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

Maria Olga VARRÀ[a], *Lenka HUSÁKOVÁ*[b*], *Jan PATOČKA*[b], *Sergio GHIDINI*[a],

Emanuela ZANARDI[a*]

[a]***Department of Food and Drug, University of Parma, Parma, Via del Taglio, 10, Parma 43126, Italy***

E-mail addresses: emanuela.zanardi@unipr.it (E. Zanardi); sergio.ghidini@unipr.it (S. Ghidini); mariaolga.varra@studenti.unipr.it (M. O. Varrà).

[b]***Department of Analytical Chemistry, Faculty of Chemical Technology, University of Pardubice, Studentská 573 HB/D, Pardubice, CZ-532 10, Czech Republic***

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

***CORRESPONDING AUTHORS:**

Lenka Husáková (L. Husáková)

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

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

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

Emanuela Zanardi (E. Zanardi)

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

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

ABSTRACT

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

Abbreviations

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

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

1. Introduction

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

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

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

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

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

The traditional anchovy transformation process by brine-ripening finds a long tradition in southern Europe. The fish, typically caught by purse seines, is quickly transported to the fish canning industry where it is beheaded, partially eviscerated and punt into ripening containers (barrels), alternating layers of fish and salt. A pressure is then applied on the top layer to facilitate the progressive elimination of water. The fish is ripened until the desired degree of maturation is reached (from 3 up to 11-12 month on average) to then be moved from the barrels. From that point on, the bulk ripened anchovies can be preserved and packaged in salt to be commercialised or further processed to obtain different products and preparations, for example by filleting and packaging-in-oil.

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

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

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

Tracing back to the origin of processed or highly processed products is considerably difficult because of the manipulation and the addition of several compounds during preparation procedures. The use of salt during anchovy manufacturing may represent the most critical point since it can potentially mask the natural elemental content of fish. Nevertheless, the multiple identification of elements using techniques such as inductively coupled plasma-optical emission spectroscopy (ICP-OES) and inductively coupled plasma-mass spectrometry (ICP-MS) have been already successfully applied to identify the origin of transformed food products such processed tomato products (Lo Feudo, Naccarato, Sindona, & Tagarelli, 2010; Fragni, Trifirò, & Nucci, 2015), fruit juices (Turra et al., 2017), wines (da Costa, Ximenez, Rodrigues, Barbosa, & Barbosa, 2020), dried beef (Franke et al., 2008), hams (Epova et al., 2018), and different types of cheese (Suhaj & Kore, 2008; Moreno-Rojas, Cámara-Martos, Sánchez-Segarra, & Amaro-López, 2012; Magdas et al., 2019). One application dealing with the use of multi-elemental analysis to authenticate seafood products is also available and it concerns the identification of caviar from different origins, which, as anchovy, is a salted product (Rodushkin et al., 2007).

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

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

2. Materials and Methods

2.1. Chemicals, standards, and reference materials

All the aqueous solutions employed for analyses were prepared using ultrapure water (0.05 µS cm^{-1}) obtained by the Milli-Q® water purification system (Millipore, Bedford, USA).

For microwave digestion, hydrogen peroxide (H_2O_2, ≥ 30% w/w) for ultra-trace analysis (Fluka Chemie AG, Buchs, Switzerland) and sub-boiled nitric acid prepared from nitric acid (65%, w/w, Selectipur quality, Lach-Ner, Neratovice, Czech Republic) by means of the sub-boiling distillation apparatus Distillacid™ BSB-939-IR (Berghof, Eningen, Germany) were used.

The working calibration solutions for ICP-MS analysis were prepared daily using the multi-element stock solutions “A”, “B1”, “B2”, and “C”. Stock solution “A” (10 mg L^{-1} of Li, B, Al, V, Cr, Fe, Mn, Ni, Cu, Zn, 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 Supelco ICP multi-element standard solution IV (Merck, Darmstadt, Germany) (containing 1 g L^{-1} of Li, B, Al, Cr, Fe, Mn, Ni, Cu, Zn, Co, Sr, Cd, Ba, Tl, Pb, and Bi) and single element standards (V, As, Se, Rb, Zr, Mo, Ru, Pd, Sn, Sb, Cs, Hf, Re, Pt, and Th) of concentration 1 ± 0.002 g L^{-1} (Analytika Ltd., Prague, Czech Republic or SCP Science, Montreal, Canada).

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

A 1 g L^{-1} stock solution of Rh (SCP Science, Montreal, Canada) was used to prepare the internal standard solution (ISTD) at concentration of 200 µg L^{-1}.

A 10 g L^{-1} stock solution prepared from urea (TraceSelect quality, Fluka Chemie AG, Buchs, Switzerland) was used to prepare carbon reference solutions.

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

2.2. Anchovy sampling and processing

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

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

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

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

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

2.2.1 Lyophilisation process

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

2.2.2 Microwave-assisted acid digestion

For subsequent ICP-MS analysis, 0.1 g of freeze-dried samples or CRMs were weighted (in triplicate) into a 10 mL perfluoroalkoxy (PFA) tube and 4 mL of 16% HNO_3 (65%, w/w HNO_3, 1:3 diluted) and 1 mL of 30% H_2O_2 were added. Three PFA tubes were placed into DAC-100S polytetrafluoroethylene vessels (Berghof, Eningen, Germany) previously filled with 25 mL of HNO_3 (16%, v/v), by ensuring that the level of liquid in the outer polytetrafluoroethylene vessel was higher than those in the PFA tubes. This way, the vapor pressures were compensated and the evaporation of the solution from the PFA tubes was avoided (Husáková et al., 2015).

Samples were decomposed using a Berghof Speedwave™ MWS-3$^+$ microwave digestion system (Berghof, Germany) with the maximum total output of the microwave generator (1450 W) via the following multistep 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 100 °C (ramp 5 min).

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

2.3. Mercury analysis

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

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

2.4. ICP-MS analysis

Element quantification in samples was performed by using an Agilent 7900 quadrupole ICP-MS apparatus (Agilent Technologies, Inc., Santa Clara, CA, USA) equipped with a quartz concentric nebulizer MicroMist (400 µL min^{-1}), the Peltier-cooled (2 °C) Scott quartz spray chamber, quartz torch with 2.5 mm internal diameter injector, standard sampling and skimmer nickel cones with orifices of 1 and 0.45 mm, and an octopole collision/reaction cell for interference removal using kinetic energy discrimination. A low-pulsation, 10-roller peristaltic pump with three separate channels was employed to precisely deliver both samples and ISTD. For the simultaneous ISTD aspiration and mixing with the sample the connecting tubing, connectors, and the "Y" piece (included into internal standard kit) were employed. Analytical conditions were enhanced before starting sample measurement by using the multi-elemental tuning solution (Agilent Technologies, Inc., Santa Clara, CA, USA) containing 1 µg L^{-1} of Ce, Co, Li, Mg, Tl and Y, to obtain the highest possible sensitivity for elements of low, middle, and high m/z. Using the typical operating conditions summarised in Table S2 (Supplementary Materials), a sensitivity of 6000 counts s^{-1} per µg L^{-1} and a resolution of 0.64 amu peak width (full width at half maximum intensity) were achieved for $^{7}Li^{+}$. The same parameters were 50000 counts s^{-1} per µg L^{-1} and 0.62 for $^{89}Y^{+}$, and 30000 counts s^{-1} per µg L^{-1} and 0.60 for $^{205}Tl^{+}$.

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

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

The calibration curves for the quantification of 51 elements (R^{2} >0.999) resulted from the acquisition of working calibration solutions prepared from multi-element stock solutions "A", "B1", "B2", and "C" described in Section 2.1. The concentration of elements for calibration were as follows: blank, 1, 5, 10, 50, 100 µg L^{-1} of 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, Hf, Ta, Re, Pt, Tl, Pb, Bi, Th; 0.1, 0.5, 1, 5, 10 µg L^{-1} of La, Ce, Pr, Nd, U; 0.02, 0.1, 0.2, 1, 2 µg L^{-1} of Y, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu; 0.5, 1, 5, 10 mg L^{-1} of Na, Mg, P, K, Ca, Mn, Cu, Zn.

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

2.5. Analytical validation

The accuracy of the methodology was checked by analysing the five certified reference materials listed in *Section 2.1* (NIST 1566 Oyster Tissue, BCR CRM 184 Bovine muscle, BCR CRM 185 Bovine Liver, CRM 12-2-01 Bovine Liver, NIST SRM 1577 Bovine Liver) intended for the evaluation of analytical methods and instruments and used for the determination of the mass fraction values of selected elements in marine tissue, foods, or similar materials. In addition, CRM 12-2-03 Essential and toxic elements in Lucerne, NCS ZC 73015 Milk Powder, and CRM 12-2-04 Wheat bread flour, were analysed to assess accuracy in determining lanthanides and actinides. The high level of agreement between target and found values demonstrated trueness of data obtained (Supplementary Materials, Table S3).

