

Article

Probabilistic Machine Learning for the Authentication of the Protected Designation of Origin of Greek Bottarga from Messolongi: A Generic Methodology to Cope with Very Small Number of Samples

George Tsirogiannis ^{1,*}, Anna-Akrivi Thomatou ¹, Eleni Psarra ¹, Eleni C. Mazarakioti ¹, Katerina Katerinopoulou ¹ , Anastasios Zotos ², Achilleas Kontogeorgos ^{3,*} , Angelos Patakas ¹  and Athanasios Ladavos ¹ 

¹ Department of Business Administration of Food and Agricultural Enterprises, University of Patras, 30100 Agrinio, Greece; athomatu@upatras.gr (A.-A.T.); e.psarra@upatras.gr (E.P.); e.mazarakioti@upnet.gr (E.C.M.); akaterin@upatras.gr (K.K.); apatakas@upatras.gr (A.P.); alantavo@upatras.gr (A.L.)

² Department of Biosystems Science and Agricultural Engineering, University of Patras, 30200 Messolongi, Greece; azotos@upatras.gr

³ Department of Agriculture, International Hellenic University, 57001 Thessaloniki, Greece

* Correspondence: gtsirogianni@upatras.gr (G.T.); akontoge@upatras.gr (A.K.); Tel.: +30-26410-74121 (G.T.)



Citation: Tsirogiannis, G.; Thomatou, A.-A.; Psarra, E.; Mazarakioti, E.C.; Katerinopoulou, K.; Zotos, A.; Kontogeorgos, A.; Patakas, A.; Ladavos, A. Probabilistic Machine Learning for the Authentication of the Protected Designation of Origin of Greek Bottarga from Messolongi: A Generic Methodology to Cope with Very Small Number of Samples. *Appl. Sci.* **2022**, *12*, 6335. <https://doi.org/10.3390/app12136335>

Academic Editors: Alessandra Biancolillo, Mengxing Li, Manping Jia and Peihua Ma

Received: 18 May 2022

Accepted: 21 June 2022

Published: 22 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Consumers are increasingly interested in the geographical origin of foodstuff, as an important characteristic of food authenticity and quality. To assure the authenticity of the geographical origin, various methods have been proposed. Stable isotope analysis is a method that has been extensively used for products like wine, oil, and meat by using large datasets and analysis. On the other hand, only few studies have been conducted for the discrimination of seafood origin and especially for mullet roes or bottarga products, and even fewer investigate a small number of samples and datasets. Stable isotopes of Carbon (C), Nitrogen (N), and Sulfur (S) analysis of bottarga samples from four different origins were carried out. The first results show that the stable isotopes ratios of C, N, and S could be used to discriminate the Greek PDO Bottarga (Messolongi) from other similar products by using a probabilistic machine learning methodology. That could use limited sample data to fit/estimate their parameters, while, at the same time, being capable of describing accurately the population and discriminate individual samples regarding their origin.

Keywords: stable isotopes; food authenticity; probabilistic machine learning; PDO products

1. Introduction

Food authenticity is of particular interest for both consumers and producers. Consumers are interested in local products of good quality. Therefore, the geographical origin and authenticity of foodstuff are of increasing interest [1]. However, the growing demand for food has led to fraud and adulteration, which are quite often phenomena in food markets. More specifically, adulteration of food with cheaper and lower quality components can lead to food fraud that could affect consumers and producers and face financial impacts and public health risks [2,3].

Over the last decades, a rapid evolution in fish production has been observed. Aquatic foods are a substantial part of the food commerce. They are composed of omega-3 fatty acids, vitamins, minerals, trace elements, high-quality protein, and essential amino acids; thereby, they improve nutrition, health, and well-being [4]. Fish consumption and demand have increased due to the health advantages and have led to the development of the aquaculture industry [5]. Therefore, the continuous rise lurks a lot of risks to security and

safety in seafood consumption, such as food fraud, diseases, and food contamination that can occur along the food chain [6–9].

In addition, the need for determining food authenticity led the European Union to institute a Traceability Regulation (178/2002/EC). It came into force in January 2005 and defines “food and feed traceability.” Moreover, for foods with Protected Geographical Indications (PGI), Protected Designations of Origin (PDO), and Traditional Specialties Guarantee (TSG), European laws EC N. 510/2006 and 1151/2012 require protection against mislabeling [8]. The mullet (*Mugil Cephalus*) is a fish species that can be found in littoral waters. Roes, which are the mullet’s eggs, are taken out from the fish and, after a treatment of salting and drying, produces seafood that is named alternatively according to the area that is produced, like Greek “Avgotaracho” (see Figure 1), Japanese Karasumi, or Italian Bottarga [10].

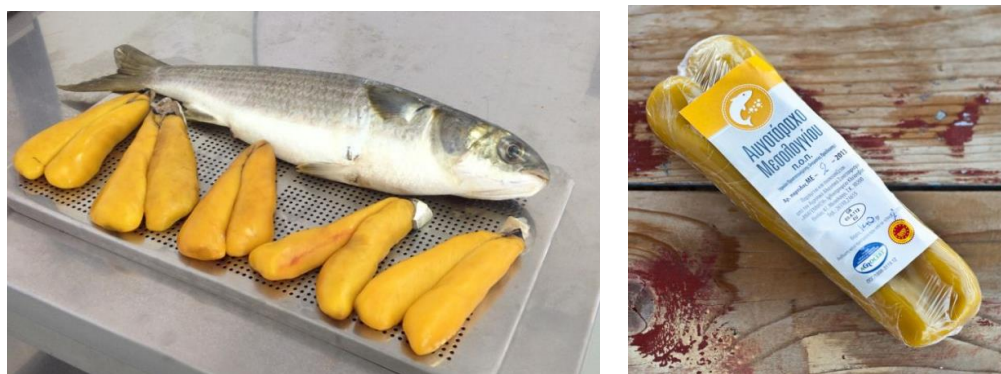


Figure 1. Greek Bottarga from *Messolongi* (Fish-Roe) Registered “Avgotaracho *Messolongiou*” PDO.

