 (Bekhit, Morton, & Dawson, 2008).

344 Despite the addition of salt as an exogenous source of contaminants, the possibility of using the elemental  
345 profile to discriminate the origin of processed salted cheeses has been previously investigated and it was found  
346 that some elements as Sr, Li, Mg, Rb and K (Magdas et al., 2019) as well as Tl, Li (Epova et al., 2018) can  
347 still be used as powerful tracers since retain a strong link with the place of origin of milk. In addition,  
348 concentrations of As, Ba, Br, I, Mo, and Se, were proven to be stable although processing and were identified  
349 as useful markers for the geographic origin of caviar (Rodushkin et al., 2007) as well as to distinguish caviar  
350 obtained from wild sturgeon (Depeters, Puschner, Taylor, & Rodzen, 2013). On the contrary, the concentration  
351 of Fe, Al, Ti, V were found to be heavily affected by handling and packaging operations and, therefore, useless  
352 for authentication of transformed products (Rodushkin et al., 2007).

353 In the present work, the elemental profile of the bulk anchovies of each origin was compared to that of the  
354 finished packaged products, to highlight possible modifications occurring during the end stages of anchovies  
355 processing. Results from Mann-Whitney test (two-tailed, confidence level 95%) highlighted the concentrations  
356 of following elements to be significantly different between the two types of products at least in two out of three  
357 origin groups: Na, Mg, Ca, Cu, Cr, V, Ru, Rb, La, Ce, Pr, Gd, Re and Tl (see Supplementary Materials,  
358 Table S7).

359 Considering that the main aim of the present research was to create machine learning based models as robust  
360 as possible in classifying samples by origin, highly variable elements identified both in the present study (Na,  
361 Mg, Ca, Cu, Cr, V, Ru, Rb, La, Ce, Pr, Gd, Re and Tl), and retrieved from the literature (Fe, Ti, and Al) were  
362 excluded and a total of 35 input variables, i.e. elements, were retained for subsequent classification analyses.

### 363 3.1.3. Hierarchical cluster analysis (HCA)

364 The inner potential of the elemental profile in guiding the creation of groups of anchovies was at first instance  
365 explored. Any possible natural difference or similarity among anchovies of both datasets was therefore  
366 uncovered by the application of HCA. Since different methods to measure distances (i.e. Euclidean and  
367 Manhattan distances) and to perform grouping (i.e. Ward's minimum variance, single linkage, complete  
368 linkage, simple average group average, median, and centroid clustering methods) are applicable for HCA

369 building, the cophenetic correlation coefficient (CCC) was employed as an index to compare the methods with  
370 each other and to evaluate the validity of the resulting dendrograms (Sokal & Rohlf, 1962; Saraçlı, Doğan, &  
371 Doğan, 2013). Further details about the tested methods can be retrieved from Everitt, Landau, Leese, & Stahl  
372 (2011).

373 By evaluating the CCC reported in Table S8 (Supplementary Materials) and taking into consideration that the  
374 closer is this index to 1, the higher is the degree of fit of clustering (Saraçlı et al., 2013) it was found that the  
375 use of the Euclidean inter-point distance and the Ward's aggregation method was the most performant HCA-  
376 based method both for bulk and packaged anchovies, with CCC values of 0.9888 and 0.9836, respectively (See  
377 Table S8, Supplementary Materials). Dendrograms resulting from the proposed methodology are shown in  
378 Fig. 2. As it can be observed, bulk anchovies (Fig. 2A) and packaged anchovies (Fig. 2B) were gathered into  
379 three major clusters at dissimilarity values of 65 % and 70 %, respectively. These three clusters mostly  
380 enclosed samples of the same origin, thus suggesting the existence of elemental patterns strong enough to  
381 reflect on the presence of geographical origin-driven groupings. Nevertheless, a few bulk samples from Croatia  
382 drifted apart from the others (Fig. 2A), as well as some samples from Croatia and Tunisia were mingled  
383 together in the second cluster of the dendrogram of packaged anchovies (Fig. 2B). Thus, Cantabrian products  
384 were in general better clustered than samples from Croatia and Tunisia which, on the contrary, were more  
385 connected each other. This is easily explained by the fact that the two fishing areas of anchovy from Croatia  
386 and Tunisia are neighbouring zones of the Mediterranean Sea (FAO fishing area 37.2.1 and 37.2.2,  
387 respectively). Therefore, these samples may share lots of compositional features linked to the similar  
388 environmental characteristics.

### 389 *3.2 Data mining for the geographic origin evaluation*

390 Considering that the main task of the present research was to verify the usefulness of the elemental profile of  
391 anchovy products to the development of rapid methods for geographical origin verification before and after  
392 the product packaging, different machine learning algorithms were explored to identify the best-suited one to  
393 this purpose.

394 Four decision trees algorithms (C5.0, CART, CHAID, and QUEST) trained on **70 % of the whole sample sets**  
395 **(270 samples per each anchovy set)** were examined, in search of the most accurate models in identifying the



396 origin labels of bulk and packaged anchovy samples in the validation set. One of the main advantages of using  
397 decision tree is that the selection of the most important variables for classification is performed automatically  
398 during training stage. Therefore, computation time is reduced, while interpretability, accessibility and  
399 handiness of the models is improved.

400 Performance outcomes of the trained and validated classification models are summarised in Table S9  
401 (Supplementary Materials). The rate of classification accuracy of the samples of the training sets ranged  
402 between 89.9 % and 99.4 %, with better results shown up by C5.0 both for the bulk and the packaged anchovy  
403 dataset. When validation of the models was performed, 98.3 % of bulk samples was correctly classified by  
404 CHAID. As for the packaged products, the most accurate model for classifying anchovies of the validation set  
405 was found to be QUEST (96.0% accuracy) (Table S9, Supplementary Materials). Based on the accuracy  
406 outcomes obtained during the validation phase, CHAID and QUEST were selected as the most appropriate  
407 algorithms to classify the origin of bulk and packaged anchovy samples, respectively. A short summary of the  
408 outputs obtained by the application of the other algorithms is however reported in Supplementary Material  
409 (Supplementary Tables S10, S11).

410 By looking at the ranks of each predictive element selected by the four models (Fig. 3) it is possible to highlight  
411 that C5.0 and CHAID models extracted a lower number of attributes compared to CART and QUEST models.  
412 Li, B, P, K, As, Sr, Zr, Pd, Cd, Cs, and Ba were shared as predictors within the two anchovy datasets. By  
413 contrast, Sb and Pb were influent only for bulk products (Fig. 3A), while Ni e U were extracted only for  
414 packaged products. Interestingly, As emerged as the variable showing the highest impact for all the  
415 classification models of bulk anchovies (Fig. 3A) and for QUEST and CHAID models of packaged anchovies  
416 (Fig. 3B). B and Cs were instead found to be the most important attributes in the CART and C5.0 models of  
417 packaged anchovies, respectively (Fig. 3B).

418 Arsenic contamination of seawaters can be related to anthropogenic pollutant activities as well as to the natural  
419 geological characteristics of the area (Garellick, Jones, Dybowska, & Valsami-Jones, 2008). As an example,  
420 As (together with Cr, Cu, Hg, Mn, Ni, Pb, Se, and V) concentrations were reported to be higher in seawater  
421 where volcanic activities exist such as the Mediterranean Basin (Juncos et al., 2016; Zkeri, Aloupi, & Gaganis,  
422 2018). Moreover, the uptake of As by fish is influenced by several natural factors including water temperature  
423 and salinity, cooccurrence of phosphate, and seasonal differences of the distribution of the inorganic and

424 organic forms of As in the aquatic environments (Ferrante et al., 2019). Regardless the natural or anthropogenic  
425 nature of As and releasing sources, the reduced exchange of water in the Mediterranean Basin and, especially  
426 in the Adriatic Sea, can facilitate the accumulation of As in the environment (Ferrante et al., 2019). This can  
427 justify the higher amounts of As in anchovy from the Mediterranean Sea compared to Cantabrian (Atlantic  
428 Ocean) anchovy reported in the present work (see Fig. 1A, Fig. 1B). Moreover, in pelagic fish species from  
429 the Adriatic Sea higher amounts of As compared to other sampling zones was previously shown (Storelli &  
430 Marcotrigiano, 2004).

### 431 3.2.1 Decision tree by CHAID algorithm for origin authenticity of bulk anchovies

432 In accordance with the optimal accuracy results achieved in training and validation, CHAID model was found  
433 to be characterised by optimal accuracy (94.1 %), sensitivity (95.6 %), and specificity (97.4 %) values also  
434 when used to classify unlabelled samples of the bulk anchovy test set (Table 1). Therefore, the method used  
435 can be effectively considered powerful enough when the analytical goal is the identification of Tunisian,  
436 Cantabrian, and Croatian bulk anchovy products origins.

437 The architecture of the decision tree obtained is illustrated in Fig. 4. As it can be observed, the tree was a **three-**  
438 level structure, with a total of 19 decision nodes and 12 classification rules created by using 6 elements only.  
439 The decision rules generated from the root node were based on 5 concentration ranges of As, which was  
440 confirmed to be the most influent element for first sample discrimination by CHAID (see Fig. 3A). CHAID-  
441 decision trees are generally more complex than those generated by other technique since it relies on a multiway  
442 splitting principle, but the higher degree of segmentation can help reducing the tree depth and speed up the  
443 classification of samples.

444 Concentrations of As  $\leq 3.38$  mg kg<sup>-1</sup> classified Cantabrian anchovies just at level 1 with 100% probability. In  
445 general, decreasing As concentrations (from 8.16 mg kg<sup>-1</sup> downwards) together with increasing concentrations  
446 of K (from 4084 mg kg<sup>-1</sup> upwards) and P (from 5078 mg kg<sup>-1</sup> upwards) were associated to the highest  
447 probability of identifying Cantabrian samples. Tunisian samples were better classified by descending As  
448 concentration ranges coupled with higher Li ( $> 0.16$  mg kg<sup>-1</sup>), Sr ( $>29.75$  mg kg<sup>-1</sup>), or Cd amounts ( $> 0.07$   
449 mg kg<sup>-1</sup>). Finally, when P, Li, Cd got lower the occurrence of Croatian samples become more probable.

### 450 3.2.2 Decision tree by QUEST algorithm for origin authenticity of packaged anchovies

451 The QUEST-based decision trees applied to packaged anchovy was composed by 13 decisions nodes stratified  
452 into **four** levels. B was selected as the first binary splitting variable. The outcomes related to predictor  
453 importance reported in Fig. 3B (according to which B had the highest influence in prediction) were confirmed  
454 by analysing the splitting variables used to generate the QUEST decision tree, where B was just selected as  
455 the first binary splitting variable (Fig. 5). B value higher than 5.13 mg kg<sup>-1</sup> generated a leaf (final) node with  
456 91.3% probability of predicting samples originating from Tunisia. Globally, 100% probability of correctly  
457 identifying Croatian samples was reached at the last tree-level using the classification rule based on B, As, and  
458 Cd or the classification rule based on B, As, K, and Pd. With increased B (> 5.13 mg kg<sup>-1</sup>), As (> 7.14 mg kg<sup>-1</sup>),  
459 and K (> 5669 mg kg<sup>-1</sup>) concentrations, also the likelihood of recognising Cantabrian anchovies increased.  
460 The set of decision rules established by QUEST further proved to be reliable and effective for the classification  
461 of packaged anchovy origin, owing to the good ability in predicting the unknown origin of samples of the test  
462 set. Compared to other decision tree algorithm, QUEST ranked first in terms of accuracy (97.7 %), sensitivity  
463 (97.6 %), specificity (98.9 %), and precision (98.0 %) (Table 1).  
464 Even though transformed anchovy implicitly represents a complex processed foodstuff, it is important to stress  
465 that using decision trees may be the quickest and the most intelligible way to solve problems related to  
466 classification of foodstuffs.

#### 467 **4. Conclusions**

468 In this work, data mining techniques were applied to transformed anchovy products to verify whether the origin  
469 of fish could be identified through the elemental patterns measured by ICP-MS and direct mercury analysis.  
470 Different machine learning algorithms relying on the principle of decision trees were applied to data and  
471 classification rules to distinguish anchovy fish of Cantabrian Sea from Tunisian and Croatian anchovies were  
472 created. Firstly, differences of elemental composition between anchovies at two stages of the production chain  
473 were investigated, to verify whether misleading elemental inclusion from the manufacturing environment was  
474 introduced. After having excluded problematic elements based on literature review and direct comparison of  
475 bulk and packaged anchovy profile and after having explored the effective presence of fish clusters related to  
476 origin, C5.0, CART, CHAID and QUEST decision trees were trained. This way, the selection of the most

477 important variables and the identification of cut-off limits for each element concentration to describe a specific  
478 group of samples were performed in tandem.

479 The results obtained showed that the concentrations of 6 elements only (As, K, P, Li, Cd, and Sr) are required  
480 to identify the origin of anchovy fish under the form of bulk products using the CHAID algorithm. Arsenic  
481 was found to be the first sorting element, whose contribution to geographical origin differentiation was  
482 remarkably reflected in the ability of the decision tree to identify the unknown label of bulk fish with accuracy,  
483 specificity, sensitivity, and precision values above 93 % on average. The origin of the packaged anchovy  
484 products for sale, was better recognised by the set of classification rules generated by the QUEST algorithm.  
485 In this case, 5 elements were sufficient to achieve accuracy, sensitivity, specificity, and precision outcomes  
486 higher than 96 %. The splitting of samples into groups was driven by the predictive influence of B, followed  
487 by As, K, Cd, and Pd.

488 In view of the above results, decision tree-based methods applied to elemental profiles of fishery products after  
489 industrial processing might be postulated as an immediate and easy-to-handly procedure to figure out how the  
490 elemental composition can help in solving many actual challenges related to fish authenticity and commercial  
491 fraud. Moreover, the cost-effectiveness of the methodology, reached by doing away with the irrelevant  
492 elements, may finally disconnect this kind of applications from the scientific research and lead to the  
493 application in the primary and secondary production sectors.