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

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

2.6. Data elaboration

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

Univariate data analysis, consisting of nonparametric Mann-Whitney test ($p \leq 0.05$) and Kruskal–Wallis test 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 × 3 replicates × 3 provenances) and the set of packaged anchovies (30 samples × 3 replicates × 3 provenances) to investigate any significant difference between groups of samples. Nonparametric tests were chosen instead of classical parametric ones because most of the data presented non normal distribution and heteroscedasticity, as verified though the application of Shapiro-Wilk's and Levene's tests, respectively ($p \leq 0.05$).

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

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

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

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

The output of the method is a set of concatenated classification rules in the form of a decision tree composed by nodes (identifying the features that need to be sorted) and branches (identifying the values assumed by nodes) (Han, Kamber, & Pei, 2011). CART, C5.0, CHAID, and QUEST represent different methods by which the architecture of the decision trees can be built up and differ each other mainly for the segmentation rules applied and the tree optimisation method. The C5.0 algorithm is based on binary splitting and works by choosing progressively the instances that allow to gain the maximum partitioning information and stopping via the pruning rule, i.e., by removing from the splits which do not add significant information. CART algorithm is also based on binary splitting but differs from C5.0 essentially for a different stopping rule in the creation of the tree, consisting of the evaluation of the purity of the node, i.e. the maximum degree of homogeneity between categories. The CHAID algorithm is a multiway splitting system based on Chi-square statistics to decide for tree ramifications and is based on the measures of the impurity of the nodes. Finally, the QUEST is a binary-split algorithm which, in the case of continuous variables, uses an ANOVA F-Test to create tree nodes. Further information can be retrieved from Han, Kamber, & Pei (2011) and from Rokach & Maimon (2008).

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

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

3. Results and discussion

3.1. Initial data evaluation

3.1.1 Global elemental profiles of transformed anchovy products

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

Complete data matrices reporting the whole concentrations of the 52 elements measured in bulk and packaged anchovies can be found in Table S5 and Table S6 (Supplementary Materials), respectively. The most abundant element was found to be Na (whose median concentrations ranged from 142 mg kg^{-1} in bulk Croatian anchovies to 177 mg kg^{-1} in Tunisian packaged anchovies) followed by P, K, Ca, and Mg. Bulk and packaged products from Cantabrian Sea differed from samples of Mediterranean origin (Croatian and Tunisian) because of significantly higher concentrations of P and K. At the same time, no differences in Na and Ca contents between Cantabrian and Croatian anchovies was encountered, which, instead, were both significantly lower compared to those found in samples originating from Tunisia (Table S5, Table S6, Supplementary Materials). Similarly, some minor, trace- and ultra-trace elements showed significant variations in relation to the provenance Besides, Cantabrian anchovies were characterised by higher Ni, Mo, Cs, and Tl concentrations and lower Al concentrations than those encountered in the two Mediterranean products (Supplementary Tables S5, S6).

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

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

3.1.2 Pre-selection of elements as potential indicators of origin

The use of salt in the form of saturated brine during the processing and packaging of ripened anchovies is one of the main complicating factors which might limit the proper identification of markers of geographical origin since it inevitably modifies the natural element profile of the product. For instance, the use of a brine for fermentation purposes was reported to decrease the total amount of certain trace elements in transformed fish roe because of partitioning phenomena between the solid and the liquid phases occurring during fermentation (Bekhit, Morton, & Dawson, 2008).

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

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

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

3.1.3. Hierarchical cluster analysis (HCA)

The inner potential of the elemental profile in guiding the creation of groups of anchovies was at first instance explored. Any possible natural difference or similarity among anchovies of both datasets was therefore uncovered by the application of HCA. Since different methods to measure distances (i.e. Euclidean and Manhattan distances) and to perform grouping (i.e. Ward's minimum variance, single linkage, complete linkage, simple average group average, median, and centroid clustering methods) are applicable for HCA building, the cophenetic correlation coefficient (CCC) was employed as an index to compare the methods with each other and to evaluate the validity of the resulting dendrograms (Sokhal & Rohlf, 1962; Saraçli, Doğan, & Doğan, 2013). Further details about the tested methods can be retrieved from Everitt, Landau, Leese, & Stahl (2011).

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

3.2 Data mining for the geographic origin evaluation

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

Four decision trees algorithms (C5.0, CART, CHAID, and QUEST) trained on 70 % of the whole sample sets (270 samples per each anchovy set) were examined, in search of the most accurate models in identifying the origin labels of bulk and packaged anchovy samples in the validation set. One of the main advantages of using decision tree is that the selection of the most important variables for classification is performed automatically during training stage. Therefore, computation time is reduced, while interpretability, accessibility and handiness of the models is improved.

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

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

Arsenic contamination of seawaters can be related to anthropogenic pollutant activities as well as to the natural geological characteristics of the area (Garelick, Jones, Dybowska, & Valsami-Jones, 2008). As an example, As (together with Cr, Cu, Hg, Mn, Ni, Pb, Se, and V) concentrations were reported to be higher in seawater where volcanic activities exist such as the Mediterranean Basin (Juncos et al., 2016; Zkeri, Aloupi, & Gaganis, 2018). Moreover, the uptake of As by fish is influenced by several natural factors including water temperature and salinity, cooccurrence of phosphate, and seasonal differences of the distribution of the inorganic and organic forms of As in the aquatic environments (Ferrante et al., 2019). Regardless the natural or anthropogenic nature of As and releasing sources, the reduced exchange of water in the Mediterranean Basin and, especially in the Adriatic Sea, can facilitate the accumulation of As in the environment (Ferrante et al., 2019). This can justify the higher amounts of As in anchovy from the Mediterranean Sea compared to Cantabrian (Atlantic Ocean) anchovy reported in the present work (see Fig. 1A, Fig. 1B). Moreover, in pelagic fish species from the Adriatic Sea higher amounts of As compared to other sampling zones was previously shown (Storelli & Marcotrigiano, 2004).

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

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

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

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

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

The QUEST-based decision trees applied to packaged anchovy was composed by 13 decisions nodes stratified into four levels. B was selected as the first binary splitting variable. The outcomes related to predictor importance reported in Fig. 3B (according to which B had the highest influence in prediction) were confirmed by analysing the splitting variables used to generate the QUEST decision tree, where B was just selected as the first binary splitting variable (Fig. 5). B value higher than 5.13 mg kg^{-1} generated a leaf (final) node with 91.3% probability of predicting samples originating from Tunisia. Globally, 100% probability of correctly identifying Croatian samples was reached at the last tree-level using the classification rule based on B, As, and Cd or the classification rule based on B, As, K, and Pd. With increased B (> 5.13 mg kg^{-1}), As (> 7.14 mg kg^{-1}), and K (> 5669 mg kg^{-1}) concentrations, also the likelihood of recognising Cantabrian anchovies increased. The set of decision rules established by QUEST further proved to be reliable and effective for the classification of packaged anchovy origin, owing to the good ability in predicting the unknown origin of samples of the test set. Compared to other decision tree algorithm, QUEST ranked first in terms of accuracy (97.7 %), sensitivity (97.6 %), specificity (98.9 %), and precision (98.0 %) (Table 1).

Even though transformed anchovy implicitly represents a complex processed foodstuff, it is important to stress that using decision trees may be the quickest and the most intelligible way to solve problems related to classification of foodstuffs.

4. Conclusions

In this work, data mining techniques were applied to transformed anchovy products to verify whether the origin of fish could be identified through the elemental patterns measured by ICP-MS and direct mercury analysis. Different machine learning algorithms relying on the principle of decision trees were applied to data and classification rules to distinguish anchovy fish of Cantabrian Sea from Tunisian and Croatian anchovies were created. Firstly, differences of elemental composition between anchovies at two stages of the production chain were investigated, to verify whether misleading elemental inclusion from the manufacturing environment was introduced. After having excluded problematic elements based on literature review and direct comparison of bulk and packaged anchovy profile and after having explored the effective presence of fish clusters related to origin, C5.0, CART, CHAID and QUEST decision trees were trained. This way, the selection of the most important variables and the identification of cut-off limits for each element concentration to describe a specific group of samples were performed in tandem.

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

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

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

Appendix A. Supplementary Materials

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

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

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

Acknowledgements

The research was carried out within the project “Development of rapid tools for anchovies authentication” funded by Rizzoli Emanuelli S.p.a. (Parma, Italy). The authors also gratefully acknowledge the University of Pardubice for the financial support (project no. SGS_2020_002).