A very important tool in food safety strategy is traceability systems. Several analytical methods have been proposed to discriminate the geographical origin of food products. Modern instruments and techniques have been used for the accurate identification of the geographical origin of products; such techniques include plasma laser induced breakdown spectroscopy (LIBS); fourier-transform infrared spectroscopy (FTIR); near-infrared (NIR) and mid-infrared (MIR) spectroscopy; chromatographic techniques, such as gas chromatography–mass spectrometry (GC-MS) or high-performance liquid chromatography (HPLC); nuclear magnetic resonance (NMR); electron spin resonance (ESR); polymerase chain reaction (PCR); enzymatic assays (ELISA); DNA based techniques; and isotope-ratio mass spectrometry (IRMS) with high measurement accuracy, which were developed to serve this need [11]. Amongst them, stable isotope ratio analysis has proven to respond well to adulteration detection of agri-food products. Stable isotope analysis gives information about the geographical origin of various food products, such as juice, wine, milk, honey, oil, and meat [11–16]. Few studies have been conducted on isotopic analysis for the determination of seafood, and most of them deal with the discrepancy between farmed and non-farmed products [17–25]. Kim et al. (2015) were the first who investigated the geographical origin of commercial fish by using stable isotope ratio analysis [12]. In particular, they used stable isotope analysis for the discrimination of shrimp and hairtail fish and found that $\delta^{13}\text{C}$ and $\delta^{15}\text{N}$ ratios can be used to differentiate the geographic origin of the samples of the same kind. Moretti et al. (2003) referred to the possibility that stable isotope ratio analysis can be applied for the determination of fish seafood products’ geographical origin [26]. Gopi, et.al. (2019) showed that isotopic profiles can be used to trace the geographic origins of farmed and wild-caught Asian seabass [9]. Turchini et al. (2009) studied the discrimination between farms in different geographical areas by using $\delta^{13}\text{C}$, $\delta^{15}\text{N}$ and $\delta^{18}\text{O}$ isotopic ratios and concluded that the combination of these ratios could help in the differentiation of the seafood origin [27].

The present work aims to study the application of stable isotope analysis using EA-IRMS to discriminate the geographical origin of bottarga samples from Messolongi and

samples from different regions by using a model for the discrimination of bottarga samples in the case of a very small number of samples.

2. Modeling Approach

In most studies, stable isotopic analysis is performed with a large dataset of samples to discriminate the geographical origin of food. However, in some cases, geographical discrimination could also be achieved with only a few samples. Carter et al. (2015) succeeded to discriminate a small number of prawns' samples [28]. In particular, they used only ten samples to decide which parameters are more useful in order to discriminate Australian fresh prawns from other, mainly imported, samples. Then, the method was applied to seventeen unknown samples to distinguish the geographical origin. Dinca et al. (2015) used stable isotopic analysis to determine honey samples from different floral sources and different areas [29]. They used a small number of samples (40 honey samples) to separate them in different growing areas and a smaller number of samples to separate them according to their botanical origin (13 samples of acacia, 4 of linden, 6 of sunflower, 4 of rape, 4 of honeydew, and 9 of multifloral honey). Thus, despite the number of samples, stable isotope ratio analysis led to good discrimination of honey samples in every case and is considered an effective tool for the authentication of honey.

The modeling approach of this paper is focused on establishing a discrimination mechanism of individual bottarga samples based on a very small number of samples available for analysis. In this research, there were available five different samples (different lot numbers) from Messolongi and two different samples (different lot number) for the other three areas included in the analysis (namely Preveza—Greece, Australia & Mauritania). Each sample was used to measure stable isotopes ratios of Carbon, Nitrogen, and Sulfur. The isotopic analyses of carbon, nitrogen, and sulfur were performed by an Elementar Isoprime 100 Isotope Ratio Mass Spectrometry (IRMS) instrument (IsoPrime Ltd., Cheadle Hulme, UK) coupled to an Elemental Analyzer (Elementar Vario Isotope EL Cube, Elementar Analysensysteme GmbH, Hanau, Germany). Samples were weighed into tin capsules for measurement and were loaded on the auto-sampler of the IRMS analyzer. The results of the isotope ratio analyses were expressed as delta values δ (‰) and calculated according to the following equation:

$$\delta X (\text{‰}) = [(R_{\text{sample}}/R_{\text{standard}}) - 1] \times 1000$$

where:

- X is the isotope being studied (e.g., ^{13}C , ^{15}N , ^{34}S),
- R_{sample} is the isotopic ratio of the measured element in its physical form (e.g., $^{13}\text{C}/^{12}\text{C}$, $^{15}\text{N}/^{14}\text{N}$, $^{34}\text{S}/^{32}\text{S}$) in the sample, and
- R_{standard} is the isotopic ratio of the reference material used in the analysis.

Each sample was analyzed 5 times resulting for 25 measurements for Messolongi and 30 measurements for other areas. That is 55 measurements for each of the elements measured in the analysis.

The whole concept of this model is based on stable isotopes ratios and quantifies the differences between the bottarga samples from Messolongi and samples from other regions (local and worldwide). Due to the probabilistic nature of the proposed machine learning model (see details in Sections 2.2–2.5), an aggregated high probability confidence is very strong statistical evidence that one sample comes from the region of interest or not (Section 2.6). The technical challenges of the problem are discussed in the next section.

2.1. Technical Challenges and Requirements

Several technical challenges arise when we model populations where sampling is very costly and/or limited by external factors. In the case of bottarga, the price of the product allows only the collection of a very small number of samples. Each sample was randomly chosen and measured from the annual local productions, the whole block is

rendered useless (though, only 1–2 g is needed, the collection process is destructive). Therefore, an attempt was made to develop a model that is very data frugal during model parameter fit, and at the same time capable of reporting its uncertainty, which is essential for characterizing individual future samples.