494 Future research including anchovy fish obtained from different countries and production systems is however  
495 desirable **not only** to clarify the involvement of multiple environmental factors on the stability over space and  
496 time of element profile of processed fish products, **but also for the creation and curation of databases storing**  
497 **and making available analytical data relating to the fish authenticity.**

## 498 **Appendix A. Supplementary Materials**

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

500 **Declaration of interest:** The authors declare that they have no known competing financial interests or  
501 personal relationships that could have appeared to influence the work reported in this paper.

502 **Data Availability:** The dataset generated during the current study is available from the corresponding author  
503 on reasonable request.

#### 504 **Acknowledgements**

505 The research was carried out within the project “Development of rapid tools for anchovies authentication”  
506 funded by Rizzoli Emanuelli S.p.a. (Parma, Italy). The authors also gratefully acknowledge the University of  
507 Pardubice for the financial support (project no. SGS\_2020\_002).

508 **References**

- 509 **Barbosa, R. M., Nacano, L. R., Freitas, R., Batista, B. L., & Barbosa Jr, F. (2014). The use of decision trees**  
510 **and naive Bayes algorithms and trace element patterns for controlling the authenticity of free- range-**  
511 **pastured hens' eggs. *Journal of food science*, 79(9), C1672-C1677. [https://doi.org/10.1111/1750-](https://doi.org/10.1111/1750-3841.12577)**  
512 **[3841.12577](https://doi.org/10.1111/1750-3841.12577)**
- 513 Bekhit, A. E. D. A., Morton, J. D., & Dawson, C. O. (2008). Effect of processing conditions on trace  
514 elements in fish roe from six commercial New Zealand fish species. *Journal of Agricultural and Food*  
515 *Chemistry*, 56(12), 4846–4853. <https://doi.org/10.1021/jf8005646>
- 516 Besteiro, I., Rodríguez, C. J., Tilve-Jar, C., & Pascual López, C. (2000). Selection of attributes for the  
517 sensory evaluation of anchovies during the ripening process. *Journal of Sensory Studies*, 15(1), 65–77.  
518 <https://doi.org/10.1111/j.1745-459X.2000.tb00410.x>
- 519 Carrera, M., & Gallardo, J. M. (2017). Determination of the Geographical Origin of All Commercial Hake  
520 Species by Stable Isotope Ratio (SIR) Analysis. *Journal of Agricultural and Food Chemistry*, 65(5),  
521 1070–1077. <https://doi.org/10.1021/acs.jafc.6b04972>
- 522 Cuadros-Rodríguez, L., Pérez-Castaño, E., & Ruiz-Samblás, C. (2016). Quality performance metrics in  
523 multivariate classification methods for qualitative analysis. *Trends in Analytical Chemistry*, 80, 612–  
524 624. <https://doi.org/10.1016/j.trac.2016.04.021>
- 525 Czerner, M., Agustinelli, S. P., Guccione, S., & Yeannes, M. I. (2015). Effect of different preservation  
526 processes on chemical composition and fatty acid profile of anchovy (*Engraulis anchoita*). *International*  
527 *Journal of Food Sciences and Nutrition*, 66(8), 887–894.  
528 <https://doi.org/10.3109/09637486.2015.1110687>
- 529 Da Costa, N. L., Ximenez, J. P. B., Rodrigues, J. L., Barbosa, F., & Barbosa, R. (2020). Characterization of  
530 Cabernet Sauvignon wines from California: determination of origin based on ICP-MS analysis and  
531 machine learning techniques. *European Food Research and Technology*, 246(6), 1193–1205.  
532 <https://doi.org/10.1007/s00217-020-03480-5>
- 533 Depeters, E. J., Puschner, B., Taylor, S. J., & Rodzen, J. A. (2013). Can fatty acid and mineral compositions

534 of sturgeon eggs distinguish between farm-raised versus wild white ( *Acipenser transmontanus* )  
535 sturgeon origins in California ? Preliminary report. *Forensic Science International*, 229(1–3), 128–132.  
536 <https://doi.org/10.1016/j.forsciint.2013.04.003>

537 Drab, K., & Daszykowski, M. (2014). Clustering in analytical chemistry. *Journal of AOAC International*,  
538 97(1), 29–38. <https://doi.org/10.5740/jaoacint.SGEDrab>

539 Epova, E. N., Bérail, S., Zuliani, T., Malherbe, J., Sarthou, L., Valiente, M., & Donard, O. F. X. (2018).  
540 <sup>87</sup>Sr/<sup>86</sup>Sr isotope ratio and multielemental signatures as indicators of origin of European cured hams:  
541 The role of salt. *Food Chemistry*, 246(October 2017), 313–322.  
542 <https://doi.org/10.1016/j.foodchem.2017.10.143>

543 European Parliament Resolution of 14 January 2014 on the food crisis, fraud in the food chain and the  
544 control thereof (2013/2091(INI)). *Official Journal of the European Union*, C482, 22-30. Retrieved  
545 from: <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A52014IP0011%2801%29>

546 Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Cluster analysis (5th ed). Wiley Series in  
547 Probability and Statistics. ISBN: 978-0-470-97844-3

548 FAO (2020) The state of world fisheries and aquaculture. Sustainability in action. Rome, Italy.  
549 <https://doi.org/10.4060/ca9229en>

550 FAO (2018) Overview of food fraud in the fisheries sector. Text by Reilly, A. Fisheries and Aquaculture  
551 Circular No. 1165, p. I. Retrieved from: <http://www.fao.org/documents/card/en/c/I8791EN/>

552 Ferrante, M., Napoli, S., Grasso, A., Zuccarello, P., Cristaldi, A., & Copat, C. (2019). Systematic review of  
553 arsenic in fresh seafood from the Mediterranean Sea and European Atlantic coasts: A health risk  
554 assessment. *Food and Chemical Toxicology*, 126(September 2018), 322–331.  
555 <https://doi.org/10.1016/j.fct.2019.01.010>

556 Fragni, R., Trifirò, A., & Nucci, A. (2015). Towards the development of a multi-element analysis by ICP-  
557 TOF-MS for tracing the geographical origin of processed tomato products. *Food Control*, 48, 96–101.  
558 <https://doi.org/10.1016/j.foodcont.2014.04.027>

559 Franke, B. M., Haldimann, M., Gremaud, G., Bosset, J. O., Hadorn, R., & Kreuzer, M. (2008). Element

- 560 signature analysis: Its validation as a tool for geographic authentication of the origin of dried beef and  
561 poultry meat. *European Food Research and Technology*, 227(3), 701–708.  
562 <https://doi.org/10.1007/s00217-007-0776-8>
- 563 Garelick, H., Jones, H., Dybowska, A., & Valsami-Jones, E. (2008). Arsenic pollution sources. *Reviews of*  
564 *Environmental Contamination and Toxicology*, 197(January 2017), 17–60. [https://doi.org/10.1007/978-](https://doi.org/10.1007/978-0-387-79284-2_2)  
565 [0-387-79284-2\\_2](https://doi.org/10.1007/978-0-387-79284-2_2)
- 566 Han, J., Kamber, M., & Pei, J. (2011). *Data Mining : Concepts and Techniques* (Third Edit). Morgan  
567 Kaufmann Publishers. <https://doi.org/10.1016/B978-0-12-381479-1.00001-0>
- 568 Hernandez-Herrero, M. M., Roig-Sagués, A. X., López-Sabater, E. I., Rodríguez-Jerez, J. J., & Mora-  
569 Ventura, M. T. (1999). Total Volatile Basic Nitrogen and other Physico- chemical and Microbiological  
570 Characteristics as. *Journal of Food Science*, 64(2), 344–347.
- 571 Husáková, L., Urbanová, I., Šrámková, J., Černohorský, T., Krejčová, A., Bednaříková, M., Frýdová, E.,  
572 Nedělková, I., & Pilařová, L. (2011). Analytical capabilities of inductively coupled plasma orthogonal  
573 acceleration time-of-flight mass spectrometry (ICP-oe-TOF-MS) for multi-element analysis of food and  
574 beverages. *Food Chemistry*, 129(3), 1287–1296. <https://doi.org/10.1016/j.foodchem.2011.05.047>
- 575 Husáková, L., Urbanová, I., Šídová, T., Cahová, T., Faltys, T., & Šrámková, J. (2015). Evaluation of  
576 ammonium fluoride for quantitative microwave-assisted extraction of silicon and boron from different  
577 solid samples. *International Journal of Environmental Analytical Chemistry*, 95(10), 922–935.  
578 <https://doi.org/10.1080/03067319.2015.1070409>
- 579 Ikem, A., & Egiebor, N. O. (2005). Assessment of trace elements in canned fishes (mackerel, tuna, salmon,  
580 sardines and herrings) marketed in Georgia and Alabama (United States of America). *Journal of Food*  
581 *Composition and Analysis*, 18(8), 771–787. <https://doi.org/10.1016/j.jfca.2004.11.002>
- 582 Juncos, R., Arcagni, M., Rizzo, A., Campbell, L., Arribére, M., & Guevara, S. R. (2016). Natural origin  
583 arsenic in aquatic organisms from a deep oligotrophic lake under the influence of volcanic eruptions.  
584 *Chemosphere*, 144, 2277–2289. <https://doi.org/10.1016/j.chemosphere.2015.10.092>
- 585 **Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4), 261-283.**



- 586 <https://doi.org/10.1007/s10462-011-9272-4>
- 587 Laso, J., Margallo, M., Fullana, P., Bala, A., Gazulla, C., Irabien, A., & Aldaco, R. (2017). Introducing life  
588 cycle thinking to define best available techniques for products: Application to the anchovy canning  
589 industry. *Journal of Cleaner Production*, 155, 139–150. <https://doi.org/10.1016/j.jclepro.2016.08.040>
- 590 Li, L., Han, C., Dong, S., & Boyd, C. E. (2019). Use of elemental profiling and isotopic signatures to  
591 differentiate Pacific white shrimp (*Litopenaeus vannamei*) from freshwater and seawater culture areas.  
592 *Food Control*, 95, 249–256. <https://doi.org/10.1016/j.foodcont.2018.08.015>
- 593 Lo Feudo, G., Naccarato, A., Sindona, G., & Tagarelli, A. (2010). Investigating the origin of tomatoes and  
594 triple concentrated tomato pastes through multielement determination by inductively coupled plasma  
595 mass spectrometry and statistical analysis. *Journal of Agricultural and Food Chemistry*, 58(6), 3801–  
596 3807. <https://doi.org/10.1021/jf903868j>
- 597 Magdas, D. A., Feher, I., Cristea, G., Voica, C., Tabaran, A., Mihaiu, M., Cordea, D. V., Bâlțeanu, V. A., &  
598 Dan, D. S. (2019). Geographical origin and species differentiation of Transylvanian cheese.  
599 Comparative study of isotopic and elemental profiling vs . DNA results. *Food Chemistry*, 277(June  
600 2018), 307–313. <https://doi.org/10.1016/j.foodchem.2018.10.103>
- 601 Moreno-Rojas, R., Cámara-Martos, F., Sánchez-Segarra, P. J., & Amaro-López, M. Á. (2012). Influence of  
602 manufacturing conditions and discrimination of Northern Spanish cheeses using multi-element  
603 analysis. *International Journal of Dairy Technology*, 65(4), 594–602. [https://doi.org/10.1111/j.1471-  
604 0307.2012.00853.x](https://doi.org/10.1111/j.1471-0307.2012.00853.x)
- 605 Rodushkin, I., Bergman, T., Douglas, G., Engström, E., Sörlin, D., & Baxter, D. C. (2007). Authentication of  
606 Kalix (N.E. Sweden) vendace caviar using inductively coupled plasma-based analytical techniques:  
607 Evaluation of different approaches. *Analytica Chimica Acta*, 583(2), 310–318.  
608 <https://doi.org/10.1016/j.aca.2006.10.038>
- 609 Rokach, L., & Maimon, O. Z. (2008). *Data mining with decision trees. Theory and applications* (2nd ed.).  
610 World Scientific Publishing Co. Pte. Ltd.
- 611 Saraçlı, S., Doğan, N., & Doğan, İ. (2013). Comparison of hierarchical cluster analysis methods by

612 cophenetic correlation. *Journal of Inequalities and Applications*, 203. <https://doi.org/10.1186/1029->  
613 242X-2013-203

614 Smith, R. G., & Watts, C. A. (2009). Determination of the country of origin of farm-raised shrimp (family  
615 penaeidae) using trace metal profiling and multivariate statistics. *Journal of Agricultural and Food*  
616 *Chemistry*, 57(18), 8244–8249. <https://doi.org/10.1021/jf901658f>

617 Sokal, R. R., & Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, 33-40.

618 Storelli, M. M., & Marcotrigiano, G. O. (2004). Interspecific variation in total arsenic body concentrations in  
619 elasmobranch fish from the Mediterranean Sea. *Marine Pollution Bulletin*, 48(11–12), 1145–1149.  
620 <https://doi.org/10.1016/j.marpolbul.2004.03.005>

621 Suhaj, M., & Kore, M. (2008). Study of some European cheeses geographical traceability by pattern  
622 recognition analysis of multielemental data. *European Food Research and Technology*, 227, 1419–  
623 1427. <https://doi.org/10.1007/s00217-008-0861-7>

624 Turra, C., Dias de Lima, M., Fernandes, E. A. D. N., Bacchi, M. A., Barbosa, F., & Barbosa, R. (2017).  
625 Multielement determination in orange juice by ICP-MS associated with data mining for the  
626 classification of organic samples. *Information Processing in Agriculture*, 4(3), 199–205.  
627 <https://doi.org/10.1016/j.inpa.2017.05.004>

628 Varrà, M. O., Ghidini, S., Zanardi, E., Badiani, A., & Ianieri, A. (2019). Authentication of European sea bass  
629 according to production method and geographical origin by light stable isotope ratio and rare earth  
630 elements analyses combined with chemometrics. *Italian Journal of Food Safety*, 8(1).  
631 <https://doi.org/10.4081/ijfs.2019.7872>

632 Velasco, A., Aldrey, A., Pérez-Martín, R. I., & Sotelo, C. G. (2016). Assessment of the labelling accuracy of  
633 spanish semipreserved anchovies products by FINS (forensically informative nucleotide sequencing).  
634 *Heliyon*, 2(6), e00124. <https://doi.org/10.1016/j.heliyon.2016.e00124>

635 Zkeri, E., Aloupi, M., & Gaganis, P. (2018). Seasonal and spatial variation of arsenic in groundwater in a  
636 rhyolithic volcanic area of Lesvos Island, Greece. *Environmental Monitoring and Assessment*, 190, 44.  
637 <https://doi.org/10.1007/s10661-017-6395-3>

638 **Figure Captions**

639 **Fig. 1.** Beeswarm box-plots with Kruskal-Wallis and Dunn's multiple comparison test results (median and  
640 quartiles) showing elements in bulk (A) and packaged (B) anchovy products varying significantly in relation  
641 to the origin ( $p < 0.05$ ).