References

Barbosa, R. M., Nacano, L. R., Freitas, R., Batista, B. L., & Barbosa Jr, F. (2014). The use of decision trees and naive Bayes algorithms and trace element patterns for controlling the authenticity of free- range-pastured hens' eggs. Journal of food science, 79(9), C1672-C1677. https://doi.org/10.1111/1750-3841.12577

Bekhit, A. E. D. A., Morton, J. D., & Dawson, C. O. (2008). Effect of processing conditions on trace elements in fish roe from six commercial New Zealand fish species. *Journal of Agricultural and Food Chemistry*, *56*(12), 4846–4853. https://doi.org/10.1021/jf8005646

Besteiro, I., Rodríguez, C. J., Tilve-Jar, C., & Pascual López, C. (2000). Selection of attributes for the sensory evaluation of anchovies during the ripening process. *Journal of Sensory Studies*, *15*(1), 65–77. https://doi.org/10.1111/j.1745-459X.2000.tb00410.x

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

Cuadros-Rodríguez, L., Pérez-Castaño, E., & Ruiz-Samblás, C. (2016). Quality performance metrics in multivariate classification methods for qualitative analysis. *Trends in Analytical Chemistry*, *80*, 612–624. https://doi.org/10.1016/j.trac.2016.04.021

Czerner, M., Agustinelli, S. P., Guccione, S., & Yeannes, M. I. (2015). Effect of different preservation processes on chemical composition and fatty acid profile of anchovy (Engraulis anchoita). *International Journal of Food Sciences and Nutrition, 66(8),* 887–894. https://doi.org/10.3109/09637486.2015.1110687

Da Costa, N. L., Ximenez, J. P. B., Rodrigues, J. L., Barbosa, F., & Barbosa, R. (2020). Characterization of Cabernet Sauvignon wines from California: determination of origin based on ICP-MS analysis and machine learning techniques. *European Food Research and Technology, 246(6),* 1193–1205. https://doi.org/10.1007/s00217-020-03480-5

Depeters, E. J., Puschner, B., Taylor, S. J., & Rodzen, J. A. (2013). Can fatty acid and mineral compositions of sturgeon eggs distinguish between farm-raised versus wild white (Acipenser transmontanus) sturgeon origins in California ? Preliminary report. *Forensic Science International*, *229*(1–3), 128–132. https://doi.org/10.1016/j.forsciint.2013.04.003

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

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

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

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

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

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

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

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

Franke, B. M., Haldimann, M., Gremaud, G., Bosset, J. O., Hadorn, R., & Kreuzer, M. (2008). Element signature analysis: Its validation as a tool for geographic authentication of the origin of dried beef and poultry meat. *European Food Research and Technology, 227(3),* 701–708. https://doi.org/10.1007/s00217-007-0776-8

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

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

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

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

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

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

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

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

https://doi.org/10.1007/s10462-011-9272-4

Laso, J., Margallo, M., Fullana, P., Bala, A., Gazulla, C., Irabien, A., & Aldaco, R. (2017). Introducing life cycle thinking to define best available techniques for products: Application to the anchovy canning industry. *Journal of Cleaner Production*, *155*, 139–150. https://doi.org/10.1016/j.jclepro.2016.08.040

Li, L., Han, C., Dong, S., & Boyd, C. E. (2019). Use of elemental profiling and isotopic signatures to differentiate Pacific white shrimp (Litopenaeus vannamei) from freshwater and seawater culture areas. *Food Control*, *95*, 249–256. https://doi.org/10.1016/j.foodcont.2018.08.015

Lo Feudo, G., Naccarato, A., Sindona, G., & Tagarelli, A. (2010). Investigating the origin of tomatoes and triple concentrated tomato pastes through multielement determination by inductively coupled plasma mass spectrometry and statistical analysis. *Journal of Agricultural and Food Chemistry*, *58*(6), 3801–3807. https://doi.org/10.1021/jf903868j

Magdas, D. A., Feher, I., Cristea, G., Voica, C., Tabaran, A., Mihaiu, M., Cordea, D. V., Bâlteanu, V. A., & Dan, D. S. (2019). Geographical origin and species differentiation of Transylvanian cheese. Comparative study of isotopic and elemental profiling vs . DNA results. *Food Chemistry*, *277*(June 2018), 307–313. https://doi.org/10.1016/j.foodchem.2018.10.103

Moreno-Rojas, R., Cámara-Martos, F., Sánchez-Segarra, P. J., & Amaro-López, M. Á. (2012). Influence of manufacturing conditions and discrimination of Northern Spanish cheeses using multi-element analysis. *International Journal of Dairy Technology*, *65*(4), 594–602. https://doi.org/10.1111/j.1471-0307.2012.00853.x

Rodushkin, I., Bergman, T., Douglas, G., Engström, E., Sörlin, D., & Baxter, D. C. (2007). Authentication of Kalix (N.E. Sweden) vendace caviar using inductively coupled plasma-based analytical techniques: Evaluation of different approaches. *Analytica Chimica Acta*, *583*(2), 310–318. https://doi.org/10.1016/j.aca.2006.10.038

Rokach, L., & Maimon, O. Z. (2008). *Data mining with decision trees. Theory and applications* (2nd ed.). World Scientific Publishing Co. Pte. Ltd.

Saraçli, S., Doğan, N., & Doğan, İ. (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications*, 203. https://doi.org/10.1186/1029-242X-2013-203

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

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

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

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

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

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

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

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

Figure Captions

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

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

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

Fig. 4. Decision classification tree resulting from the application of the CHAID algorithm for the classification of bulk anchovies using the element profile. Concentrations are reported in µg kg^{-1}.

Fig. 5. Decision classification tree resulting from the application of the QUEST algorithm for the classification of packaged anchovies using the element profile. Concentrations are reported in µg kg^{-1}.

Click here to view linked References

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

Maria Olga VARRÀ[a], *Lenka HUSÁKOVÁ*[b*], *Jan PATOČKA*[b], *Sergio GHIDINI*[a], *Emanuela ZANARDI*[a*]

[a]***Department of Food and Drug, University of Parma, Parma, Via del Taglio, 10, Parma 43126, Italy***

E-mail addresses: emanuela.zanardi@unipr.it (E. Zanardi); sergio.ghidini@unipr.it (S. Ghidini); mariaolga.varra@studenti.unipr.it (M. O. Varrà).

[b]***Department of Analytical Chemistry, Faculty of Chemical Technology, University of Pardubice, Studentská 573 HB/D, Pardubice, CZ-532 10, Czech Republic***

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

***CORRESPONDING AUTHORS:**

Lenka Husáková (L. Husáková)

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

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

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

Emanuela Zanardi (E. Zanardi)

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

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

ABSTRACT

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

Abbreviations

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

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

1. Introduction

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

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

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

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

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

The traditional anchovy transformation process by brine-ripening finds a long tradition in southern Europe. The fish, typically caught by purse seines, is quickly transported to the fish canning industry where it is beheaded, partially eviscerated and punt into ripening containers (barrels), alternating layers of fish and salt. A pressure is then applied on the top layer to facilitate the progressive elimination of water. The fish is ripened until the desired degree of maturation is reached (from 3 up to 11-12 month on average) to then be moved from the barrels. From that point on, the bulk ripened anchovies can be preserved and packaged in salt to be commercialised or further processed to obtain different products and preparations, for example by filleting and packaging-in-oil.

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

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

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

Tracing back to the origin of processed or highly processed products is considerably difficult because of the manipulation and the addition of several compounds during preparation procedures. The use of salt during anchovy manufacturing may represent the most critical point since it can potentially mask the natural elemental content of fish. Nevertheless, the multiple identification of elements using techniques such as inductively coupled plasma-optical emission spectroscopy (ICP-OES) and inductively coupled plasma-mass spectrometry (ICP-MS) have been already successfully applied to identify the origin of transformed food products such processed tomato products (Lo Feudo, Naccarato, Sindona, & Tagarelli, 2010; Fragni, Trifirò, & Nucci, 2015), fruit juices (Turra et al., 2017), wines (da Costa, Ximenez, Rodrigues, Barbosa, & Barbosa, 2020), dried beef (Franke et al., 2008), hams (Epova et al., 2018), and different types of cheese (Suhaj & Kore, 2008; Moreno-Rojas, Cámara-Martos, Sánchez-Segarra, & Amaro-López, 2012; Magdas et al., 2019). One application dealing with the use of multi-elemental analysis to authenticate seafood products is also available and it concerns the identification of caviar from different origins, which, as anchovy, is a salted product (Rodushkin et al., 2007).

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

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

2. Materials and Methods

2.1. Chemicals, standards, and reference materials

All the aqueous solutions employed for analyses were prepared using ultrapure water (0.05 µS cm^{-1}) obtained by the Milli-Q® water purification system (Millipore, Bedford, USA).

For microwave digestion, hydrogen peroxide (H_2O_2, ≥ 30% w/w) for ultra-trace analysis (Fluka Chemie AG, Buchs, Switzerland) and sub-boiled nitric acid prepared from nitric acid (65%, w/w, Selectipur quality, Lach-Ner, Neratovice, Czech Republic) by means of the sub-boiling distillation apparatus Distillacid™ BSB-939-IR (Berghof, Eningen, Germany) were used.

The working calibration solutions for ICP-MS analysis were prepared daily using the multi-element stock solutions “A”, “B1”, “B2”, and “C”. Stock solution “A” (10 mg L^{-1} of Li, B, Al, V, Cr, Fe, Mn, Ni, Cu, Zn, 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 Supelco ICP multi-element standard solution IV (Merck, Darmstadt, Germany) (containing 1 g L^{-1} of Li, B, Al, Cr, Fe, Mn, Ni, Cu, Zn, Co, Sr, Cd, Ba, Tl, Pb, and Bi) and single element standards (V, As, Se, Rb, Zr, Mo, Ru, Pd, Sn, Sb, Cs, Hf, Re, Pt, and Th) of concentration 1 ± 0.002 g L^{-1} (Analytika Ltd., Prague, Czech Republic or SCP Science, Montreal, Canada).