Another technical challenge is the variability of the sample measurements. Even small variations in sample preparation and in combination with laboratory equipment reproducibility properties may result in measurement error. This source of variability hampers the performance of the predictive model. When we certify or reject a new sample, we need a predictive model that can incorporate this variability source into its decision and report its confidence levels.

Finally, an explainable model is preferable to a black box. We aim for models that can easily explain how they make a prediction, not only because we can trust them due to our deep understanding, but also because they provide us valuable knowledge regarding the problem we solve (in our case which ratio of stable isotopes is more dominant in Messolongi region and triggers a “why” question).

2.2. Model Details

Modern machine learning offers a plethora of models, algorithms, and software frameworks [30,31]. They are based on a wide range of mathematical/statistical ideas and methods, while they vary significantly in their complexity and performance. From this multitude of available models, only a tiny subset fulfills and addresses the technical challenges of Section 2.1. The very small size of the dataset (11 different samples) significantly restricts the model complexity. In addition, the small size implies a significant uncertainty of the model parameters, which should be visible during the decision process; thus, the proposed model should explicitly report it. The measurement error, which is not negligible, is considered by a hierarchical structure, as well. Such a machine learning model does not explicitly exist in a predefined form, but as a generic part of a probabilistic machine learning approach [32].

Before presenting the model structure and parameters, the features and target variable should be clarified. In machine learning, features are the data used as input to the model, while the target is the desired output. In this case, primary features are the ratios of stable isotopes while their combinations up to second-order are the derived ones. The list of primary features is summarized in Table 1. It is noted that δ will be used instead of δ notation, on the caption of plots due to software font support limitations.

Table 1. List of primary features.

Feature	Is Influenced By
$\delta^{15}\text{N}_{\text{AIR}} (\text{‰})$	Nutrient uptake, digestion, metabolism, Cultivation Conditions (i.e., fertilizers)—Techniques
$\delta^{13}\text{C}_{\text{V-PDB}} (\text{‰})$	Nutrient uptake (e.g., C3 versus C4), digestion and metabolism, adulteration of foods
$\delta^{34}\text{S}_{\text{V-CDT}} (\text{‰})$	Bacterial action, Distance to Sea (e.g., inland versus coastal geography)

The target variable is the sample origin which follows the following encoding:

$$\text{target} = \begin{cases} 1. & \text{if origin of sample is Messolongi, Greece} \\ 0. & \text{if origin of sample is not Messolongi, Greece} \end{cases}$$

By this encoding, the problem is transformed into an equivalent binary classification. The proposed probabilistic model is a generalized linear model with parameters that will be estimated by their full distribution rather than point estimates (full distribution estimation will reveal and quantify the model uncertainty). By construction, it has the desired simplicity due to the small data set, it is hierarchical and able to cope with external sources of variations (the measurement error in our case).

The standard notation of probabilistic modeling is used [33,34], where $p(\cdot)$ which denotes a marginal distribution, and $p(\cdot|\cdot)$ a conditional probability density for both continues and discrete random variables. The “ $\sim$ ” symbol is used to denote that a random variable follows known distributions. In this paper $\theta \sim N(\mu, \sigma^2)$ denotes that θ has a normal distribution with mean μ and variance σ^2 , which is equivalent to $p(\theta) \sim N(\theta|\mu, \sigma^2)$ or even more explicitly $p(\theta|\mu, \sigma^2) \sim N(\theta|\mu, \sigma^2)$. $\theta \sim B(p)$ denotes a Bernoulli distribution with success probability p , or equivalently a binomial distribution when $n = 1$, $\theta \sim \text{Bin}(p, n)$.

The basic idea of the proposed probabilistic classification model is that we consider each sample of bottarga as a Bernoulli trial, where success is outcome 1 (i.e., origin of sample is Messolongi—Greece according to our encoding) with success probability p that is defined for each trial by a non-linear function which depends on the linear combinations of our features (i.e., linear combinations of ratios of stable isotopes up to second-order). The mathematical definition of this model for n samples is the following:

$$\prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}$$

where $p_i = \frac{1}{1+e^{-z_i}}$, is the probability of success for each sample (i.e., a nonlinear function which depends on z_i). z_i is the linear part of the model, and effectively the one that contains the model parameters we need to infer from the data (see [32] (ch. 12)).

To be able to infer the distributions of the proposed probabilistic machine learning model, we need to define explicitly all its parameters. The full model is the following:

$$m_{\delta 13\text{CV-PDB}} \sim N(0, \infty) \quad (1)$$

$$m_{\delta 15\text{NAIR} (\text{‰})} \sim N(0, \infty) \quad (2)$$

$$m_{\delta 34\text{SV-CDT} (\text{‰})} \sim N(0, \infty) \quad (3)$$

$$m_{\delta 13\text{CV-PDB}^2} \sim N(0, \infty) \quad (4)$$

$$m_{\delta 15\text{NAIR} (\text{‰})^2} \sim N(0, \infty) \quad (5)$$

$$m_{\delta 34\text{SV-CDT} (\text{‰})^2} \sim N(0, \infty) \quad (6)$$

$$c \sim N(0, \infty) \quad (7)$$

$$y_i \sim B\left(\frac{1}{1 + e^{-z_i}}\right) \quad (8)$$

$$z_i = f(m_{\delta 13\text{CV-PDB}}, m_{\delta 15\text{NAIR} (\text{‰})}, m_{\delta 34\text{SV-CDT} (\text{‰})}, m_{\delta 13\text{CV-PDB}^2},$$

$$m_{\delta 15\text{NAIR} (\text{‰})^2}, m_{\delta 34\text{SV-CDT} (\text{‰})^2}, c) \quad (9)$$

where $m_{\delta 13\text{CV-PDB}}$, $m_{\delta 15\text{NAIR} (\text{‰})}$, $m_{\delta 34\text{SV-CDT} (\text{‰})}$, $m_{\delta 13\text{CV-PDB}^2}$, $m_{\delta 15\text{NAIR} (\text{‰})^2}$, and $m_{\delta 34\text{SV-CDT} (\text{‰})^2}$ are the model parameters (i.e., multipliers of the features), c is the intercept parameter, and $f(\cdot)$ is one of the linear functions that are given in the Appendix A. This is a generic family of multiple models with similar structure, but different on how the ratios of stable isotopes influence the probability of characterization that a sample originates from Messolongi, Greece.