642

643 **Fig. 2.** Hierarchical cluster analysis simplified dendrograms for bulk anchovy dataset (A) and packaged  
644 anchovy dataset (B) based on 35 elements.

645 **Fig. 3.** Comparison of the most important elemental predictors in C5.0, CART, CHAID, and QUEST models  
646 for bulk anchovies (A) and packaged anchovies (B). Values are scaled from 0 (no influence) to 1 (maximum  
647 influence).

648 **Fig. 4.** Decision classification tree resulting from the application of the CHAID algorithm for the classification  
649 of bulk anchovies using the element profile. Concentrations are reported in  $\mu\text{g kg}^{-1}$ .

650 **Fig. 5.** Decision classification tree resulting from the application of the QUEST algorithm for the classification  
651 of packaged anchovies using the element profile. Concentrations are reported in  $\mu\text{g kg}^{-1}$ .

652

653

1 **Classification of transformed anchovy products based on the use of**  
2 **element patterns and decision trees to assess traceability and country**  
3 **of origin labelling**

4  
5 *Maria Olga VARRÀ<sup>a</sup>, Lenka HUSÁKOVÁ<sup>b\*</sup>, Jan PATOČKA<sup>b</sup>, Sergio GHIDINI<sup>a</sup>,*  
6 *Emanuela ZANARDI<sup>a\*</sup>*

7  
8 *<sup>a</sup>Department of Food and Drug, University of Parma, Parma, Via del Taglio, 10, Parma 43126, Italy*

9 E-mail addresses: emanuela.zanardi@unipr.it (E. Zanardi); sergio.ghidini@unipr.it (S. Ghidini);

10 mariaolga.varra@studenti.unipr.it (M. O. Varrà).

11 *<sup>b</sup>Department of Analytical Chemistry, Faculty of Chemical Technology, University of Pardubice,*

12 *Studentská 573 HB/D, Pardubice, CZ-532 10, Czech Republic*

13 E-mail address: jan.patočka@upce.cz (J. Patočka)

14

15 **\*CORRESPONDING AUTHORS:**

16 **Lenka Husáková** (L. Husáková)

17 Department of Analytical Chemistry, Faculty of Chemical Technology, University of Pardubice

18 Studentská 573 HB/D, Pardubice, CZ-532 10 (Czech Republic)

19 Tel. +420 466 037 029; Fax: +420 466 037 068; E-mail address: lenka.husakova@upce.cz

20

21 **Emanuela Zanardi** (E. Zanardi)

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

23 Tel. +39.0521.902.760; E-mail address: emanuela.zanardi@unipr.it

24

25 **ABSTRACT**

26 Quadrupole inductively coupled plasma mass spectrometry (Q-ICP-MS) and direct mercury analysis were used  
27 to determine the elemental composition of 180 transformed (salt-ripened) anchovies from three different  
28 fishing areas before and after packaging. To this purpose, four decision trees-based algorithms, corresponding  
29 to C5.0, classification and regression trees (CART), chi-square automatic interaction detection (CHAID), and  
30 quick unbiased efficient statistical tree (QUEST) were applied to the elemental datasets to find the most  
31 accurate data mining procedure to achieve the ultimate goal of fish origin prediction. Classification rules  
32 generated by the trained CHAID model optimally identified unlabelled testing bulk anchovies (93.9% F-score)  
33 by using just 6 out 52 elements (As, K, P, Cd, Li, and Sr). The finished packaged product was better modelled  
34 by the QUEST algorithm which recognised the origin of anchovies with F-score of 97.7%, considering the  
35 information carried out by 5 elements (B, As, K, Cd, and Pd). Results obtained suggested that the traceability  
36 system in the fishery sector may be supported by simplified machine learning techniques applied to a limited  
37 but effective number of inorganic predictors of origin.

38

39 **Abbreviations**

40 Certified reference materials, CRMs; classification and regression trees, CART; chi-square automatic  
41 interaction detection, CHAID; hierarchical cluster analysis, HCA; high energy Helium mode, HE He;  
42 inductively coupled plasma mass spectrometry, ICP-MS; inductively coupled plasma optical emission  
43 spectrometry ICP-OES; internal standard, ISTD; kinetic energy discrimination, KED; method detection limit,  
44 MDL; method limit of quantification, MLOQ; quick unbiased efficient statistical tree, QUEST.

45

46 **Keywords:** *Engraulis encrasicolus*; fish products; decision trees; geographical origin; data mining; ICP-MS.

47

## 48 **1. Introduction**

49 Foodstuff free-trade between nations all over the world, together with increasing diversification into food-  
50 related products, recently made the development of easy, rapid, cheap, and robust tools to assess traceability  
51 of foodstuffs to become a hot topic in the scientific community as well as in an industrial context.

52 The fishery sector is particularly prone to fraudulent practices but, on the other hand, it is insufficiently  
53 protected. The high complexity of the fish supply chain, the high number of stakeholders involved, and the  
54 fast perishability of fish, are a few of the many factors hampering the fight against fraud, which, in turn, reflect  
55 negatively on producers, transformers and final consumers from both economical and sanitary point of view  
56 (FAO, 2018; European Parliament Resolution 2013/2091(INI), 2014).

57 The perception of quality fresh or processed fish and seafood products by consumers is the sum of several  
58 different objective and subjective factors and it directly influences the global economic and market values of  
59 the product. At present, mislabelling or misrepresenting the origin of fish products keep getting encouraged by  
60 the so-called country-of-origin effect, according to which the consumers increasingly tend to associate high  
61 quality fish products with specific production areas because of specific sensorial characteristics, ethical or  
62 ecological motivations.

63 In this context, processed fish products deriving from the industrial transformation of the highly valuable  
64 European anchovy (*Engraulis encrasicolus*, L. 1758) are frequently subjected to fraud (Velasco, Aldrey,  
65 Pérez-Martín, & Sotelo, 2016).

66 European anchovy is a small pelagic fish that is mainly fished in the Mediterranean Sea and Black Sea, as well  
67 as in Eastern Central Africa (alongside the Moroccan coasts) and in Northeast Atlantic, especially in the  
68 Cantabrian Sea (FAO, 2020). In addition to the direct consumption as fresh fish, the product is frequently  
69 found in the European marketplace in the form of transformed, brine-fermented anchovy or filleted and canned  
70 (preserved in oil) anchovy (FAO, 2020).

71 The traditional anchovy transformation process by brine-ripening finds a long tradition in southern Europe.  
72 The fish, typically caught by purse seines, is quickly transported to the fish canning industry where it is  
73 beheaded, partially eviscerated and put into ripening containers (barrels), alternating layers of fish and salt.  
74 A pressure is then applied on the top layer to facilitate the progressive elimination of water. The fish is ripened  
75 until the desired degree of maturation is reached (from 3 up to 11-12 month on average) to then be moved from

76 the barrels. From that point on, the bulk ripened anchovies can be preserved and packaged in salt to be  
77 commercialised or further processed to obtain different products and preparations, for example by filleting and  
78 packaging-in-oil.

79 During the ripening, several chemical and physical modifications occur, including lipolysis, lipid oxidation,  
80 and proteolysis (Hernandez-Herrero, Roig-Sagués, López-Sabater, Rodriguez-Jerez, & Mora-Ventura, 1999;  
81 Czerner, Agustinelli, Guccione, & Yeannes, 2015). These modifications are of fundamental importance to  
82 prolong the shelf life and reduce the microbiological-associated risks and, at the same time, they influence the  
83 final organoleptic characteristics of the products (Besteiro, Rodríguez, Tilve-Jar, & Pascual López, 2000).

84 Salted anchovy from the Cantabrian Sea (Northern Spain) is worldwide appreciated as a high-quality product  
85 thanks to the sensorial characteristics of the raw fish, the strong link with the territory, and the long artisanal  
86 tradition behind its manufacturing (Laso et al., 2017). Taking into consideration the Cantabrian anchovy  
87 overall reputation and its high commercial value, it is therefore assumed to be object of fraud by substitution  
88 with fish from other sources. Therefore, developing methods that aim at providing concrete protection to the  
89 product is a matter of the utmost importance.

90 Up to now, the scientific research dealing with the identification of fish and seafood origin has been mainly  
91 focused on raw untransformed fish and seafood and has made use of different approaches. Among these,  
92 approaches based on the use of the inorganic components, such as stable isotopes (Carrera & Gallardo, 2017),  
93 mineral, trace- and/or ultra-trace elements (Smith & Watts, 2009), and a combination of stable isotopes and  
94 trace elements (Li, Han, Dong, & Boyd, 2019; Varrà, Ghidini, Zanardi, Badiani, & Ianieri, 2019) have been  
95 demonstrated to be successful strategies since offering several advantages depending on the reflection of  
96 seawater overall compositions on fish flesh.

97 Tracing back to the origin of processed or highly processed products is considerably difficult because of the  
98 manipulation and the addition of several compounds during preparation procedures. The use of salt during  
99 anchovy manufacturing may represent the most critical point since it can potentially mask the natural elemental  
100 content of fish. Nevertheless, the multiple identification of elements using techniques such as inductively  
101 coupled plasma-optical emission spectroscopy (ICP-OES) and inductively coupled plasma-mass spectrometry  
102 (ICP-MS) have been already successfully applied to identify the origin of transformed food products such  
103 processed tomato products (Lo Feudo, Naccarato, Sindona, & Tagarelli, 2010; Fragni, Trifirò, & Nucci, 2015),

104 fruit juices (Turra et al., 2017), wines (da Costa, Ximenez, Rodrigues, Barbosa, & Barbosa, 2020), dried beef  
105 (Franke et al., 2008), hams (Epova et al., 2018), and different types of cheese (Suhaj & Kore, 2008; Moreno-  
106 Rojas, Cámara-Martos, Sánchez-Segarra, & Amaro-López, 2012; Magdas et al., 2019). One application  
107 dealing with the use of multi-elemental analysis to authenticate seafood products is also available and it  
108 concerns the identification of caviar from different origins, which, as anchovy, is a salted product (Rodushkin  
109 et al., 2007).

110 The success of most of these applications was anyway strictly dependent on the support provided by  
111 chemometrics and machine learning methods for the identification of those elemental patterns echoing the  
112 original environment.

113 In this study, four decision tree algorithms corresponding to C5.0, classification and regression trees (CART),  
114 chi-square automatic interaction detection (CHAID), and quick unbiased efficient statistical tree (QUEST)  
115 were applied to elemental content of transformed anchovy products to predict the country of origin of the raw  
116 materials and support traceability of the products before and after packaging. The reasons behind the selection  
117 of decision trees as the basis method of this study are since decision tree models are easily understandable and  
118 interpretable, quick to build and, in general, low training times are associated with their use. Moreover, these  
119 techniques have high prediction accuracy in many fields so that makes them preferable and trustable choices  
120 for this kind of task.

## 121 **2. Materials and Methods**

### 122 *2.1. Chemicals, standards, and reference materials*

123 All the aqueous solutions employed for analyses were prepared using ultrapure water ( $0.05 \mu\text{S cm}^{-1}$ ) obtained  
124 by the Milli-Q<sup>®</sup> water purification system (Millipore, Bedford, USA).

125 For microwave digestion, hydrogen peroxide ( $\text{H}_2\text{O}_2$ ,  $\geq 30\%$  w/w) for ultra-trace analysis (Fluka Chemie AG,  
126 Buchs, Switzerland) and sub-boiled nitric acid prepared from nitric acid (65%, w/w, Selectipur quality, Lach-  
127 Ner, Neratovice, Czech Republic) by means of the sub-boiling distillation apparatus Distillacid<sup>™</sup> BSB-939-  
128 IR (Berghof, Eningen, Germany) were used.

129 The working calibration solutions for ICP-MS analysis were prepared daily using the multi-element stock  
130 solutions “A”, “B1”, “B2”, and “C”. Stock solution “A” ( $10 \text{ mg L}^{-1}$  of Li, B, Al, V, Cr, Fe, Mn, Ni, Cu, Zn,



131 Co, As, Se, Rb, Sr, Zr, Mo, Ru, Pd, Cd, Sn, Sb, Cs, Ba, Hf, Re, Pt, Tl, Pb, Bi, and Th) was prepared from the  
132 Supelco ICP multi-element standard solution IV (Merck, Darmstadt, Germany) (containing 1 g L<sup>-1</sup> of Li, B,  
133 Al, Cr, Fe, Mn, Ni, Cu, Zn, Co, Sr, Cd, Ba, Tl, Pb, and Bi) and single element standards (V, As, Se, Rb, Zr,  
134 Mo, Ru, Pd, Sn, Sb, Cs, Hf, Re, Pt, and Th) of concentration 1 ± 0.002 g L<sup>-1</sup> (Analytika Ltd., Prague, Czech  
135 Republic or SCP Science, Montreal, Canada).

136 Stock solutions “B1” (1 mg L<sup>-1</sup> of La, Ce, Pr, Nd, and U) and “B2” (0.20 mg L<sup>-1</sup> of Y, Tb, Ho, Yb, Sm, Eu,  
137 Gd, Er, Tm, Lu, and Dy) were prepared from the stock solution of rare earth elements Astatol mix “M008”  
138 (Analytika Ltd., Prague, Czech Republic). Stock solution “C” (50 mg L<sup>-1</sup> of Na, Mg, P, K, Ca, Mn, Cu, and  
139 Zn) was prepared from single element standards of 1 g L<sup>-1</sup> (Analytika Ltd., Prague, Czech Republic).