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

A 1 g L^{-1} stock solution of Rh (SCP Science, Montreal, Canada) was used to prepare the internal standard solution (ISTD) at concentration of 200 µg L^{-1}.

A 10 g L^{-1} stock solution prepared from urea (TraceSelect quality, Fluka Chemie AG, Buchs, Switzerland) was used to prepare carbon reference solutions.

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

2.2. Anchovy sampling and processing

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

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

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

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

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

2.2.1 Lyophilisation process

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

2.2.2 Microwave-assisted acid digestion

For subsequent ICP-MS analysis, 0.1 g of freeze-dried samples or CRMs were weighted (in triplicate) into a 10 mL perfluoroalkoxy (PFA) tube and 4 mL of 16% HNO_3 (65%, w/w HNO_3, 1:3 diluted) and 1 mL of 30% H_2O_2 were added. Three PFA tubes were placed into DAC-100S polytetrafluoroethylene vessels (Berghof, Eningen, Germany) previously filled with 25 mL of HNO_3 (16%, v/v), by ensuring that the level of liquid in the outer polytetrafluoroethylene vessel was higher than those in the PFA tubes. This way, the vapor pressures were compensated and the evaporation of the solution from the PFA tubes was avoided (Husáková et al., 2015).

Samples were decomposed using a Berghof Speedwave™ MWS-3^+ microwave digestion system (Berghof, Germany) with the maximum total output of the microwave generator (1450 W) via the following multistep 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 100 °C (ramp 5 min).

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

2.3. Mercury analysis

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

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

2.4. ICP-MS analysis

Element quantification in samples was performed by using an Agilent 7900 quadrupole ICP-MS apparatus (Agilent Technologies, Inc., Santa Clara, CA, USA) equipped with a quartz concentric nebulizer MicroMist (400 µL min^{-1}), the Peltier-cooled (2 °C) Scott quartz spray chamber, quartz torch with 2.5 mm internal diameter injector, standard sampling and skimmer nickel cones with orifices of 1 and 0.45 mm, and an octopole collision/reaction cell for interference removal using kinetic energy discrimination. A low-pulsation, 10-roller peristaltic pump with three separate channels was employed to precisely deliver both samples and ISTD. For the simultaneous ISTD aspiration and mixing with the sample the connecting tubing, connectors, and the “Y” piece (included into internal standard kit) were employed. Analytical conditions were enhanced before starting sample measurement by using the multi-elemental tuning solution (Agilent Technologies, Inc., Santa Clara, CA, USA) containing 1 µg L^{-1} of Ce, Co, Li, Mg, Tl and Y, to obtain the highest possible sensitivity for elements of low, middle, and high m/z. Using the typical operating conditions summarised in Table S2 (Supplementary Materials), a sensitivity of 6000 counts s^{-1} per µg L^{-1} and a resolution of 0.64 amu peak width (full width at half maximum intensity) were achieved for $^{7}Li^{+}$. The same parameters were 50000 counts s^{-1} per µg L^{-1} and 0.62 for $^{89}Y^{+}$, and 30000 counts s^{-1} per µg L^{-1} and 0.60 for $^{205}Tl^{+}$.

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

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

The calibration curves for the quantification of 51 elements ($R^2 > 0.999$) resulted from the acquisition of working calibration solutions prepared from multi-element stock solutions "A", "B1", "B2", and "C" described in Section 2.1. The concentration of elements for calibration were as follows: blank, 1, 5, 10, 50, 100 µg L^{-1} of 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, Hf, Ta, Re, Pt, Tl, Pb, Bi, Th; 0.1, 0.5, 1, 5, 10 µg L^{-1} of La, Ce, Pr, Nd, U; 0.02, 0.1, 0.2, 1, 2 µg L^{-1} of Y, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu; 0.5, 1, 5, 10 mg L^{-1} of Na, Mg, P, K, Ca, Mn, Cu, Zn.

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

2.5. Analytical validation

The accuracy of the methodology was checked by analysing the five certified reference materials listed in *Section 2.1* (NIST 1566 Oyster Tissue, BCR CRM 184 Bovine muscle, BCR CRM 185 Bovine Liver, CRM 12-2-01 Bovine Liver, NIST SRM 1577 Bovine Liver) intended for the evaluation of analytical methods and instruments and used for the determination of the mass fraction values of selected elements in marine tissue, foods, or similar materials. In addition, CRM 12-2-03 Essential and toxic elements in Lucerne, NCS ZC 73015 Milk Powder, and CRM 12-2-04 Wheat bread flour, were analysed to assess accuracy in determining lanthanides and actinides. The high level of agreement between target and found values demonstrated trueness of data obtained (Supplementary Materials, Table S3).

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

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

2.6. Data elaboration

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

Univariate data analysis, consisting of nonparametric Mann-Whitney test ($p \leq 0.05$) and Kruskal–Wallis test 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 × 3 replicates × 3 provenances) and the set of packaged anchovies (30 samples × 3 replicates × 3 provenances) to investigate any significant difference between groups of samples. Nonparametric tests were chosen instead of classical parametric ones because most of the data presented non normal distribution and heteroscedasticity, as verified though the application of Shapiro-Wilk's and Levene's tests, respectively ($p \leq 0.05$).

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

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

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

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

The output of the method is a set of concatenated classification rules in the form of a decision tree composed by nodes (identifying the features that need to be sorted) and branches (identifying the values assumed by nodes) (Han, Kamber, & Pei, 2011). CART, C5.0, CHAID, and QUEST represent different methods by which the architecture of the decision trees can be built up and differ each other mainly for the segmentation rules applied and the tree optimisation method. The C5.0 algorithm is based on binary splitting and works by choosing progressively the instances that allow to gain the maximum partitioning information and stopping via the pruning rule, i.e., by removing from the splits which do not add significant information. CART algorithm is also based on binary splitting but differs from C5.0 essentially for a different stopping rule in the creation of the tree, consisting of the evaluation of the purity of the node, i.e. the maximum degree of homogeneity between categories. The CHAID algorithm is a multiway splitting system based on Chi-square statistics to decide for tree ramifications and is based on the measures of the impurity of the nodes. Finally, the QUEST is a binary-split algorithm which, in the case of continuous variables, uses an ANOVA F-Test to create tree nodes. Further information can be retrieved from Han, Kamber, & Pei (2011) and from Rokach & Maimon (2008).

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

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

3. Results and discussion

3.1. Initial data evaluation

3.1.1 Global elemental profiles of transformed anchovy products

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

Complete data matrices reporting the whole concentrations of the 52 elements measured in bulk and packaged anchovies can be found in Table S5 and Table S6 (Supplementary Materials), respectively. The most abundant element was found to be Na (whose median concentrations ranged from 142 mg kg^{-1} in bulk Croatian anchovies to 177 mg kg^{-1} in Tunisian packaged anchovies) followed by P, K, Ca, and Mg. Bulk and packaged products from Cantabrian Sea differed from samples of Mediterranean origin (Croatian and Tunisian) because of significantly higher concentrations of P and K. At the same time, no differences in Na and Ca contents between Cantabrian and Croatian anchovies was encountered, which, instead, were both significantly lower compared to those found in samples originating from Tunisia (Table S5, Table S6, Supplementary Materials). Similarly, some minor, trace- and ultra-trace elements showed significant variations in relation to the provenance Besides, Cantabrian anchovies were characterised by higher Ni, Mo, Cs, and Tl concentrations and lower Al concentrations than those encountered in the two Mediterranean products (Supplementary Tables S5, S6).

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

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

3.1.2 Pre-selection of elements as potential indicators of origin

The use of salt in the form of saturated brine during the processing and packaging of ripened anchovies is one of the main complicating factors which might limit the proper identification of markers of geographical origin since it inevitably modifies the natural element profile of the product. For instance, the use of a brine for fermentation purposes was reported to decrease the total amount of certain trace elements in transformed fish roe because of partitioning phenomena between the solid and the liquid phases occurring during fermentation (Bekhit, Morton, & Dawson, 2008).

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

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

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

3.1.3. Hierarchical cluster analysis (HCA)

The inner potential of the elemental profile in guiding the creation of groups of anchovies was at first instance explored. Any possible natural difference or similarity among anchovies of both datasets was therefore uncovered by the application of HCA. Since different methods to measure distances (i.e. Euclidean and Manhattan distances) and to perform grouping (i.e. Ward's minimum variance, single linkage, complete linkage, simple average group average, median, and centroid clustering methods) are applicable for HCA building, the cophenetic correlation coefficient (CCC) was employed as an index to compare the methods with each other and to evaluate the validity of the resulting dendrograms (Sokhal & Rohlf, 1962; Saraçli, Doğan, & Doğan, 2013). Further details about the tested methods can be retrieved from Everitt, Landau, Leese, & Stahl (2011).