Equations (1)–(7) implement the prior knowledge we introduce to the model, and we have chosen very wide noninformative ones. This way, we let the data, solely, define them [32,33]. Equations (8) and (9) define the likelihood of the model, a distribution function assigned to an observed variable, where y_i is the target variable of origin [34,35]. Equations (1)–(9) are the Bayesian formulation of Generalized Linear Models with binomial link functions [32].

2.3. Model Fitting Methodology

The family of models described in the previous section varies in the number of parameters from two to seven depending on the number of features we use at the linear function $f(\cdot)$. Since we need to quantify the model uncertainty (Section 2.1), we estimate the whole distribution of its parameters. This can be conducted by using the Bayes formula:

$$p(\theta|y_i) = \frac{p(y_i|\theta) * p(\theta)}{p(y_i)} \quad (10)$$

where $p(y_i|\theta)$ is the likelihood function of Equations (8) and (9), $p(\theta)$ are the prior probabilities of the parameters, and $p(y_i)$ is the marginal likelihood [32–35]. The likelihood function links the observations/measurements to the parameters, and it is multiplied by the prior distributions (i.e., beliefs before taken any measurements), posterior distribution of the unknown parameter is obtained (distributions incorporate the model uncertainty). Although this is conceptually simple, in practice it is challenging to compute due to the normalization factor $p(y_i)$. If we re-write the marginal likelihood as $p(y_i) = \int_{\Theta} p(y_i|\theta) * p(\theta) d(\theta)$, it is clear that it is needed to be integrated over the domain of Θ , i.e., all possible values of θ , which is demanding.

The general modern computational approach to calculate the posterior probabilities of model parameters avoids the labor-intensive approximation of $p(y_i)$, and to draw samples directly from $p(\theta|y_i)$. This computational efficient technique, called Markov Chain Monte Carlo (MCMC), constitutes a cornerstone of probabilistic machine learning and Bayesian statistics [32,33]. Recent computational frameworks of probabilistic machine learning and Bayesian statistics [36,37], where high accuracy and speed are critical factors, incorporate gradient-based information from $\nabla \log p(\theta|y_i)$. Those algorithms are called Hamiltonian Monte Carlo (HMC) and can significantly speed-up the calculations [38].

By using MCMC and HMC, we can efficiently estimate the distribution of all model parameters and quantify their and model's uncertainty. Moreover, this Bayesian treatment of the candidate models enables us to select the best performing on for future bottarga samples (see next section). Due to the very small number of the samples, we cannot perform the usual cross-validation as we would have done in a traditional machine learning model [30,39].

2.4. Model Comparison

From the list of potential functions in Appendix A, if we choose one of those and substitute it to Equation (9), it is shown that the model complexity varies from two to seven parameters. A first selection of a model that fits better to the measurements of stable isotope ratios analyzed, i.e., with the smallest error residual, would favor the model(s) of larger complexity. Such a model would indeed describe the current samples with the smallest total error, but it would be very likely not a suitable one describing the general population of the bottarga. In machine learning, this is described as model overfitting and must be avoided. On the other extreme, if a very simple model is selected, then it could probably have the capacity to perform well, a situation known as underfitting [29–32,39].

Therefore, the desired models could use limited sample data to fit/estimate their parameters, but, at the same time, should be capable of generalizing beyond the sample (i.e., they describe the population based on the samples) and performing well. In the common machine learning models and algorithms, such as decision trees, random forests, boosting trees, and neural networks [31], the usual practice to avoid overfitting is to perform cross-validation. By splitting the data into training/evaluation parts, we fit the model based on training data only, and we measure the performance on the data the model has never seen before (evaluation part), simulating the model performance on the general population. Unfortunately, when the available data are extremely small, as in this case where the production is very limited and the measurement destructive, cross-validation is not an option.

This is a very challenging situation, and we need to turn it into theoretical aspects of stochastic modeling. Information theory and information-based criteria help us select the best performing model [33,34]. The seminal work of S. Watanabe [40] provides a model comparison mechanism called Widely Applicable Information Criterion (WAIC) for models that belong to the same family. WAIC provides an approximation of the out-of-sample deviance, and according to its value, we select the best-expected model from Appendix A (i.e., the model of lower WAIC).

2.5. Incorporating Measurement Error

Small variations might appear during preparation and measurement procedure of the bottarga samples. Even if these variations are expected to be limited in a laboratory routine/protocol where equipment is well-calibrated and the sample preparation process stable, we should take this variation into account in our decision/characterization methodology. For this reason, each bottarga sample was analyzed and measured five (5) times independently. Under the assumption, that future samples (measured in different laboratories with similar methods) will demonstrate similar variability, the effect of measurement variations could be incorporated into the methodology by considering them as perturbations of the features.

The proposed generative model of the perturbations has the following form:

$$\sigma \sim \text{halfN}(0, 1) \quad (11)$$

$$d_i \sim N(0, \sigma) \quad (12)$$

where d_i is the measurement residual when we remove the group mean value, and $\text{halfN}(\cdot)$ is a truncated normal distribution without the negative values of the random variable. In principle, this is a model of the standard deviation of the measurements with very mild prior probability (this mild prior comes from the assumption/knowledge that the laboratory setup is stable and equipment well calibrated). Once, we compute the posterior probability distribution of σ using MCMC based on the data of Section 3.1, we generate perturbations based on the $N(0, \sigma_i)$ distribution.