140 A 1 g L<sup>-1</sup> stock solution of Rh (SCP Science, Montreal, Canada) was used to prepare the internal standard  
141 solution (ISTD) at concentration of 200 µg L<sup>-1</sup>.

142 A 10 g L<sup>-1</sup> stock solution prepared from urea (TraceSelect quality, Fluka Chemie AG, Buchs, Switzerland) was  
143 used to prepare carbon reference solutions.

144 The element quantification accuracy was evaluated using the following certified reference materials (CRMs):  
145 NIST SRM 1577 Bovine Liver (National Institute of Science and Technology, NIST, Gaithersburg,  
146 MD, USA); NIST SRM 1566 Oyster Tissue (NIST, Gaithersburg, MD, USA); BCR<sup>®</sup> certified reference  
147 material (CRM)184 Bovine muscle (Institute for Reference Materials and Measurements, IRMM, Geel,  
148 Belgium); BCR<sup>®</sup> 185 Bovine Liver (IRMM, Geel, Belgium); CRM NCS ZC73015 Milk Powder (National  
149 Research Centre for Certified Reference Materials, NRCRM, Beijing, China); P-WBF CRM 12-2-04 Essential  
150 and Toxic Elements in Wheat Bread Flour (pb-anal, Kosice, Slovakia); CRM12-2-03 P-Alfalfa Essential and  
151 toxic elements in Lucerne (pb-anal, Kosice, Slovakia); SMU CRM 12-02-01 Bovine liver (pb-anal, Kosice,  
152 Slovakia).

## 153 2.2. Anchovy sampling and processing

154 Salt-ripened anchovies as bulk product (semi-finished, non-packaged) and as packaged (finished) product were  
155 obtained from the processing of European anchovy (*Engraulis encrasicolus* L.) and provided by the same fish  
156 preserves company.

157 A total of 90 bulk specimens were randomly collected from different ripening barrels' batches after maturation  
158 and suddenly vacuum-packaged into plastic bags. Similarly, a total of 90 finished specimens were obtained  
159 after salt-packaging of bulk anchovies and provided packaged into glass jars.

160 Both types of products were prepared from salting process (using not iodised sea salt of the same origin) of  
161 raw fish caught in the following geographical areas: Cantabrian Sea (Spain, FAO fishing area 27.8, n=30),  
162 upper Central Mediterranean Sea (Croatia, FAO fishing area 37.2.1, n=30) and lower Central Mediterranean  
163 Sea, (Tunisia, FAO fishing area 37.2.2, n=30).

164 Detailed information on the sampling and characteristics of the transformed fish used in the present study is  
165 reported in Table S1 (Supplementary Materials).

166 Before analysis, each individual fish was carefully cleansed with filter paper to remove external salt and  
167 manually peeled, eviscerated, and deboned, and finally minced with a ceramic knife. After that, samples were  
168 individually stored into glass vials and frozen at -20 °C.

#### 169 *2.2.1 Lyophilisation process*

170 Around 3.5 g of each trimmed fish sample were transferred into 5mL lyophilisation vials (borosilicate glass  
171 Vacule® equipped with 3-leg stopper, Wheaton, USA) wherein the material was dried. Before the freeze-  
172 drying process, the samples were deep-frozen at -80 °C for 24 hours to provide a necessary conditioning for  
173 low temperature drying. CoolSafe 4-15 L benchtop freeze dryer (LaboGene, Lyngø, Denmark) was employed  
174 for the lyophilisation of samples, with the CoolSafe condenser working temperature held at -110°C and a total  
175 chamber pressure of 3 hPa. The freeze-dried samples were subsequently homogenised directly inside the glass  
176 vials using a plastic rod to obtain a fine powder.

#### 177 *2.2.2 Microwave-assisted acid digestion*

178 For subsequent ICP-MS analysis, 0.1 g of freeze-dried samples or CRMs were weighted (in triplicate) into a  
179 10 mL perfluoroalkoxy (PFA) tube and 4 mL of 16% HNO<sub>3</sub> (65%, w/w HNO<sub>3</sub>, 1:3 diluted) and 1 mL of 30%  
180 H<sub>2</sub>O<sub>2</sub> were added. Three PFA tubes were placed into DAC-100S polytetrafluoroethylene vessels (Berghof,  
181 Eningen, Germany) previously filled with 25 mL of HNO<sub>3</sub> (16%, v/v), by ensuring that the level of liquid in  
182 the outer polytetrafluoroethylene vessel was higher than those in the PFA tubes. This way, the vapor pressures  
183 were compensated and the evaporation of the solution from the PFA tubes was avoided (Husáková et al., 2015).

184 Samples were decomposed using a Berghof Speedwave™ MWS-3<sup>+</sup> microwave digestion system (Berghof,  
185 Germany) with the maximum total output of the microwave generator (1450 W) via the following multistep  
186 program: step 1, 20 min at 180 °C (ramp 5 min); step 2, 20 min at 220 °C (ramp 5 min); steps 3, 5 min at  
187 100 °C (ramp 5 min).

188 The clear digested samples were diluted with deionised water up to 25 mL and the residual carbon content  
189 quantified at  $5.58 \pm 0.12\%$  by ICP-OES, following the method previously reported by Husáková et al. (2011).

### 190 2.3. Mercury analysis

191 Total Hg content was determined directly on lyophilised solid samples or CRMs using a single-purpose atomic  
192 absorption spectrometer AMA 254 (Altec Ltd., Prague, Czech Republic).

193 Analytical operation conditions as follows: sample mass, 50 mg; drying step, 60 s at 120 °C; decomposition  
194 step, 150 s at 750 °C; Hg release step, 45 s at 900 °C; reading step, 60 s monitoring the 253.6 nm absorbance  
195 peak. The flow rate of oxygen (99.5%) carrier gas was 170 mL min<sup>-1</sup>.

### 196 2.4. ICP-MS analysis

197 Element quantification in samples was performed by using an Agilent 7900 quadrupole ICP-MS apparatus  
198 (Agilent Technologies, Inc., Santa Clara, CA, USA) equipped with a quartz concentric nebulizer MicroMist  
199 (400 µL min<sup>-1</sup>), the Peltier-cooled (2 °C) Scott quartz spray chamber, quartz torch with 2.5 mm internal  
200 diameter injector, standard sampling and skimmer nickel cones with orifices of 1 and 0.45 mm, and an octopole  
201 collision/reaction cell for interference removal using kinetic energy discrimination. A low-pulsation, 10-roller  
202 peristaltic pump with three separate channels was employed to precisely deliver both samples and ISTD. For  
203 the simultaneous ISTD aspiration and mixing with the sample the connecting tubing, connectors, and the “Y”  
204 piece (included into internal standard kit) were employed. Analytical conditions were enhanced before starting  
205 sample measurement by using the multi-elemental tuning solution (Agilent Technologies, Inc., Santa Clara,  
206 CA, USA) containing 1 µg L<sup>-1</sup> of Ce, Co, Li, Mg, Tl and Y, to obtain the highest possible sensitivity for  
207 elements of low, middle, and high m/z. Using the typical operating conditions summarised in Table S2  
208 (Supplementary Materials), a sensitivity of 6000 counts s<sup>-1</sup> per µg L<sup>-1</sup> and a resolution of 0.64 amu peak width

209 (full width at half maximum intensity) were achieved for  ${}^7\text{Li}^+$ . The same parameters were 50000 counts  $\text{s}^{-1}$  per  
210  $\mu\text{g L}^{-1}$  and 0.62 for  ${}^{89}\text{Y}^+$ , and 30000 counts  $\text{s}^{-1}$  per  $\mu\text{g L}^{-1}$  and 0.60 for  ${}^{205}\text{Tl}^+$ .

211 While the quantification of certain elements was performed without pressurising the collision cell (i.e., “no-  
212 gas” mode), “Helium” mode (He) and “High Energy Helium” mode (HE He) were instead used for the  
213 quantification of problematic elements mostly suffering from polyatomic interferences.

214 The acquisition mode for elements as follows:  ${}^7\text{Li}$ ,  ${}^{11}\text{B}$ ,  ${}^{24}\text{Mg}$ ,  ${}^{66}\text{Zn}$ ,  ${}^{85}\text{Rb}$ ,  ${}^{88}\text{Sr}$ ,  ${}^{89}\text{Y}$ ,  ${}^{90}\text{Zr}$ ,  ${}^{95}\text{Mo}$ ,  ${}^{101}\text{Ru}$ ,  ${}^{103}\text{Rh}$ ,  
215  ${}^{105}\text{Pd}$ ,  ${}^{111}\text{Cd}$ ,  ${}^{118}\text{Sn}$ ,  ${}^{121}\text{Sb}$ ,  ${}^{133}\text{Cs}$ ,  ${}^{138}\text{Ba}$ ,  ${}^{139}\text{La}$ ,  ${}^{140}\text{Ce}$ ,  ${}^{141}\text{Pr}$ ,  ${}^{146}\text{Nd}$ ,  ${}^{147}\text{Sm}$ ,  ${}^{153}\text{Eu}$ ,  ${}^{157}\text{Gd}$ ,  ${}^{159}\text{Tb}$ ,  ${}^{163}\text{Dy}$ ,  ${}^{165}\text{Ho}$ ,  ${}^{166}\text{Er}$ ,  
216  ${}^{172}\text{Yb}$ ,  ${}^{175}\text{Lu}$ ,  ${}^{178}\text{Hf}$ ,  ${}^{185}\text{Re}$ ,  ${}^{195}\text{Pt}$ ,  ${}^{205}\text{Tl}$ ,  ${}^{206+207+208}\text{Pb}$ ,  ${}^{209}\text{Bi}$ ,  ${}^{232}\text{Th}$ ,  ${}^{238}\text{U}$ , all by “No gas mode”;  ${}^{23}\text{Na}$ ,  ${}^{27}\text{Al}$ ,  ${}^{39}\text{K}$ ,  
217  ${}^{44}\text{Ca}$ ,  ${}^{51}\text{V}$ ,  ${}^{52}\text{Cr}$ ,  ${}^{55}\text{Mn}$ ,  ${}^{56}\text{Fe}$ ,  ${}^{59}\text{Co}$ ,  ${}^{60}\text{Ni}$ ,  ${}^{63}\text{Cu}$ ,  ${}^{103}\text{Rh}$ , all by “He mode”;  ${}^{31}\text{P}$ ,  ${}^{75}\text{As}$ ,  ${}^{78}\text{Se}$ ,  ${}^{103}\text{Rh}$  all by “HE He  
218 mode”.

219 The calibration curves for the quantification of 51 elements ( $R^2 > 0.999$ ) resulted from the acquisition of  
220 working calibration solutions prepared from multi-element stock solutions “A”, “B1”, “B2”, and “C” described  
221 in Section 2.1. The concentration of elements for calibration were as follows: blank, 1, 5, 10, 50, 100  $\mu\text{g L}^{-1}$  of  
222 Li, Be, B, Al, V, Cr, Fe, Mn, Ni, Cu, Zn, Co, Ga, Ge, As, Se, Rb, Sr, Zr, Mo, Ru, Cd, In, Sn, Sb, Te, Cs, Ba,  
223 Hf, Ta, Re, Pt, Tl, Pb, Bi, Th; 0.1, 0.5, 1, 5, 10  $\mu\text{g L}^{-1}$  of La, Ce, Pr, Nd, U; 0.02, 0.1, 0.2, 1, 2  $\mu\text{g L}^{-1}$  of Y, Sm,  
224 Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu; 0.5, 1, 5, 10  $\text{mg L}^{-1}$  of Na, Mg, P, K, Ca, Mn, Cu, Zn.

225 Samples were measured in triplicate. Standards and blanks were analysed after a single initial calibration.  
226 Continuing calibration blanks and calibration verification standards were automatically run after every 25  
227 samples. To compensate possible instrumental drift and matrix effects, a 200  $\mu\text{g L}^{-1}$  Rh ISTD was  
228 simultaneously aspired and mixed with samples. The percent recovery of the ISTD responses for the entire 12-  
229 hour sequence normalised to the calibration blanks shows that there was no gradual loss of sensitivity over  
230 time, even when running high matrix samples. Long term stability measured by comparing ISTD responses  
231 from the beginning of the sequence to the end was better than 5 %.

## 232 2.5. Analytical validation

233 The accuracy of the methodology was checked by analysing the five certified reference materials listed in  
234 Section 2.1 (NIST 1566 Oyster Tissue, BCR CRM 184 Bovine muscle, BCR CRM 185 Bovine Liver, CRM  
235 12-2-01 Bovine Liver, NIST SRM 1577 Bovine Liver) intended for the evaluation of analytical methods and

236 instruments and used for the determination of the mass fraction values of selected elements in marine tissue,  
237 foods, or similar materials. In addition, CRM 12-2-03 Essential and toxic elements in Lucerne, NCS ZC 73015  
238 Milk Powder, and CRM 12-2-04 Wheat bread flour, were analysed to assess accuracy in determining  
239 lanthanides and actinides. The high level of agreement between target and found values demonstrated trueness  
240 of data obtained (Supplementary Materials, Table S3).

241 Intra-day and inter-day precisions were calculated to assess the overall precision of the method and were  
242 determined by analysing single CRMs three times during the same day and during three different days over  
243 one month, respectively. The method was found to be precise enough due to the percent relative standard  
244 deviations (RDS %) of intra-day and inter-day precision which were mostly below 14 % (Supplementary  
245 Materials, Table S3).

246 The 3 sigma method detection limits (MDL) and the method limits of quantification (MLOQ) reported in Table  
247 S4 (Supplementary Materials), were experimentally determined through the measurement of ten replicates of  
248 a blank sample and by calculating triple and tenfold standard deviations divided by the slope of the calibration  
249 curve, respectively. In all cases, detection limits were found significantly below the typical requirements for  
250 this analysis so that selected elements could be determined at the background level. Table S4 (Supplementary  
251 Materials) also summarises relative sensitivities of Q-ICP-MS for analysis of individual elements with the use  
252 of Rh ISTD.