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

3.2 Data mining for the geographic origin evaluation

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

Four decision trees algorithms (C5.0, CART, CHAID, and QUEST) trained on 70 % of the whole sample sets (270 samples per each anchovy set) were examined, in search of the most accurate models in identifying the origin labels of bulk and packaged anchovy samples in the validation set. One of the main advantages of using decision tree is that the selection of the most important variables for classification is performed automatically during training stage. Therefore, computation time is reduced, while interpretability, accessibility and handiness of the models is improved.

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

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

Arsenic contamination of seawaters can be related to anthropogenic pollutant activities as well as to the natural geological characteristics of the area (Garelick, Jones, Dybowska, & Valsami-Jones, 2008). As an example, As (together with Cr, Cu, Hg, Mn, Ni, Pb, Se, and V) concentrations were reported to be higher in seawater where volcanic activities exist such as the Mediterranean Basin (Juncos et al., 2016; Zkeri, Aloupi, & Gaganis, 2018). Moreover, the uptake of As by fish is influenced by several natural factors including water temperature and salinity, cooccurrence of phosphate, and seasonal differences of the distribution of the inorganic and organic forms of As in the aquatic environments (Ferrante et al., 2019). Regardless the natural or anthropogenic nature of As and releasing sources, the reduced exchange of water in the Mediterranean Basin and, especially in the Adriatic Sea, can facilitate the accumulation of As in the environment (Ferrante et al., 2019). This can justify the higher amounts of As in anchovy from the Mediterranean Sea compared to Cantabrian (Atlantic Ocean) anchovy reported in the present work (see Fig. 1A, Fig. 1B). Moreover, in pelagic fish species from the Adriatic Sea higher amounts of As compared to other sampling zones was previously shown (Storelli & Marcotrigiano, 2004).

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

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

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

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

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

The QUEST-based decision trees applied to packaged anchovy was composed by 13 decisions nodes stratified into four levels. B was selected as the first binary splitting variable. The outcomes related to predictor importance reported in Fig. 3B (according to which B had the highest influence in prediction) were confirmed by analysing the splitting variables used to generate the QUEST decision tree, where B was just selected as the first binary splitting variable (Fig. 5). B value higher than 5.13 mg kg^{-1} generated a leaf (final) node with 91.3% probability of predicting samples originating from Tunisia. Globally, 100% probability of correctly identifying Croatian samples was reached at the last tree-level using the classification rule based on B, As, and Cd or the classification rule based on B, As, K, and Pd. With increased B (> 5.13 mg kg^{-1}), As (> 7.14 mg kg^{-1}), and K (> 5669 mg kg^{-1}) concentrations, also the likelihood of recognising Cantabrian anchovies increased. The set of decision rules established by QUEST further proved to be reliable and effective for the classification of packaged anchovy origin, owing to the good ability in predicting the unknown origin of samples of the test set. Compared to other decision tree algorithm, QUEST ranked first in terms of accuracy (97.7 %), sensitivity (97.6 %), specificity (98.9 %), and precision (98.0 %) (Table 1).

Even though transformed anchovy implicitly represents a complex processed foodstuff, it is important to stress that using decision trees may be the quickest and the most intelligible way to solve problems related to classification of foodstuffs.

4. Conclusions

In this work, data mining techniques were applied to transformed anchovy products to verify whether the origin of fish could be identified through the elemental patterns measured by ICP-MS and direct mercury analysis. Different machine learning algorithms relying on the principle of decision trees were applied to data and classification rules to distinguish anchovy fish of Cantabrian Sea from Tunisian and Croatian anchovies were created. Firstly, differences of elemental composition between anchovies at two stages of the production chain were investigated, to verify whether misleading elemental inclusion from the manufacturing environment was introduced. After having excluded problematic elements based on literature review and direct comparison of bulk and packaged anchovy profile and after having explored the effective presence of fish clusters related to origin, C5.0, CART, CHAID and QUEST decision trees were trained. This way, the selection of the most important variables and the identification of cut-off limits for each element concentration to describe a specific group of samples were performed in tandem.

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

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

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

Appendix A. Supplementary Materials

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

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

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

Acknowledgements

The research was carried out within the project "Development of rapid tools for anchovies authentication" funded by Rizzoli Emanuelli S.p.a. (Parma, Italy). The authors also gratefully acknowledge the University of Pardubice for the financial support (project no. SGS_2020_002).

References

Barbosa, R. M., Nacano, L. R., Freitas, R., Batista, B. L., & Barbosa Jr, F. (2014). The use of decision trees and naive Bayes algorithms and trace element patterns for controlling the authenticity of free- range-pastured hens’ eggs. Journal of food science, 79(9), C1672-C1677. https://doi.org/10.1111/1750-3841.12577

Bekhit, A. E. D. A., Morton, J. D., & Dawson, C. O. (2008). Effect of processing conditions on trace elements in fish roe from six commercial New Zealand fish species. *Journal of Agricultural and Food Chemistry*, *56*(12), 4846–4853. https://doi.org/10.1021/jf8005646

Besteiro, I., Rodríguez, C. J., Tilve-Jar, C., & Pascual López, C. (2000). Selection of attributes for the sensory evaluation of anchovies during the ripening process. *Journal of Sensory Studies*, *15*(1), 65–77. https://doi.org/10.1111/j.1745-459X.2000.tb00410.x

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

Cuadros-Rodríguez, L., Pérez-Castaño, E., & Ruiz-Samblás, C. (2016). Quality performance metrics in multivariate classification methods for qualitative analysis. *Trends in Analytical Chemistry*, *80*, 612–624. https://doi.org/10.1016/j.trac.2016.04.021

Czerner, M., Agustinelli, S. P., Guccione, S., & Yeannes, M. I. (2015). Effect of different preservation processes on chemical composition and fatty acid profile of anchovy (Engraulis anchoita). *International Journal of Food Sciences and Nutrition, 66(8),* 887–894. https://doi.org/10.3109/09637486.2015.1110687

Da Costa, N. L., Ximenez, J. P. B., Rodrigues, J. L., Barbosa, F., & Barbosa, R. (2020). Characterization of Cabernet Sauvignon wines from California: determination of origin based on ICP-MS analysis and machine learning techniques. *European Food Research and Technology, 246(6),* 1193–1205. https://doi.org/10.1007/s00217-020-03480-5

Depeters, E. J., Puschner, B., Taylor, S. J., & Rodzen, J. A. (2013). Can fatty acid and mineral compositions of sturgeon eggs distinguish between farm-raised versus wild white (Acipenser transmontanus) sturgeon origins in California ? Preliminary report. *Forensic Science International*, *229*(1–3), 128–132. https://doi.org/10.1016/j.forsciint.2013.04.003

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

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

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

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

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

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

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

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

Franke, B. M., Haldimann, M., Gremaud, G., Bosset, J. O., Hadorn, R., & Kreuzer, M. (2008). Element signature analysis: Its validation as a tool for geographic authentication of the origin of dried beef and poultry meat. *European Food Research and Technology, 227(3)*, 701–708. https://doi.org/10.1007/s00217-007-0776-8

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

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

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

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

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

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

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

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

https://doi.org/10.1007/s10462-011-9272-4

Laso, J., Margallo, M., Fullana, P., Bala, A., Gazulla, C., Irabien, A., & Aldaco, R. (2017). Introducing life cycle thinking to define best available techniques for products: Application to the anchovy canning industry. *Journal of Cleaner Production*, *155*, 139–150. https://doi.org/10.1016/j.jclepro.2016.08.040

Li, L., Han, C., Dong, S., & Boyd, C. E. (2019). Use of elemental profiling and isotopic signatures to differentiate Pacific white shrimp (Litopenaeus vannamei) from freshwater and seawater culture areas. *Food Control*, *95*, 249–256. https://doi.org/10.1016/j.foodcont.2018.08.015

Lo Feudo, G., Naccarato, A., Sindona, G., & Tagarelli, A. (2010). Investigating the origin of tomatoes and triple concentrated tomato pastes through multielement determination by inductively coupled plasma mass spectrometry and statistical analysis. *Journal of Agricultural and Food Chemistry*, *58*(6), 3801–3807. https://doi.org/10.1021/jf903868j

Magdas, D. A., Feher, I., Cristea, G., Voica, C., Tabaran, A., Mihaiu, M., Cordea, D. V., Bâlteanu, V. A., & Dan, D. S. (2019). Geographical origin and species differentiation of Transylvanian cheese. Comparative study of isotopic and elemental profiling vs . DNA results. *Food Chemistry*, *277*(June 2018), 307–313. https://doi.org/10.1016/j.foodchem.2018.10.103

Moreno-Rojas, R., Cámara-Martos, F., Sánchez-Segarra, P. J., & Amaro-López, M. Á. (2012). Influence of manufacturing conditions and discrimination of Northern Spanish cheeses using multi-element analysis. *International Journal of Dairy Technology*, *65*(4), 594–602. https://doi.org/10.1111/j.1471-0307.2012.00853.x

Rodushkin, I., Bergman, T., Douglas, G., Engström, E., Sörlin, D., & Baxter, D. C. (2007). Authentication of Kalix (N.E. Sweden) vendace caviar using inductively coupled plasma-based analytical techniques: Evaluation of different approaches. *Analytica Chimica Acta*, *583*(2), 310–318. https://doi.org/10.1016/j.aca.2006.10.038

Rokach, L., & Maimon, O. Z. (2008). *Data mining with decision trees. Theory and applications* (2nd ed.). World Scientific Publishing Co. Pte. Ltd.