2.6. Characterization/Decision Methodology

By estimating the distributions of the model parameters described in Equations (1)–(9) based on measured bottarga samples in the laboratory by MCMC/HMC, we are able to draw parameter samples from the chains of the best model obtained from the comparison of Section 2.4 (note: here “sample” term has a dual meaning; bottarga samples and parameter samples which are vectors of numbers we get if probe/sample the posterior distributions). This way we can obtain tens of thousands of parameter values, that capture the probabilistic nature and the associated uncertainty regarding their exact/real value. These parameter combinations, when substituted to $\frac{1}{1+e^{-z_i}}$ for a z_i of our choice based on the methodology of next section, yields the probability the model “believes” a bottarga sample comes from Messolongi.

Parameter combinations may differ regarding the yielding probability of a bottarga sample. We report this difference as characterization/decision uncertainty. We set a decision threshold α (e.g., 95%); probability above this value is considered as “high confidence of Messolongi origin” for the bottarga sample. If we apply this rule to all the tens of thousands of model instances which was discussed in the previous paragraph, we obtain an equal number of model beliefs. Then we aggregate those into a single number, that we interpret as the characterization probability (or confidence level) of being from Messolongi. A characterization probability above α , can be considered almost certain.

3. Numerical Results

3.1. Dataset: Measured Bottarga Samples

It has been already mentioned that the bottarga samples come from four different areas: two of them located in Greece namely Messolongi and Preveza. Both areas are producing Bottarga products, however only the Bottarga originating in Messolongi has been registered since 1996 as a Protected Designations of Origin product (PDO-GR-0446) with the name “Avgotaracho Messolongiou”. The rest of samples come from two other regions of the world (Australia and Mauritania). The samples were labeled by codes “A1” to “A4”, see Table 1. Bottarga production in the geographical region of Messolongi is very limited and consequently expensive; thus, a random sampling scheme of five samples (derived from batches with different lot numbers) has been used. For regions outside Messolongi, a random selection performed on a retailer level trying to pick up different samples coming from different lot number batches.

The measured mean values of each area are given in the following Table 2. The scatter plots of the mean value of each sample are shown in Figure 1 (each dot corresponds to the mean values of a sample), while the main diagonal is the kernel density estimation of each pair of the primary features. A clear separation of bottarga samples (orange dots) from Messolongi is observed.

Table 2. Bottarga Analysis Results (mean δ values and standard error).

Code	Area	Mean $\delta^{15}\text{N}_{\text{AIR}}$ (‰)	Mean $\delta^{13}\text{C}_{\text{V-PDB}}$	Mean $\delta^{34}\text{S}_{\text{V-CDT}}$ (‰)
A1	Mauritania	9.6442 (0.0073)	−19.0075 (0.0084)	16.0344 (0.0109)
A2	Preveza Greece	10.8551 (0.0146)	−18.0034 (0.0187)	15.1658 (0.0073)
A3	Messolongi Greece	7.5246 (0.0056)	−18.4259 (0.0068)	11.9966 (0.0135)
A4	Australia	5.6429 (0.0050)	−14.5148 (0.0088)	4.5341 (0.0094)

Table 3 shows the pairwise Pearson correlation coefficients of the mean values of all samples [41]. It is obvious that there is a very strong linear association between all combinations of primary features; this is a strong indication of multicollinearity [34]. A practical interpretation would be that additional ratios of stable isotopes might not add predictive power to the model.

Table 3. Pearson coefficient correlation matrix of all measured bottarga samples.

	$\delta^{15}\text{N}_{\text{AIR}}$ (‰)	$\delta^{13}\text{C}_{\text{V-PDB}}$ (‰)	$\delta^{34}\text{S}_{\text{V-CDT}}$ (‰)
$\delta^{15}\text{N}_{\text{AIR}}$ (‰)	1.00	−0.85	0.95
$\delta^{13}\text{C}_{\text{V-PDB}}$ (‰)	−0.85	1.00	−0.97
$\delta^{34}\text{S}_{\text{V-CDT}}$ (‰)	0.95	−0.97	1.00

3.2. Proposed Probabilistic Machine Learning for the Characterization of Protected Designation of Origin of Bottarga from Messolongi

By using the family of models described in Equations (1)–(9), the data (i.e., mean value of subsamples per sample), and MCMC/HMC (software framework PYMC3 [36]), we estimate the posterior probability distributions of all candidate models in Appendix A. The sampler of choice is NUTS [38], we use two chains for each model, and we draw 15,000 parameter samples (after 3000 warming up samples). Based on the model comparison methodology of Section 2.4, the proposed probabilistic machine learning model for the characterization of protected designation of origin of bottarga from Messolongi is the following one:

$$m_{\delta^{34}\text{S}_{\text{V-CDT}} (\text{‰})^2} \sim \text{N}(0, \infty) \quad (13)$$

$$c \sim \text{N}(0, \infty) \quad (14)$$

$$y_i \sim B\left(n, \frac{1}{1 + e^{-m_{\delta^{34}\text{S}_{\text{V-CDT}} (\text{‰})^2} * \delta^{34}\text{S}_{\text{V-CDT}} (\text{‰})^2 + c}}\right) \quad (15)$$

This is a model that uses the squared value of $\delta^{34}\text{S}_{\text{V-CDT}}$ (‰) and an intercept in the linear part of the logistic function (15) to define the probability of success of the Bernoulli trials. The presence of a single ratio of stable isotopes out of three measured, confirms the observation of the strong linear association among them (Figure 2).