## 253 2.6. Data elaboration

254 Statistics was applied to elemental concentrations referring to anchovy dry matter (d.m.) and carried out using  
255 IBM SPSS software (v. 23.0, SPSS Inc., Chicago, IL, USA), SIMCA software (v. 16.0.2, Sartorius  
256 Stedim Data Analytics AB, Sweden), NCSS 2020 software (v. 20.0.3, NCSS LLC., Kaysville, UT, USA), and  
257 IBM SPSS Modeler software (v. 18.2, SPSS Inc., Chicago, IL, USA).

258 Univariate data analysis, consisting of nonparametric Mann-Whitney test ( $p \leq 0.05$ ) and Kruskal–Wallis test  
259 plus Dunn’s post hoc test ( $p \leq 0.05$ ) was applied to the whole data of both the set of bulk anchovies (30 samples  
260  $\times$  3 replicates  $\times$  3 provenances) and the set of packaged anchovies (30 samples  $\times$  3 replicates  $\times$  3 provenances)  
261 to investigate any significant difference between groups of samples. Nonparametric tests were chosen instead

262 of classical parametric ones because most of the data presented non normal distribution and heteroscedasticity,  
263 as verified though the application of Shapiro-Wilk's and Levene's tests, respectively ( $p \leq 0.05$ ).

264 As a classical unsupervised chemometric method, hierarchical cluster analysis (HCA) using factorial  
265 coordinates of principal component analysis was applied to the two anchovy datasets in order to scout data  
266 structures and identify group of samples by the similarity of their variables (Drab & Daszykowski, 2014).

267 To create classification models with good validity and consistency, the holdout technique (stratified randomly  
268 sampling) was adopted when machine learning was applied. The bulk and the packaged anchovy datasets were  
269 organised into three different data matrices each in a 70:15:15 partition ratio to create training, validation, and  
270 testing and sets, respectively. The training sets were used to estimate the models, the validation set to test and  
271 select the best models and the testing sets to confirm the reliability of the selected models.

272 Different learning algorithms relying on the principle of decision trees and which do not require assumptions  
273 about data distribution were chosen to learn how the data of the training set were classifiable according to  
274 origin of samples and to create proper prediction models. These were C5.0, CART, CHAID, and QUEST.

275 Briefly, decision trees work on the division/classification of samples driven by the values of the variables under  
276 examination, by creating subsets of samples which are progressively split across the structure of the tree,  
277 independently of the distribution of the predictor. Therefore, decision trees are nonparametric machine learning  
278 algorithms specifically seeking for data partitioning into response-homogeneous zones (Kotsiantis, 2013;  
279 Barbosa, Nacano, Freitas, Batista, & Barbosa, 2014).

280 The output of the method is a set of concatenated classification rules in the form of a decision tree composed  
281 by nodes (identifying the features that need to be sorted) and branches (identifying the values assumed by  
282 nodes) (Han, Kamber, & Pei, 2011). CART, C5.0, CHAID, and QUEST represent different methods by which  
283 the architecture of the decision trees can be built up and differ each other mainly for the segmentation rules  
284 applied and the tree optimisation method. The C5.0 algorithm is based on binary splitting and works by  
285 choosing progressively the instances that allow to gain the maximum partitioning information and stopping  
286 via the pruning rule, i.e., by removing from the splits which do not add significant information. CART  
287 algorithm is also based on binary splitting but differs from C5.0 essentially for a different stopping rule in the  
288 creation of the tree, consisting of the evaluation of the purity of the node, i.e. the maximum degree of  
289 homogeneity between categories. The CHAID algorithm is a multiway splitting system based on Chi-square

290 statistics to decide for tree ramifications and is based on the measures of the impurity of the nodes. Finally, the  
291 QUEST is a binary-split algorithm which, in the case of continuous variables, uses an ANOVA F-Test to create  
292 tree nodes. Further information can be retrieved from Han, Kamber, & Pei (2011) and from Rokach & Maimon  
293 (2008).

294 In the present work, CART was built using the Gini Impurity Index to determine the nodes impurity and select  
295 input variables. For CART, CHAID, and QUEST a maximum of five tree levels was set to avoid excessive  
296 splitting. Building settings for CHAID included the use of Pearson's Chi-square statistics and a Bonferroni  
297 adjustment to calculate the adjusted  $p$ -values. Significant level for splitting for both CHAID and QUEST was  
298 set at 95%.

299 Training models were compared each other by analysing classical metrics in multivariate classification  
300 methods, corresponding to accuracy (%), sensitivity (%), specificity (%), precision (%), and F-score (%)  
301 indexes, calculated from the unlabelled test set. Calculation were performed according to Cuadros-Rodríguez,  
302 Pérez-Castaño, & Ruiz-Samblás (2016).

### 303 **3. Results and discussion**

#### 304 *3.1. Initial data evaluation*

##### 305 3.1.1 Global elemental profiles of transformed anchovy products

306 The anchovy samples distributions according to the measured elemental concentrations which varied  
307 significantly in relation to the three investigated geographical provenances are shown Fig. 1. The beeswarm  
308 boxplots reported revealed that, according to Kruskal-Wallis and Dunn's multiple comparison test,  
309 concentrations of B, V, As, and Hg were different ( $p < 0.05$ ) among Cantabrian, Tunisian and Croatian bulk  
310 anchovies (Fig. 1A). Packaged anchovies, indeed, differed for the same element concentrations plus those of  
311 Li (Fig. 1B). Regardless the type of product, the highest amounts of B and V were found in Tunisian samples  
312 and those of As and Hg in Croatian samples (Fig. 1A, Fig. 1B).

313 Complete data matrices reporting the whole concentrations of the 52 elements measured in bulk and packaged  
314 anchovies can be found in Table S5 and Table S6 (Supplementary Materials), respectively. The most abundant  
315 element was found to be Na (whose median concentrations ranged from 142 mg kg<sup>-1</sup> in bulk Croatian

316 anchovies to 177 mg kg<sup>-1</sup> in Tunisian packaged anchovies) followed by P, K, Ca, and Mg. Bulk and packaged  
317 products from Cantabrian Sea differed from samples of Mediterranean origin (Croatian and Tunisian) because  
318 of significantly higher concentrations of P and K. At the same time, no differences in Na and Ca contents  
319 between Cantabrian and Croatian anchovies was encountered, which, instead, were both significantly lower  
320 compared to those found in samples originating from Tunisia (Table S5, Table S6, Supplementary Materials).  
321 Similarly, some minor, trace- and ultra-trace elements showed significant variations in relation to the  
322 provenance Besides, Cantabrian anchovies were characterised by higher Ni, Mo, Cs, and Tl concentrations  
323 and lower Al concentrations than those encountered in the two Mediterranean products (Supplementary  
324 Tables S5, S6).

325 Considering the peculiarity of the products investigated, no published works dealing with the same food  
326 matrices and purposes were found, except for similar canned products including, however, other ingredients  
327 (Ikem & Egiebor, 2005). For this reason, a more in-depth analysis of the elemental concentrations was not  
328 possible.

329 Beyond statistical differences related to the geographical origin, a high degree of variations of elemental  
330 concentrations within fish of the same provenance was however found. Considering equal characteristics of  
331 marine environment for each group, this variation is likely to be attributable to the natural biological diversity  
332 among individuals. Moreover, the technology behind the processing of anchovies does suggest that all  
333 manufacturing operations (including handling, treatment, production, and distribution) between the time of  
334 fishing and the end-product stage can impact the final elemental profile of anchovies. Therefore, potential  
335 markers of geographical origin need to go through a more-in-depth evaluation to be correctly identified,  
336 thereby preventing misinterpretations.

### 337 3.1.2 Pre-selection of elements as potential indicators of origin

338 The use of salt in the form of saturated brine during the processing and packaging of ripened anchovies is one  
339 of the main complicating factors which might limit the proper identification of markers of geographical origin  
340 since it inevitably modifies the natural element profile of the product. For instance, the use of a brine for  
341 fermentation purposes was reported to decrease the total amount of certain trace elements in transformed fish



342 roe because of partitioning phenomena between the solid and the liquid phases occurring during fermentation  
343 (Bekhit, Morton, & Dawson, 2008).

344 Despite the addition of salt as an exogenous source of contaminants, the possibility of using the elemental  
345 profile to discriminate the origin of processed salted cheeses has been previously investigated and it was found  
346 that some elements as Sr, Li, Mg, Rb and K (Magdas et al., 2019) as well as Tl, Li (Epova et al., 2018) can  
347 still be used as powerful tracers since retain a strong link with the place of origin of milk. In addition,  
348 concentrations of As, Ba, Br, I, Mo, and Se, were proven to be stable although processing and were identified  
349 as useful markers for the geographic origin of caviar (Rodushkin et al., 2007) as well as to distinguish caviar  
350 obtained from wild sturgeon (Depeters, Puschner, Taylor, & Rodzen, 2013). On the contrary, the concentration  
351 of Fe, Al, Ti, V were found to be heavily affected by handling and packaging operations and, therefore, useless  
352 for authentication of transformed products (Rodushkin et al., 2007).

353 In the present work, the elemental profile of the bulk anchovies of each origin was compared to that of the  
354 finished packaged products, to highlight possible modifications occurring during the end stages of anchovies  
355 processing. Results from Mann-Whitney test (two-tailed, confidence level 95%) highlighted the concentrations  
356 of following elements to be significantly different between the two types of products at least in two out of three  
357 origin groups: Na, Mg, Ca, Cu, Cr, V, Ru, Rb, La, Ce, Pr, Gd, Re and Tl (see Supplementary Materials,  
358 Table S7).

359 Considering that the main aim of the present research was to create machine learning based models as robust  
360 as possible in classifying samples by origin, highly variable elements identified both in the present study (Na,  
361 Mg, Ca, Cu, Cr, V, Ru, Rb, La, Ce, Pr, Gd, Re and Tl), and retrieved from the literature (Fe, Ti, and Al) were  
362 excluded and a total of 35 input variables, i.e. elements, were retained for subsequent classification analyses.

### 363 3.1.3. Hierarchical cluster analysis (HCA)

364 The inner potential of the elemental profile in guiding the creation of groups of anchovies was at first instance  
365 explored. Any possible natural difference or similarity among anchovies of both datasets was therefore  
366 uncovered by the application of HCA. Since different methods to measure distances (i.e. Euclidean and  
367 Manhattan distances) and to perform grouping (i.e. Ward's minimum variance, single linkage, complete  
368 linkage, simple average group average, median, and centroid clustering methods) are applicable for HCA

369 building, the cophenetic correlation coefficient (CCC) was employed as an index to compare the methods with  
370 each other and to evaluate the validity of the resulting dendrograms (Sokal & Rohlf, 1962; Saraçlı, Doğan, &  
371 Doğan, 2013). Further details about the tested methods can be retrieved from Everitt, Landau, Leese, & Stahl  
372 (2011).

373 By evaluating the CCC reported in Table S8 (Supplementary Materials) and taking into consideration that the  
374 closer is this index to 1, the higher is the degree of fit of clustering (Saraçlı et al., 2013) it was found that the  
375 use of the Euclidean inter-point distance and the Ward's aggregation method was the most performant HCA-  
376 based method both for bulk and packaged anchovies, with CCC values of 0.9888 and 0.9836, respectively (See  
377 Table S8, Supplementary Materials). Dendrograms resulting from the proposed methodology are shown in  
378 Fig. 2. As it can be observed, bulk anchovies (Fig. 2A) and packaged anchovies (Fig. 2B) were gathered into  
379 three major clusters at dissimilarity values of 65 % and 70 %, respectively. These three clusters mostly  
380 enclosed samples of the same origin, thus suggesting the existence of elemental patterns strong enough to  
381 reflect on the presence of geographical origin-driven groupings. Nevertheless, a few bulk samples from Croatia  
382 drifted apart from the others (Fig. 2A), as well as some samples from Croatia and Tunisia were mingled  
383 together in the second cluster of the dendrogram of packaged anchovies (Fig. 2B). Thus, Cantabrian products  
384 were in general better clustered than samples from Croatia and Tunisia which, on the contrary, were more  
385 connected each other. This is easily explained by the fact that the two fishing areas of anchovy from Croatia  
386 and Tunisia are neighbouring zones of the Mediterranean Sea (FAO fishing area 37.2.1 and 37.2.2,  
387 respectively). Therefore, these samples may share lots of compositional features linked to the similar  
388 environmental characteristics.

### 389 *3.2 Data mining for the geographic origin evaluation*

390 Considering that the main task of the present research was to verify the usefulness of the elemental profile of  
391 anchovy products to the development of rapid methods for geographical origin verification before and after  
392 the product packaging, different machine learning algorithms were explored to identify the best-suited one to  
393 this purpose.

394 Four decision trees algorithms (C5.0, CART, CHAID, and QUEST) trained on 70 % of the whole sample sets  
395 (270 samples per each anchovy set) were examined, in search of the most accurate models in identifying the

396 origin labels of bulk and packaged anchovy samples in the validation set. One of the main advantages of using  
397 decision tree is that the selection of the most important variables for classification is performed automatically  
398 during training stage. Therefore, computation time is reduced, while interpretability, accessibility and  
399 handiness of the models is improved.

400 Performance outcomes of the trained and validated classification models are summarised in Table S9  
401 (Supplementary Materials). The rate of classification accuracy of the samples of the training sets ranged  
402 between 89.9 % and 99.4 %, with better results shown up by C5.0 both for the bulk and the packaged anchovy  
403 dataset. When validation of the models was performed, 98.3 % of bulk samples was correctly classified by  
404 CHAID. As for the packaged products, the most accurate model for classifying anchovies of the validation set  
405 was found to be QUEST (96.0% accuracy) (Table S9, Supplementary Materials). Based on the accuracy  
406 outcomes obtained during the validation phase, CHAID and QUEST were selected as the most appropriate  
407 algorithms to classify the origin of bulk and packaged anchovy samples, respectively. A short summary of the  
408 outputs obtained by the application of the other algorithms is however reported in Supplementary Material  
409 (Supplementary Tables S10, S11).