Saraçli, S., Doğan, N., & Doğan, İ. (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications*, 203. https://doi.org/10.1186/1029-242X-2013-203

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

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

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

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

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

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

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

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

Figure Captions

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

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

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

Fig. 4. Decision classification tree resulting from the application of the CHAID algorithm for the classification of bulk anchovies using the element profile. Concentrations are reported in µg kg^{-1}.

Fig. 5. Decision classification tree resulting from the application of the QUEST algorithm for the classification of packaged anchovies using the element profile. Concentrations are reported in µg kg^{-1}.

Fig. 1

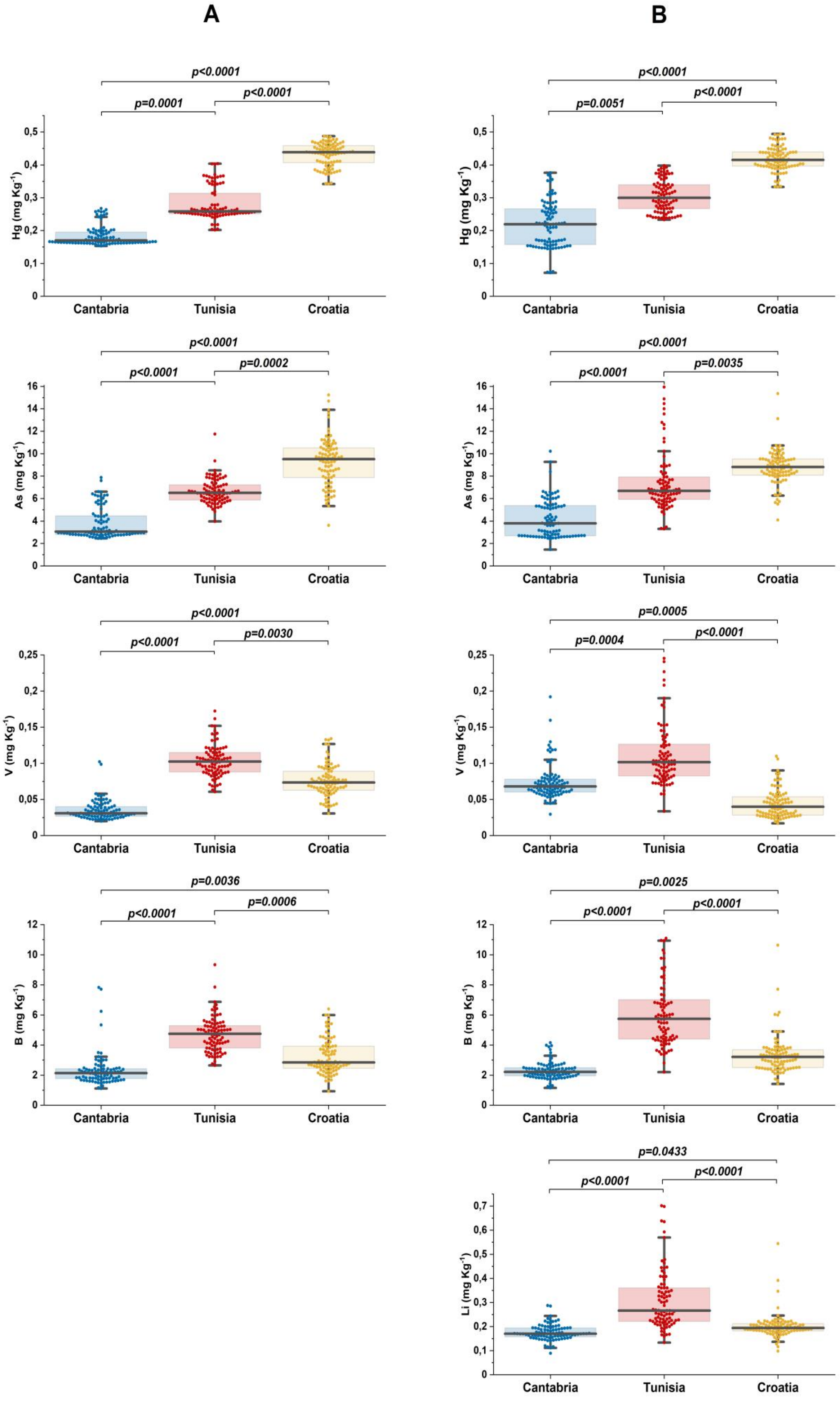

A
B
p<0.0001
p=0.0001
p<0.0001
Hg (mg Kg-1)
0,5
0,4
0,3
0,2
0,1
0
Cantabria
Tunisia
Croatia
p<0.0001
p=0.0051
p<0.0001
Hg (mg Kg-1)
p<0.0001
p<0.0001
p=0.0002
As (mg Kg-1)
16
14
12
10
8
6
4
2
0
p<0.0001
p<0.0001
p=0.0035
As (mg Kg-1)
p<0.0001
p<0.0001
p=0.0030
V (mg Kg-1)
0,25
0,2
0,15
0,1
0,05
0
p=0.0005
p=0.0004
p<0.0001
V (mg Kg-1)
p=0.0036
p<0.0001
p=0.0006
B (mg Kg-1)
12
10
8
6
4
2
0
p=0.0025
p<0.0001
p<0.0001
B (mg Kg-1)
p=0.0433
p<0.0001
p<0.0001
Li (mg Kg-1)
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0

Fig. 2.

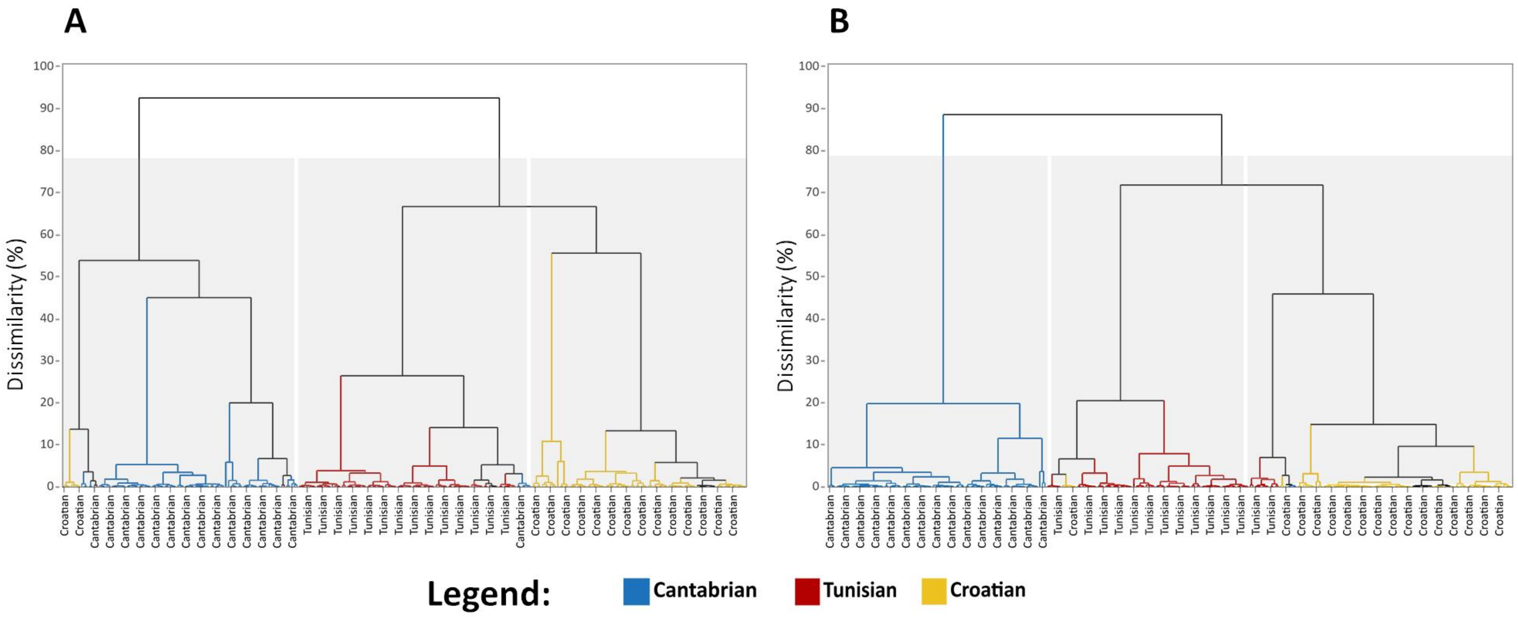

A
100
90
80
70
60
50
40
30
20
10
0
Dissimilarity (%)
Croatian
Croatian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Tunisian
Tunisian
Tunisian
Tunisian
Tunisian
Tunisian
Tunisian
Tunisian
Tunisian
Tunisian
Tunisian
Tunisian
Tunisian
Tunisian
Cantabrian
Croatian
Croatian
Croatian
Croatian
Croatian
Croatian
Croatian
Croatian
Croatian
Croatian
Croatian
Croatian
Croatian
Croatian
B
100
90
80
70
60
50
40
30
20
10
0
Dissimilarity (%)
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Cantabrian
Tunisian
Croatian
Tunisian
Tunisian
Tunisian
Tunisian
Tunisian
Tunisian
Tunisian
Tunisian
Tunisian
Tunisian
Tunisian
Tunisian
Tunisian
Croatian
Croatian
Croatian
Croatian
Croatian
Croatian
Croatian
Croatian
Croatian
Croatian
Croatian
Croatian
Croatian
Croatian
Croatian
Legend:
Cantabrian
Tunisian
Croatian

Fig.3.