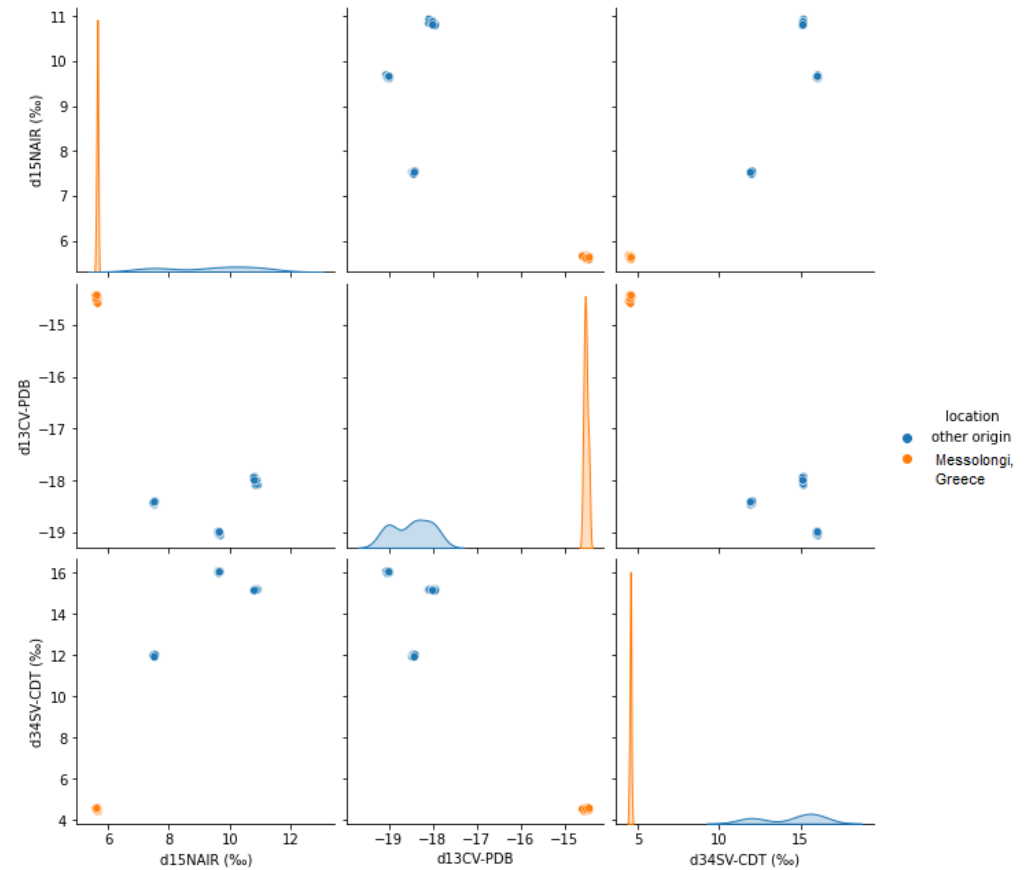


Figure 2. Scatter plots of all measured bottarga samples. Main diagonal shows kernel density estimations.

The posterior probability distributions of the model are shown in Figure 3. We can see a very wide range for both parameters, which is expected due to the very small number of data points. The two lines of Figure 3 correspond to the two chains of MCMC. As we can see in Figure 4, the chains show no signs of autocorrelation (it is confirmed and by the value of $\hat{\rho} \approx 1.00$), which verifies the correctness of the parameter estimation.

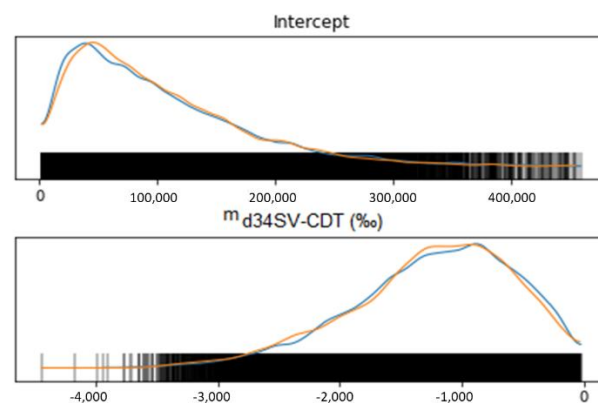


Figure 3. Posterior probability distributions of the proposed characterization model.

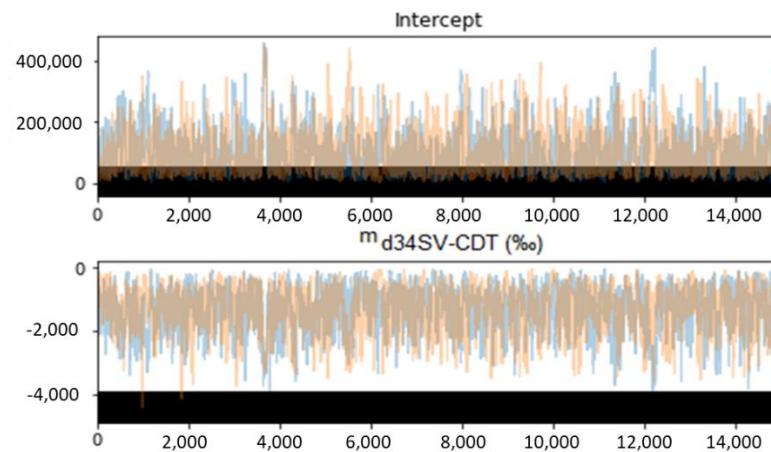


Figure 4. Trace plots of the proposed characterization model.