410 By looking at the ranks of each predictive element selected by the four models (Fig. 3) it is possible to highlight  
411 that C5.0 and CHAID models extracted a lower number of attributes compared to CART and QUEST models.  
412 Li, B, P, K, As, Sr, Zr, Pd, Cd, Cs, and Ba were shared as predictors within the two anchovy datasets. By  
413 contrast, Sb and Pb were influent only for bulk products (Fig. 3A), while Ni e U were extracted only for  
414 packaged products. Interestingly, As emerged as the variable showing the highest impact for all the  
415 classification models of bulk anchovies (Fig. 3A) and for QUEST and CHAID models of packaged anchovies  
416 (Fig. 3B). B and Cs were instead found to be the most important attributes in the CART and C5.0 models of  
417 packaged anchovies, respectively (Fig. 3B).

418 Arsenic contamination of seawaters can be related to anthropogenic pollutant activities as well as to the natural  
419 geological characteristics of the area (Garellick, Jones, Dybowska, & Valsami-Jones, 2008). As an example,  
420 As (together with Cr, Cu, Hg, Mn, Ni, Pb, Se, and V) concentrations were reported to be higher in seawater  
421 where volcanic activities exist such as the Mediterranean Basin (Juncos et al., 2016; Zkeri, Aloupi, & Gaganis,  
422 2018). Moreover, the uptake of As by fish is influenced by several natural factors including water temperature  
423 and salinity, cooccurrence of phosphate, and seasonal differences of the distribution of the inorganic and

424 organic forms of As in the aquatic environments (Ferrante et al., 2019). Regardless the natural or anthropogenic  
425 nature of As and releasing sources, the reduced exchange of water in the Mediterranean Basin and, especially  
426 in the Adriatic Sea, can facilitate the accumulation of As in the environment (Ferrante et al., 2019). This can  
427 justify the higher amounts of As in anchovy from the Mediterranean Sea compared to Cantabrian (Atlantic  
428 Ocean) anchovy reported in the present work (see Fig. 1A, Fig. 1B). Moreover, in pelagic fish species from  
429 the Adriatic Sea higher amounts of As compared to other sampling zones was previously shown (Storelli &  
430 Marcotrigiano, 2004).

### 431 3.2.1 Decision tree by CHAID algorithm for origin authenticity of bulk anchovies

432 In accordance with the optimal accuracy results achieved in training and validation, CHAID model was found  
433 to be characterised by optimal accuracy (94.1 %), sensitivity (95.6 %), and specificity (97.4 %) values also  
434 when used to classify unlabelled samples of the bulk anchovy test set (Table 1). Therefore, the method used  
435 can be effectively considered powerful enough when the analytical goal is the identification of Tunisian,  
436 Cantabrian, and Croatian bulk anchovy products origins.

437 The architecture of the decision tree obtained is illustrated in Fig. 4. As it can be observed, the tree was a three-  
438 level structure, with a total of 19 decision nodes and 12 classification rules created by using 6 elements only.  
439 The decision rules generated from the root node were based on 5 concentration ranges of As, which was  
440 confirmed to be the most influent element for first sample discrimination by CHAID (see Fig. 3A). CHAID-  
441 decision trees are generally more complex than those generated by other technique since it relies on a multiway  
442 splitting principle, but the higher degree of segmentation can help reducing the tree depth and speed up the  
443 classification of samples.

444 Concentrations of As  $\leq 3.38$  mg kg<sup>-1</sup> classified Cantabrian anchovies just at level 1 with 100% probability. In  
445 general, decreasing As concentrations (from 8.16 mg kg<sup>-1</sup> downwards) together with increasing concentrations  
446 of K (from 4084 mg kg<sup>-1</sup> upwards) and P (from 5078 mg kg<sup>-1</sup> upwards) were associated to the highest  
447 probability of identifying Cantabrian samples. Tunisian samples were better classified by descending As  
448 concentration ranges coupled with higher Li ( $> 0.16$  mg kg<sup>-1</sup>), Sr ( $>29.75$  mg kg<sup>-1</sup>), or Cd amounts ( $> 0.07$   
449 mg kg<sup>-1</sup>). Finally, when P, Li, Cd got lower the occurrence of Croatian samples become more probable.

### 450 3.2.2 Decision tree by QUEST algorithm for origin authenticity of packaged anchovies

451 The QUEST-based decision trees applied to packaged anchovy was composed by 13 decisions nodes stratified  
452 into four levels. B was selected as the first binary splitting variable. The outcomes related to predictor  
453 importance reported in Fig. 3B (according to which B had the highest influence in prediction) were confirmed  
454 by analysing the splitting variables used to generate the QUEST decision tree, where B was just selected as  
455 the first binary splitting variable (Fig. 5). B value higher than 5.13 mg kg<sup>-1</sup> generated a leaf (final) node with  
456 91.3% probability of predicting samples originating from Tunisia. Globally, 100% probability of correctly  
457 identifying Croatian samples was reached at the last tree-level using the classification rule based on B, As, and  
458 Cd or the classification rule based on B, As, K, and Pd. With increased B (> 5.13 mg kg<sup>-1</sup>), As (> 7.14 mg kg<sup>-1</sup>),  
459 and K (> 5669 mg kg<sup>-1</sup>) concentrations, also the likelihood of recognising Cantabrian anchovies increased.  
460 The set of decision rules established by QUEST further proved to be reliable and effective for the classification  
461 of packaged anchovy origin, owing to the good ability in predicting the unknown origin of samples of the test  
462 set. Compared to other decision tree algorithm, QUEST ranked first in terms of accuracy (97.7 %), sensitivity  
463 (97.6 %), specificity (98.9 %), and precision (98.0 %) (Table 1).  
464 Even though transformed anchovy implicitly represents a complex processed foodstuff, it is important to stress  
465 that using decision trees may be the quickest and the most intelligible way to solve problems related to  
466 classification of foodstuffs.

#### 467 **4. Conclusions**

468 In this work, data mining techniques were applied to transformed anchovy products to verify whether the origin  
469 of fish could be identified through the elemental patterns measured by ICP-MS and direct mercury analysis.  
470 Different machine learning algorithms relying on the principle of decision trees were applied to data and  
471 classification rules to distinguish anchovy fish of Cantabrian Sea from Tunisian and Croatian anchovies were  
472 created. Firstly, differences of elemental composition between anchovies at two stages of the production chain  
473 were investigated, to verify whether misleading elemental inclusion from the manufacturing environment was  
474 introduced. After having excluded problematic elements based on literature review and direct comparison of  
475 bulk and packaged anchovy profile and after having explored the effective presence of fish clusters related to  
476 origin, C5.0, CART, CHAID and QUEST decision trees were trained. This way, the selection of the most

477 important variables and the identification of cut-off limits for each element concentration to describe a specific  
478 group of samples were performed in tandem.

479 The results obtained showed that the concentrations of 6 elements only (As, K, P, Li, Cd, and Sr) are required  
480 to identify the origin of anchovy fish under the form of bulk products using the CHAID algorithm. Arsenic  
481 was found to be the first sorting element, whose contribution to geographical origin differentiation was  
482 remarkably reflected in the ability of the decision tree to identify the unknown label of bulk fish with accuracy,  
483 specificity, sensitivity, and precision values above 93 % on average. The origin of the packaged anchovy  
484 products for sale, was better recognised by the set of classification rules generated by the QUEST algorithm.  
485 In this case, 5 elements were sufficient to achieve accuracy, sensitivity, specificity, and precision outcomes  
486 higher than 96 %. The splitting of samples into groups was driven by the predictive influence of B, followed  
487 by As, K, Cd, and Pd.

488 In view of the above results, decision tree-based methods applied to elemental profiles of fishery products after  
489 industrial processing might be postulated as an immediate and easy-to-handly procedure to figure out how the  
490 elemental composition can help in solving many actual challenges related to fish authenticity and commercial  
491 fraud. Moreover, the cost-effectiveness of the methodology, reached by doing away with the irrelevant  
492 elements, may finally disconnect this kind of applications from the scientific research and lead to the  
493 application in the primary and secondary production sectors.

494 Future research including anchovy fish obtained from different countries and production systems is however  
495 desirable not only to clarify the involvement of multiple environmental factors on the stability over space and  
496 time of element profile of processed fish products, but also for the creation and curation of databases storing  
497 and making available analytical data relating to the fish authenticity.

#### 498 **Appendix A. Supplementary Materials**

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

500 **Declaration of interest:** The authors declare that they have no known competing financial interests or  
501 personal relationships that could have appeared to influence the work reported in this paper.

502 **Data Availability:** The dataset generated during the current study is available from the corresponding author  
503 on reasonable request.

#### 504 **Acknowledgements**

505 The research was carried out within the project “Development of rapid tools for anchovies authentication”  
506 funded by Rizzoli Emanuelli S.p.a. (Parma, Italy). The authors also gratefully acknowledge the University of  
507 Pardubice for the financial support (project no. SGS\_2020\_002).

508 **References**

- 509 Barbosa, R. M., Nacano, L. R., Freitas, R., Batista, B. L., & Barbosa Jr, F. (2014). The use of decision trees  
510 and naive Bayes algorithms and trace element patterns for controlling the authenticity of free- range-  
511 pastured hens' eggs. *Journal of food science*, *79*(9), C1672-C1677. [https://doi.org/10.1111/1750-](https://doi.org/10.1111/1750-3841.12577)  
512 [3841.12577](https://doi.org/10.1111/1750-3841.12577)
- 513 Bekhit, A. E. D. A., Morton, J. D., & Dawson, C. O. (2008). Effect of processing conditions on trace  
514 elements in fish roe from six commercial New Zealand fish species. *Journal of Agricultural and Food*  
515 *Chemistry*, *56*(12), 4846–4853. <https://doi.org/10.1021/jf8005646>
- 516 Besteiro, I., Rodríguez, C. J., Tilve-Jar, C., & Pascual López, C. (2000). Selection of attributes for the  
517 sensory evaluation of anchovies during the ripening process. *Journal of Sensory Studies*, *15*(1), 65–77.  
518 <https://doi.org/10.1111/j.1745-459X.2000.tb00410.x>
- 519 Carrera, M., & Gallardo, J. M. (2017). Determination of the Geographical Origin of All Commercial Hake  
520 Species by Stable Isotope Ratio (SIR) Analysis. *Journal of Agricultural and Food Chemistry*, *65*(5),  
521 1070–1077. <https://doi.org/10.1021/acs.jafc.6b04972>
- 522 Cuadros-Rodríguez, L., Pérez-Castaño, E., & Ruiz-Samblás, C. (2016). Quality performance metrics in  
523 multivariate classification methods for qualitative analysis. *Trends in Analytical Chemistry*, *80*, 612–  
524 624. <https://doi.org/10.1016/j.trac.2016.04.021>
- 525 Czerner, M., Agustinelli, S. P., Guccione, S., & Yeannes, M. I. (2015). Effect of different preservation  
526 processes on chemical composition and fatty acid profile of anchovy (*Engraulis anchoita*). *International*  
527 *Journal of Food Sciences and Nutrition*, *66*(8), 887–894.  
528 <https://doi.org/10.3109/09637486.2015.1110687>
- 529 Da Costa, N. L., Ximenez, J. P. B., Rodrigues, J. L., Barbosa, F., & Barbosa, R. (2020). Characterization of  
530 Cabernet Sauvignon wines from California: determination of origin based on ICP-MS analysis and  
531 machine learning techniques. *European Food Research and Technology*, *246*(6), 1193–1205.  
532 <https://doi.org/10.1007/s00217-020-03480-5>
- 533 Depeters, E. J., Puschner, B., Taylor, S. J., & Rodzen, J. A. (2013). Can fatty acid and mineral compositions



534 of sturgeon eggs distinguish between farm-raised versus wild white ( *Acipenser transmontanus* )  
535 sturgeon origins in California ? Preliminary report. *Forensic Science International*, 229(1–3), 128–132.  
536 <https://doi.org/10.1016/j.forsciint.2013.04.003>

537 Drab, K., & Daszykowski, M. (2014). Clustering in analytical chemistry. *Journal of AOAC International*,  
538 97(1), 29–38. <https://doi.org/10.5740/jaoacint.SGEDrab>

539 Epova, E. N., Bérail, S., Zuliani, T., Malherbe, J., Sarthou, L., Valiente, M., & Donard, O. F. X. (2018).  
540 <sup>87</sup>Sr/<sup>86</sup>Sr isotope ratio and multielemental signatures as indicators of origin of European cured hams:  
541 The role of salt. *Food Chemistry*, 246(October 2017), 313–322.  
542 <https://doi.org/10.1016/j.foodchem.2017.10.143>

543 European Parliament Resolution of 14 January 2014 on the food crisis, fraud in the food chain and the  
544 control thereof (2013/2091(INI)). *Official Journal of the European Union*, C482, 22-30. Retrieved  
545 from: <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A52014IP0011%2801%29>

546 Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Cluster analysis (5th ed). Wiley Series in  
547 Probability and Statistics. ISBN: 978-0-470-97844-3

548 FAO (2020) The state of world fisheries and aquaculture. Sustainability in action. Rome, Italy.  
549 <https://doi.org/10.4060/ca9229en>

550 FAO (2018) Overview of food fraud in the fisheries sector. Text by Reilly, A. Fisheries and Aquaculture  
551 Circular No. 1165, p. I. Retrieved from: <http://www.fao.org/documents/card/en/c/I8791EN/>

552 Ferrante, M., Napoli, S., Grasso, A., Zuccarello, P., Cristaldi, A., & Copat, C. (2019). Systematic review of  
553 arsenic in fresh seafood from the Mediterranean Sea and European Atlantic coasts: A health risk  
554 assessment. *Food and Chemical Toxicology*, 126(September 2018), 322–331.  
555 <https://doi.org/10.1016/j.fct.2019.01.010>

556 Fragni, R., Trifirò, A., & Nucci, A. (2015). Towards the development of a multi-element analysis by ICP-  
557 TOF-MS for tracing the geographical origin of processed tomato products. *Food Control*, 48, 96–101.  
558 <https://doi.org/10.1016/j.foodcont.2014.04.027>

559 Franke, B. M., Haldimann, M., Gremaud, G., Bosset, J. O., Hadorn, R., & Kreuzer, M. (2008). Element

560 signature analysis: Its validation as a tool for geographic authentication of the origin of dried beef and  
561 poultry meat. *European Food Research and Technology*, 227(3), 701–708.  
562 <https://doi.org/10.1007/s00217-007-0776-8>

563 Garelick, H., Jones, H., Dybowska, A., & Valsami-Jones, E. (2008). Arsenic pollution sources. *Reviews of*  
564 *Environmental Contamination and Toxicology*, 197(January 2017), 17–60. [https://doi.org/10.1007/978-](https://doi.org/10.1007/978-0-387-79284-2_2)  
565 [0-387-79284-2\\_2](https://doi.org/10.1007/978-0-387-79284-2_2)

566 Han, J., Kamber, M., & Pei, J. (2011). *Data Mining : Concepts and Techniques* (Third Edit). Morgan  
567 Kaufmann Publishers. <https://doi.org/10.1016/B978-0-12-381479-1.00001-0>

568 Hernandez-Herrero, M. M., Roig-Sagués, A. X., López-Sabater, E. I., Rodríguez-Jerez, J. J., & Mora-  
569 Ventura, M. T. (1999). Total Volatile Basic Nitrogen and other Physico- chemical and Microbiological  
570 Characteristics as. *Journal of Food Science*, 64(2), 344–347.