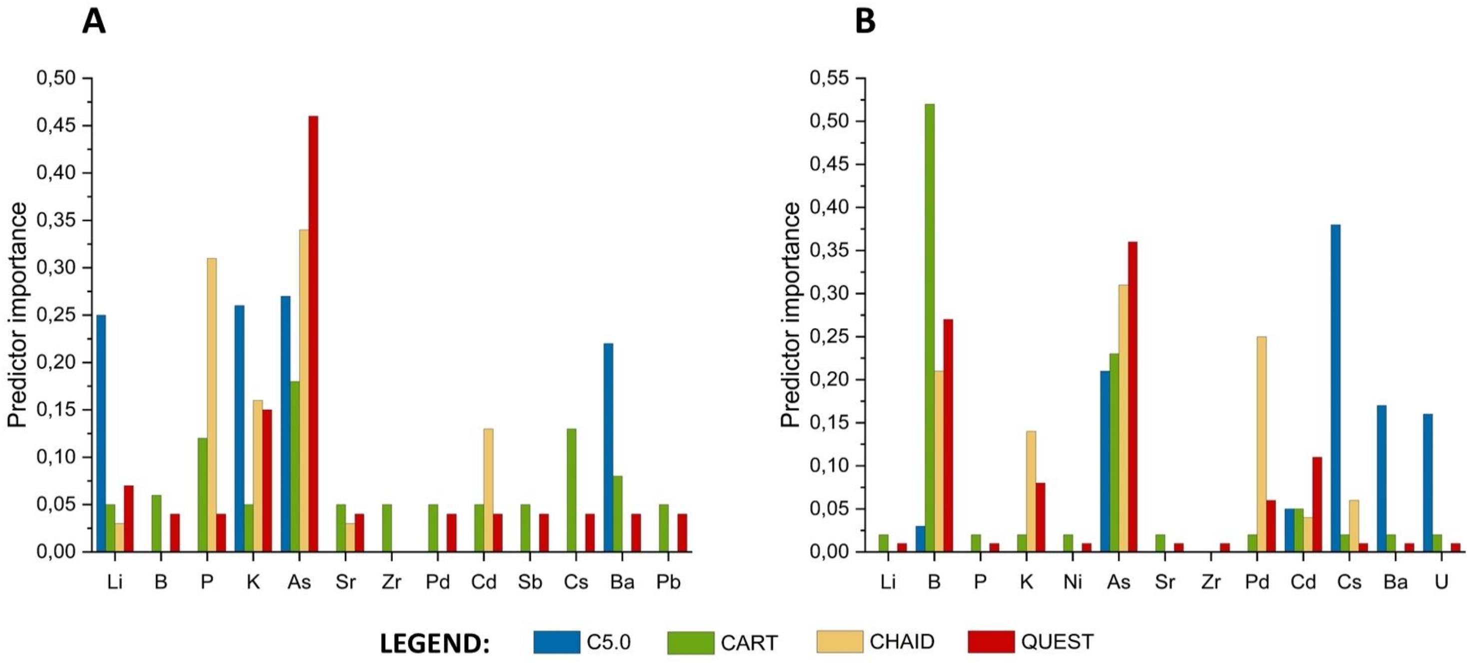

A
B
Predictor importance
0,00
0,05
0,10
0,15
0,20
0,25
0,30
0,35
0,40
0,45
0,50
0,55
Li
B
P
K
Ni
As
Sr
Zr
Pd
Cd
Sb
Cs
Ba
Pb
U
LEGEND:
C5.0
CART
CHAID
QUEST

Fig. 4.

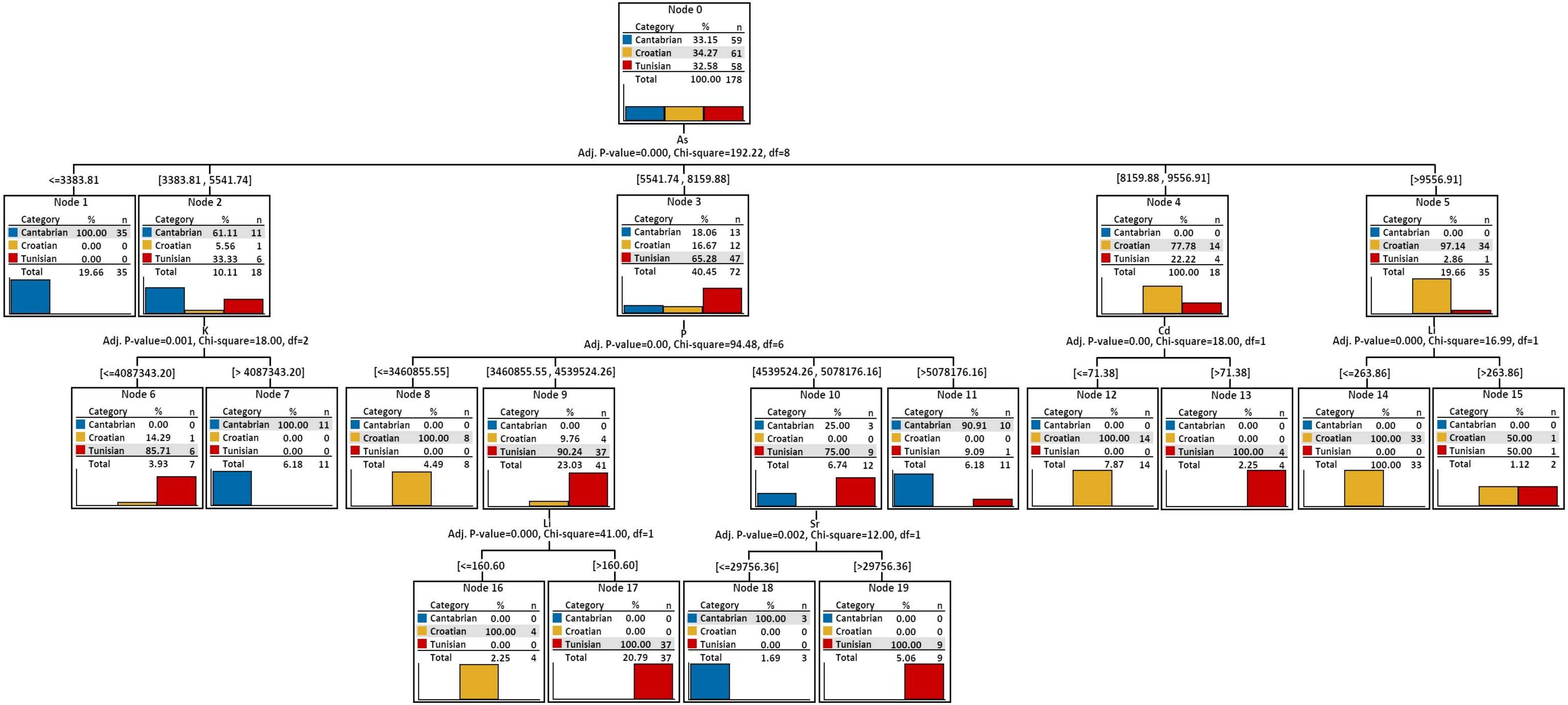
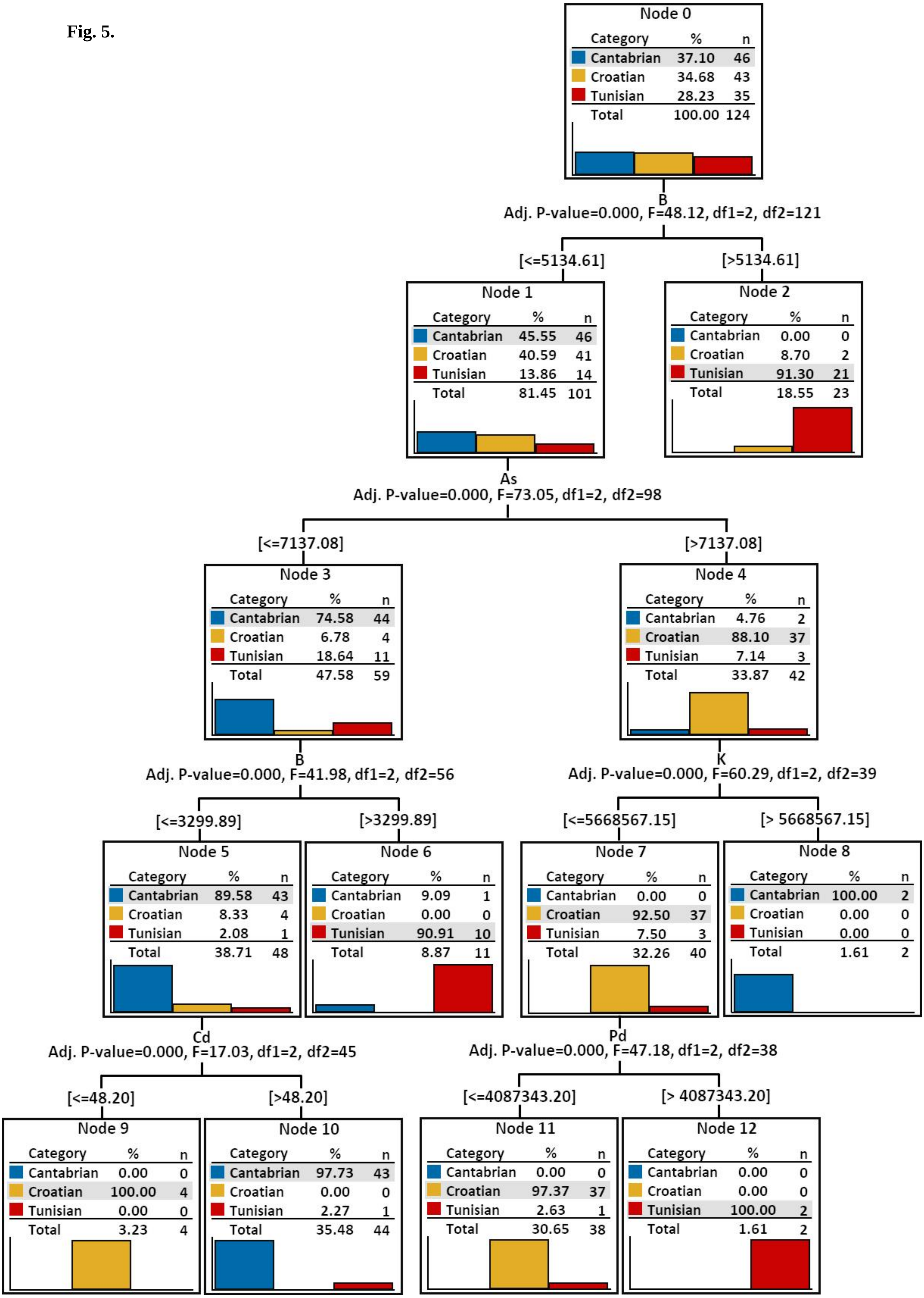