3.3. Uncertainty of the Proposed Model

A wide range of model parameter distributions, as seen in Figure 3, is expected to introduce significant uncertainty to the model output. Furthermore, it is expected that the spatial distribution be nonuniform across $\delta^{34}\text{S}_{\text{V-CDT}} (\text{‰})$ with higher uncertainty at the positions of sparse measurements. If many parameter pairs are drawn, as described in Section 2.4 (e.g., 30,000 parameter samples) and evaluated on a range of $\delta^{34}\text{S}_{\text{V-CDT}} (\text{‰})$ [4, 17.5] with step 1, the variability of the model output could be visualized. This variability is shown via violin plots (see [12]) in Figure 5 and provides the whole picture regarding the model uncertainty. Violin plots are shown for integer values of $\delta^{34}\text{S}_{\text{V-CDT}} (\text{‰})$, i.e., 4, 5 . . . 17, when all models of 30,000 parameter samples, agree on either 1 or 0, we do not see any grey area because there is no variability. On the other hand, at regions of higher uncertainty, a split of the distribution is observed in two segments (i.e., some model parameters opt for 0 and others for 1). In addition, low values of $\delta^{34}\text{S}_{\text{V-CDT}} (\text{‰})$ (where samples from Messolongi lay), have very low uncertainty (grey areas) and the blue/orange line of 10%/90% probability coincide (the green line corresponds to 50%). The larger uncertainty is reported from 5 to 11. For $\delta^{34}\text{S}_{\text{V-CDT}} (\text{‰})$ larger than 12, where are the values of bottarga samples from other origins, the uncertainty is low again. The vertical dotted black and red lines indicate the ranges of $\delta^{34}\text{S}_{\text{V-CDT}} (\text{‰})$ for samples from Messolongi and other regions respectively (the two black lines are so close that are seen as a single thick vertical line).

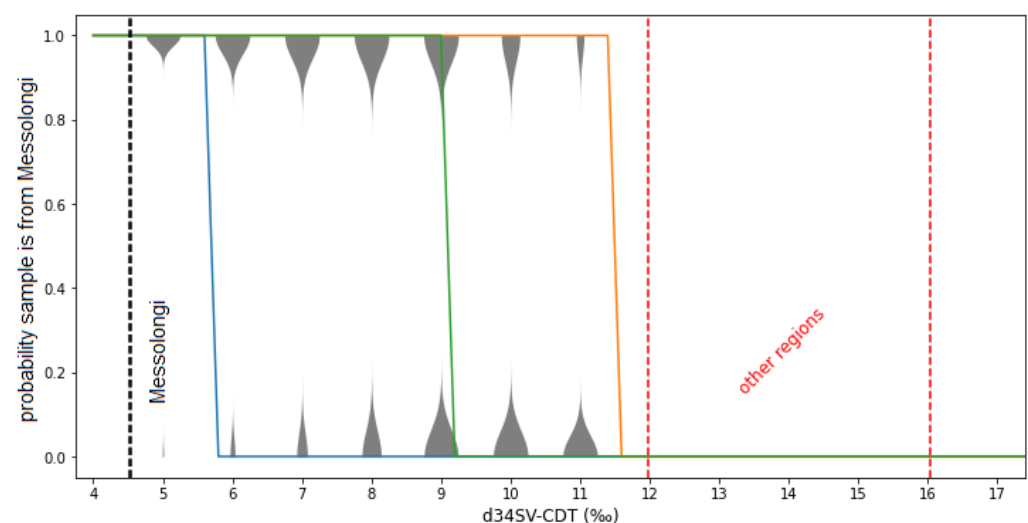


Figure 5. Uncertainty distribution of the proposed characterization model.

3.4. Decision Diagram

The proposed characterization model of Equations (13)–(15) in combination with the computed posterior probability parameter distributions of Figure 3, perform as the main tools to develop a decision/characterization mechanism for the bottarga samples. As described in Section 2.4, we draw samples (pairs of $m_{d34SV-CDT} (\%)^2$ and intercept) from the parameter distributions (30,000 in this study), we substitute their values in Equations (13)–(15), and subsequently we obtain an ensemble of models. Then we evaluate all those 30,000 models for a range of values for $\delta^{34}S_{V-CDT} (\%)$. By using decision threshold $\alpha = 0.95$, we evaluate and aggregate decisions of the ensemble for $\delta^{34}S_{V-CDT} (\%)$ and step 0.2 as shown in Figure 6 (blue line plots the confidence level).

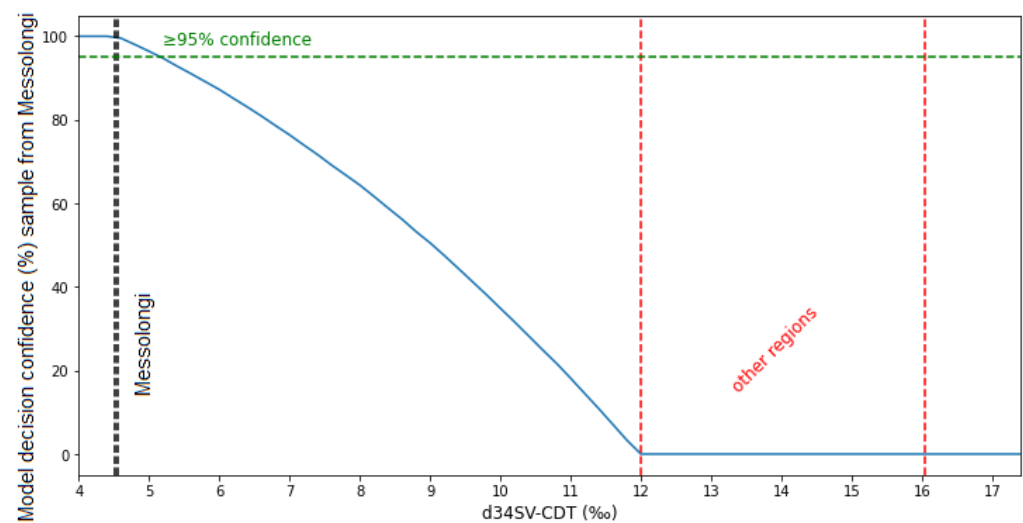


Figure 6. Decision confidence level of probabilistic model of a sample has been collected from Messolongi (blue line).

This characterization methodology consolidates the uncertainty of the model ensemble and transforms it into a level of confidence. For example, for measured value $\delta^{34}S_{V-CDT} (\%) = 11$ the sample has a probability of only 20% of having it collected from Missolonghi. This 20% probability has been computed by the ratio “number models that return probability $\geq a$ ”/“size of the ensemble”. If we adopt a conservative approach, and seek at least 95% model confidence, the critical value for $\delta^{34}S_{V-CDT} (\%)$ is around 5.2. As we see in Figure 6, at this value the blue line of model confidence crosses the green dotted line of 95%.