571 Husáková, L., Urbanová, I., Šrámková, J., Černohorský, T., Krejčová, A., Bednaříková, M., Frýdová, E.,  
572 Nedělková, I., & Pilařová, L. (2011). Analytical capabilities of inductively coupled plasma orthogonal  
573 acceleration time-of-flight mass spectrometry (ICP-oe-TOF-MS) for multi-element analysis of food and  
574 beverages. *Food Chemistry*, 129(3), 1287–1296. <https://doi.org/10.1016/j.foodchem.2011.05.047>

575 Husáková, L., Urbanová, I., Šídová, T., Cahová, T., Faltys, T., & Šrámková, J. (2015). Evaluation of  
576 ammonium fluoride for quantitative microwave-assisted extraction of silicon and boron from different  
577 solid samples. *International Journal of Environmental Analytical Chemistry*, 95(10), 922–935.  
578 <https://doi.org/10.1080/03067319.2015.1070409>

579 Ikem, A., & Egiebor, N. O. (2005). Assessment of trace elements in canned fishes (mackerel, tuna, salmon,  
580 sardines and herrings) marketed in Georgia and Alabama (United States of America). *Journal of Food*  
581 *Composition and Analysis*, 18(8), 771–787. <https://doi.org/10.1016/j.jfca.2004.11.002>

582 Juncos, R., Arcagni, M., Rizzo, A., Campbell, L., Arribére, M., & Guevara, S. R. (2016). Natural origin  
583 arsenic in aquatic organisms from a deep oligotrophic lake under the influence of volcanic eruptions.  
584 *Chemosphere*, 144, 2277–2289. <https://doi.org/10.1016/j.chemosphere.2015.10.092>

585 Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4), 261-283.

- 586 <https://doi.org/10.1007/s10462-011-9272-4>
- 587 Laso, J., Margallo, M., Fullana, P., Bala, A., Gazulla, C., Irabien, A., & Aldaco, R. (2017). Introducing life  
588 cycle thinking to define best available techniques for products: Application to the anchovy canning  
589 industry. *Journal of Cleaner Production*, 155, 139–150. <https://doi.org/10.1016/j.jclepro.2016.08.040>
- 590 Li, L., Han, C., Dong, S., & Boyd, C. E. (2019). Use of elemental profiling and isotopic signatures to  
591 differentiate Pacific white shrimp (*Litopenaeus vannamei*) from freshwater and seawater culture areas.  
592 *Food Control*, 95, 249–256. <https://doi.org/10.1016/j.foodcont.2018.08.015>
- 593 Lo Feudo, G., Naccarato, A., Sindona, G., & Tagarelli, A. (2010). Investigating the origin of tomatoes and  
594 triple concentrated tomato pastes through multielement determination by inductively coupled plasma  
595 mass spectrometry and statistical analysis. *Journal of Agricultural and Food Chemistry*, 58(6), 3801–  
596 3807. <https://doi.org/10.1021/jf903868j>
- 597 Magdas, D. A., Feher, I., Cristea, G., Voica, C., Tabaran, A., Mihaiu, M., Cordea, D. V., Bâlțeanu, V. A., &  
598 Dan, D. S. (2019). Geographical origin and species differentiation of Transylvanian cheese.  
599 Comparative study of isotopic and elemental profiling vs . DNA results. *Food Chemistry*, 277(June  
600 2018), 307–313. <https://doi.org/10.1016/j.foodchem.2018.10.103>
- 601 Moreno-Rojas, R., Cámara-Martos, F., Sánchez-Segarra, P. J., & Amaro-López, M. Á. (2012). Influence of  
602 manufacturing conditions and discrimination of Northern Spanish cheeses using multi-element  
603 analysis. *International Journal of Dairy Technology*, 65(4), 594–602. [https://doi.org/10.1111/j.1471-  
604 0307.2012.00853.x](https://doi.org/10.1111/j.1471-0307.2012.00853.x)
- 605 Rodushkin, I., Bergman, T., Douglas, G., Engström, E., Sörlin, D., & Baxter, D. C. (2007). Authentication of  
606 Kalix (N.E. Sweden) vendace caviar using inductively coupled plasma-based analytical techniques:  
607 Evaluation of different approaches. *Analytica Chimica Acta*, 583(2), 310–318.  
608 <https://doi.org/10.1016/j.aca.2006.10.038>
- 609 Rokach, L., & Maimon, O. Z. (2008). *Data mining with decision trees. Theory and applications* (2nd ed.).  
610 World Scientific Publishing Co. Pte. Ltd.
- 611 Saraçlı, S., Doğan, N., & Doğan, İ. (2013). Comparison of hierarchical cluster analysis methods by

612 cophenetic correlation. *Journal of Inequalities and Applications*, 203. <https://doi.org/10.1186/1029->  
613 242X-2013-203

614 Smith, R. G., & Watts, C. A. (2009). Determination of the country of origin of farm-raised shrimp (family  
615 penaeide) using trace metal profiling and multivariate statistics. *Journal of Agricultural and Food*  
616 *Chemistry*, 57(18), 8244–8249. <https://doi.org/10.1021/jf901658f>

617 Sokal, R. R., & Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, 33-40.

618 Storelli, M. M., & Marcotrigiano, G. O. (2004). Interspecific variation in total arsenic body concentrations in  
619 elasmobranch fish from the Mediterranean Sea. *Marine Pollution Bulletin*, 48(11–12), 1145–1149.  
620 <https://doi.org/10.1016/j.marpolbul.2004.03.005>

621 Suhaj, M., & Kore, M. (2008). Study of some European cheeses geographical traceability by pattern  
622 recognition analysis of multielemental data. *European Food Research and Technology*, 227, 1419–  
623 1427. <https://doi.org/10.1007/s00217-008-0861-7>

624 Turra, C., Dias de Lima, M., Fernandes, E. A. D. N., Bacchi, M. A., Barbosa, F., & Barbosa, R. (2017).  
625 Multielement determination in orange juice by ICP-MS associated with data mining for the  
626 classification of organic samples. *Information Processing in Agriculture*, 4(3), 199–205.  
627 <https://doi.org/10.1016/j.inpa.2017.05.004>

628 Varrà, M. O., Ghidini, S., Zanardi, E., Badiani, A., & Ianieri, A. (2019). Authentication of European sea bass  
629 according to production method and geographical origin by light stable isotope ratio and rare earth  
630 elements analyses combined with chemometrics. *Italian Journal of Food Safety*, 8(1).  
631 <https://doi.org/10.4081/ijfs.2019.7872>

632 Velasco, A., Aldrey, A., Pérez-Martín, R. I., & Sotelo, C. G. (2016). Assessment of the labelling accuracy of  
633 spanish semipreserved anchovies products by FINS (forensically informative nucleotide sequencing).  
634 *Heliyon*, 2(6), e00124. <https://doi.org/10.1016/j.heliyon.2016.e00124>

635 Zkeri, E., Aloupi, M., & Gaganis, P. (2018). Seasonal and spatial variation of arsenic in groundwater in a  
636 rhyolitic volcanic area of Lesvos Island, Greece. *Environmental Monitoring and Assessment*, 190, 44.  
637 <https://doi.org/10.1007/s10661-017-6395-3>

638 **Figure Captions**

639 **Fig. 1.** Beeswarm box-plots with Kruskal-Wallis and Dunn's multiple comparison test results (median and  
640 quartiles) showing elements in bulk (A) and packaged (B) anchovy products varying significantly in relation  
641 to the origin ( $p < 0.05$ ).

642

643 **Fig. 2.** Hierarchical cluster analysis simplified dendrograms for bulk anchovy dataset (A) and packaged  
644 anchovy dataset (B) based on 35 elements.

645 **Fig. 3.** Comparison of the most important elemental predictors in C5.0, CART, CHAID, and QUEST models  
646 for bulk anchovies (A) and packaged anchovies (B). Values are scaled from 0 (no influence) to 1 (maximum  
647 influence).

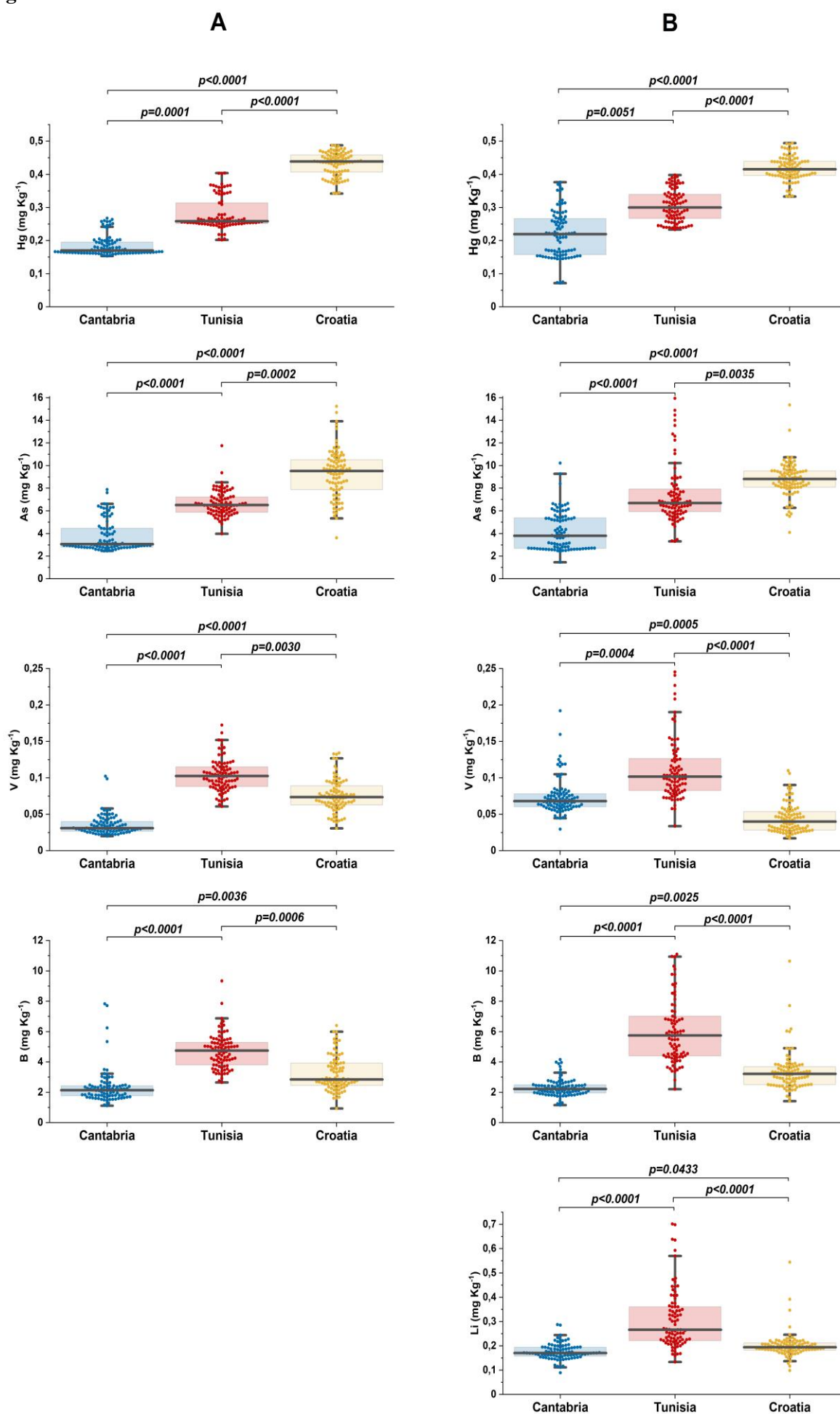
648 **Fig. 4.** Decision classification tree resulting from the application of the CHAID algorithm for the classification  
649 of bulk anchovies using the element profile. Concentrations are reported in  $\mu\text{g kg}^{-1}$ .

650 **Fig. 5.** Decision classification tree resulting from the application of the QUEST algorithm for the classification  
651 of packaged anchovies using the element profile. Concentrations are reported in  $\mu\text{g kg}^{-1}$ .