Node 0
Category % n
Cantabrian 33.15 59
Croatian 34.27 61
Tunisian 32.58 58
Total 100.00 178
As
Adj. P-value=0.000, Chi-square=192.22, df=8
<=3383.81
[3383.81 , 5541.74]
[5541.74 , 8159.88]
[8159.88 , 9556.91]
[>9556.91]
Node 1
Category % n
Cantabrian 100.00 35
Croatian 0.00 0
Tunisian 0.00 0
Total 19.66 35
Node 2
Category % n
Cantabrian 61.11 11
Croatian 5.56 1
Tunisian 33.33 6
Total 10.11 18
Node 3
Category % n
Cantabrian 18.06 13
Croatian 16.67 12
Tunisian 65.28 47
Total 40.45 72
Node 4
Category % n
Cantabrian 0.00 0
Croatian 77.78 14
Tunisian 22.22 4
Total 100.00 18
Node 5
Category % n
Cantabrian 0.00 0
Croatian 97.14 34
Tunisian 2.86 1
Total 19.66 35
K
Adj. P-value=0.001, Chi-square=18.00, df=2
P
Adj. P-value=0.00, Chi-square=94.48, df=6
Cd
Adj. P-value=0.00, Chi-square=18.00, df=1
Li
Adj. P-value=0.000, Chi-square=16.99, df=1
[<=4087343.20]
[> 4087343.20]
[<=3460855.55]
[3460855.55 , 4539524.26]
[4539524.26 , 5078176.16]
[>5078176.16]
[<=71.38]
[>71.38]
[<=263.86]
[>263.86]
Node 6
Category % n
Cantabrian 0.00 0
Croatian 14.29 1
Tunisian 85.71 6
Total 3.93 7
Node 7
Category % n
Cantabrian 100.00 11
Croatian 0.00 0
Tunisian 0.00 0
Total 6.18 11
Node 8
Category % n
Cantabrian 0.00 0
Croatian 100.00 8
Tunisian 0.00 0
Total 4.49 8
Node 9
Category % n
Cantabrian 0.00 0
Croatian 9.76 4
Tunisian 90.24 37
Total 23.03 41
Node 10
Category % n
Cantabrian 25.00 3
Croatian 0.00 0
Tunisian 75.00 9
Total 6.74 12
Node 11
Category % n
Cantabrian 90.91 10
Croatian 0.00 0
Tunisian 9.09 1
Total 6.18 11
Node 12
Category % n
Cantabrian 0.00 0
Croatian 100.00 14
Tunisian 0.00 0
Total 7.87 14
Node 13
Category % n
Cantabrian 0.00 0
Croatian 0.00 0
Tunisian 100.00 4
Total 2.25 4
Node 14
Category % n
Cantabrian 0.00 0
Croatian 100.00 33
Tunisian 0.00 0
Total 100.00 33
Node 15
Category % n
Cantabrian 0.00 0
Croatian 50.00 1
Tunisian 50.00 1
Total 1.12 2
Li
Adj. P-value=0.000, Chi-square=41.00, df=1
Sr
Adj. P-value=0.002, Chi-square=12.00, df=1
[<=160.60]
[>160.60]
[<=29756.36]
[>29756.36]
Node 16
Category % n
Cantabrian 0.00 0
Croatian 100.00 4
Tunisian 0.00 0
Total 2.25 4
Node 17
Category % n
Cantabrian 0.00 0
Croatian 0.00 0
Tunisian 100.00 37
Total 20.79 37
Node 18
Category % n
Cantabrian 100.00 3
Croatian 0.00 0
Tunisian 0.00 0
Total 1.69 3
Node 19
Category % n
Cantabrian 0.00 0
Croatian 0.00 0
Tunisian 100.00 9
Total 5.06 9

Fig. 5.



Node 0

Category	%	n
🟦 Cantabrian	37.10	46
🟨 Croatian	34.68	43
🟥 Tunisian	28.23	35
Total	100.00	124

Node 1

Category	%	n
🟦 Cantabrian	45.55	46
🟨 Croatian	40.59	41
🟥 Tunisian	13.86	14
Total	81.45	101

Node 2

Category	%	n
🟦 Cantabrian	0.00	0
🟨 Croatian	8.70	2
🟥 Tunisian	91.30	21
Total	18.55	23

Node 3

Category	%	n
🟦 Cantabrian	74.58	44
🟨 Croatian	6.78	4
🟥 Tunisian	18.64	11
Total	47.58	59

Node 4

Category	%	n
🟦 Cantabrian	4.76	2
🟨 Croatian	88.10	37
🟥 Tunisian	7.14	3
Total	33.87	42

Node 5

Category	%	n
🟦 Cantabrian	89.58	43
🟨 Croatian	8.33	4
🟥 Tunisian	2.08	1
Total	38.71	48

Node 6

Category	%	n
🟦 Cantabrian	9.09	1
🟨 Croatian	0.00	0
🟥 Tunisian	90.91	10
Total	8.87	11

Node 7

Category	%	n
🟦 Cantabrian	0.00	0
🟨 Croatian	92.50	37
🟥 Tunisian	7.50	3
Total	32.26	40

Node 8

Category	%	n
🟦 Cantabrian	100.00	2
🟨 Croatian	0.00	0
🟥 Tunisian	0.00	0
Total	1.61	2

Node 9

Category	%	n
🟦 Cantabrian	0.00	0
🟨 Croatian	100.00	4
🟥 Tunisian	0.00	0
Total	3.23	4

Node 10

Category	%	n
🟦 Cantabrian	97.73	43
🟨 Croatian	0.00	0
🟥 Tunisian	2.27	1
Total	35.48	44

Node 11

Category	%	n
🟦 Cantabrian	0.00	0
🟨 Croatian	97.37	37
🟥 Tunisian	2.63	1
Total	30.65	38

Node 12

Category	%	n
🟦 Cantabrian	0.00	0
🟨 Croatian	0.00	0
🟥 Tunisian	100.00	2
Total	1.61	2

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.