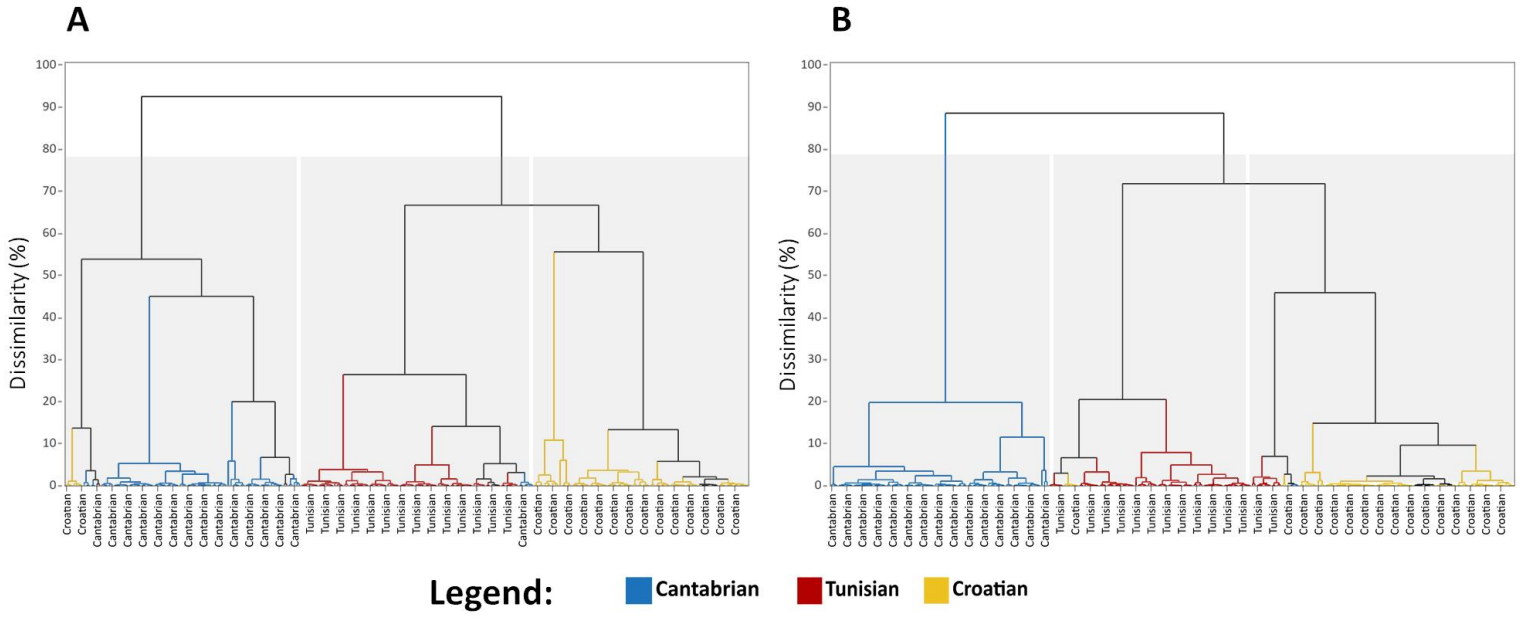
652

653

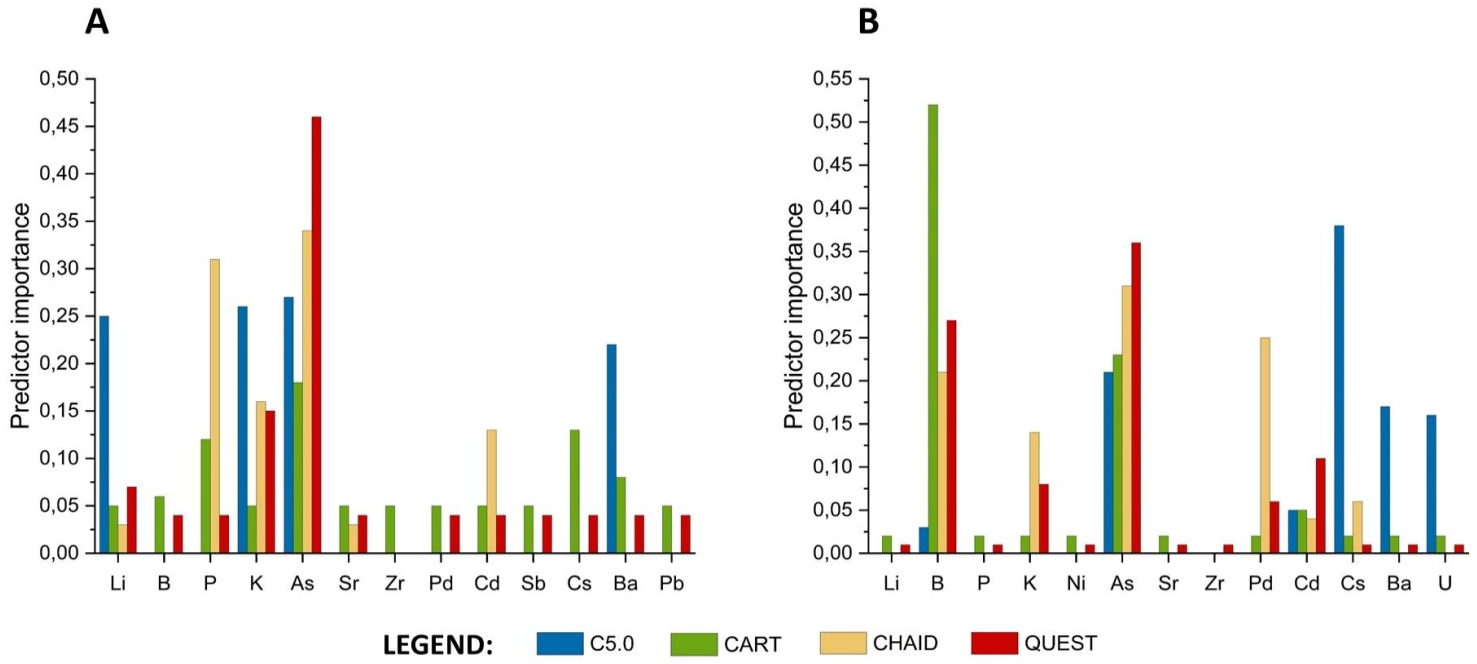
Fig. 1



**Fig. 2.**



**Fig.3.**





**Fig. 4.**

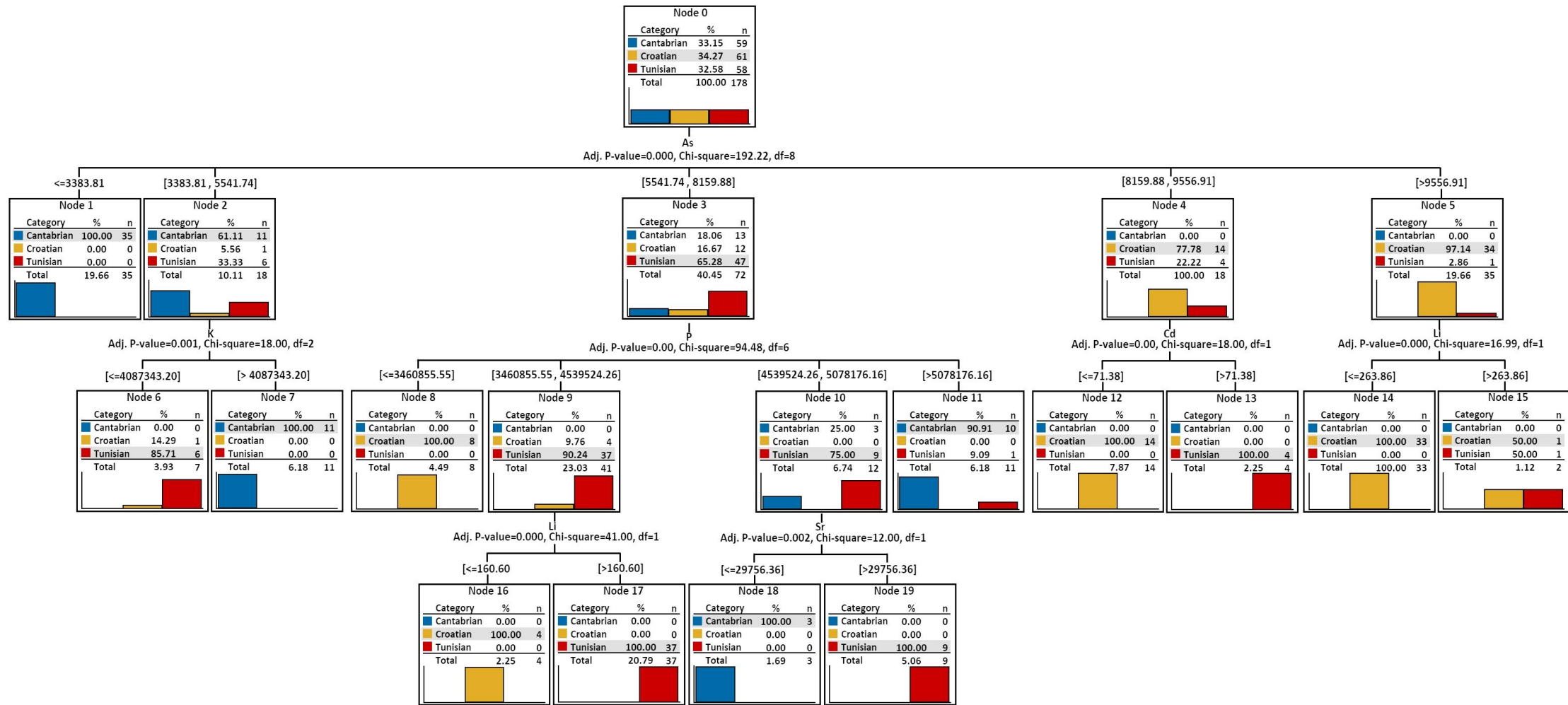
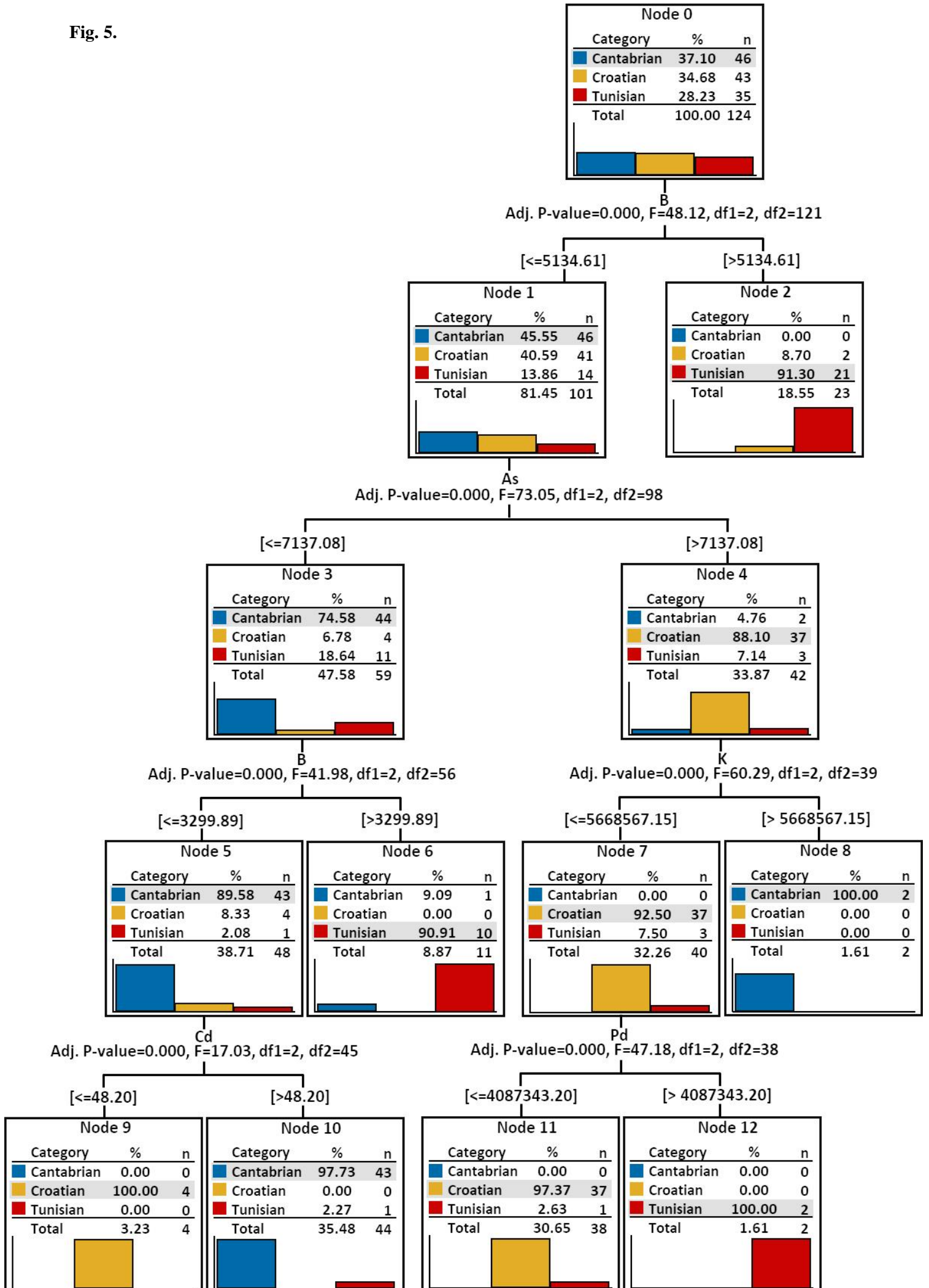


Fig. 5.



**Table 1**

Summary of performance parameters of decision tree models for classification of anchovy products of the testing datasets.

Performance index	Bulk anchovies				Packaged anchovies			
	C5.0	CHAID	CART	QUEST	C5.0	CHAID	CART	QUEST
Accuracy (%)	91.2	94.1	91.2	94.1	90.5	91.9	96.8	97.7
Sensitivity (%)	91.4	95.6	89.4	93.6	92.6	92.3	96.7	97.6
Specificity (%)	95.3	97.4	95.1	96.7	95.4	95.5	98.5	98.9
Precision (%)	90.7	93.3	92.6	95.0	92.2	92.7	96.7	97.9
F-score (%)	90.9	93.9	90.2	94.1	91.4	91.8	96.5	97.7

**Declaration of interests**

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

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

## **CRedit authorship contribution statement**

**Maria Olga Varra:** Writing – original draft, Investigation, Data curation, Formal analysis, Validation, Resources. **Lenka Husáková:** Writing - review & editing, Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Supervision, Funding acquisition. **Jan Patočka:** Methodology, Investigation, Validation, Formal analysis. **Emanuela Zanardi:** Conceptualization, Writing - review & editing, Project administration, Funding acquisition, Supervision. **Sergio Ghidini:** Resources, Conceptualization.