In Section 2.5 the incorporation of measurement error into the model was discussed. Variations of bottarga sample preparation and equipment might be a source of small differences among subsamples. This is confirmed when we estimate the parameters of the model described in Equations (11) and (12). In Figure 7 we present the distribution of standard deviation, and we conclude that is small but not negligible. By considering it, the confidence level of the probabilistic model decision that a sample has been collected from Messolongi is shown in Figure 8. The error bar (small vertical lines of the maximum range $\sim 2\%$) of the model decision is larger for the values of $\delta^{34}S_{V-CDT} (\%)$ where we do not have bottarga samples, and thus uncertainty becomes higher (i.e., $\delta^{34}S_{V-CDT} (\%) \in [5.5, 11.5]$).

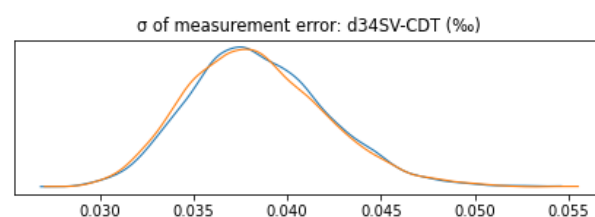


Figure 7. Distribution of standard deviation of measurement error.

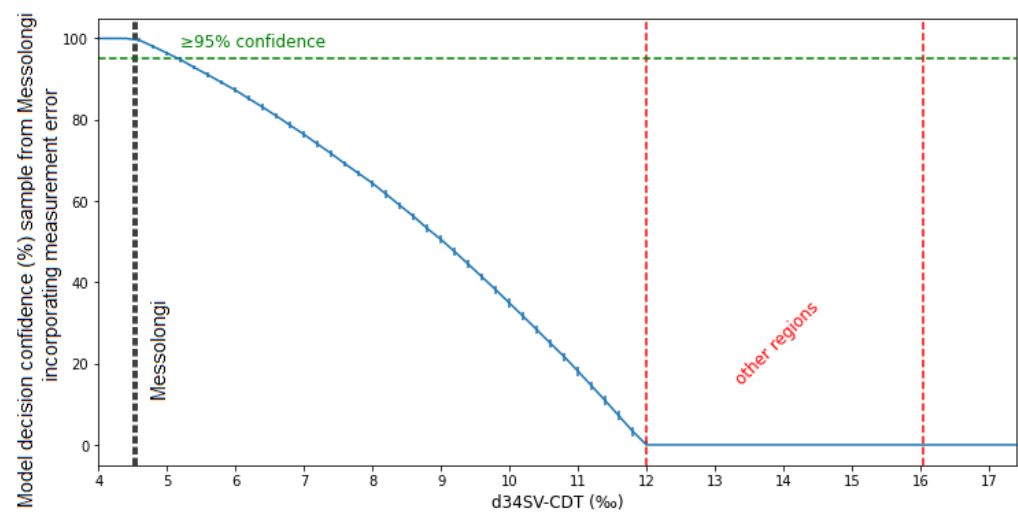


Figure 8. Decision confidence level of probabilistic model that incorporates measurement error of a sample has been collected from Messolongi (blue line).

3.5. Decision Table

Decision diagrams of the previous section show the whole picture of the probability a sample has been collected from Messolongi. However, when it comes to certification of samples of unknown origin, an easy and reliable procedure is required. We propose a tabular lookup table that gives the confidence level based on $\delta^{34}\text{S}_{\text{V-CDT}} (\text{‰})$ measured values. We have two options regarding $\delta^{34}\text{S}_{\text{V-CDT}} (\text{‰})$ measurements: (a) we divide into five or more subsamples and compute the mean value of them, and (b) measure only once the bottarga sample. Then we search for the row with the nearest value in the first column of Table A2 of Appendix B. If we have chosen option (a) the confidence level is the number reported in the second column. For option (b), the confidence level is a number between the last two columns. We propose to certify a sample of bottarga as protected designation of origin from Messolongi for values equal to or higher than 95%.

4. Discussion

The characterization of protected designation of origin is a critical factor for the viability of local products where the increased production cost is counterbalanced by a higher price. At the same time, PDO products become the target of fraudsters who want to benefit from the higher value, and fake products of different origins. Fraudsters hide behind the “declarative” nature of the designation characterization. With the proposed model and methodology, we quantified the clear difference of the bottarga from Messolongi in general, while at the same time we turned it to decision mechanism for the certification of individual samples.

Stable isotope ratios of $\delta^{34}\text{S}_{\text{V-CDT}} (\text{‰})$ are very distinct for the case of the Bottarga originated from the area of Messolongi. Carbon and nitrogen stable isotopes depend on nutrient uptake, digestion, and metabolism of the organisms [42–44]. Hence, the diet of each organism can explain the differences in these results. On the other hand, the $\delta^{34}\text{S}_{\text{V-CDT}}$ ranged from 4.53‰ to 16.03‰ which are typical values for marine fish. Similar results for $\delta^{34}\text{S}_{\text{V-CDT}}$ (4.6–20.1‰) have been reported from Sayle et al. (2013) [45]. However, the lowest value has been reported for Messolongi samples. Messolongi lagoon is a shallow area connected with two narrow openings in the north side with Etoliko lagoon and in the south with Gulf of Patras allowing seawater influx in the lagoons. In addition, two major rivers of the area (Acheloos and Evinos), have also participated in the formation of the lagoons through the deposition of sediment. Towns and highly cultivated agricultural areas in the vicinity of the lagoons produce an influx of sewage and irrigation run-off rich in nutrients. The influx of nutrients causes a high oxygen demand during decomposition of the organic matter. Moreover, large gypsum ($\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$) deposits to the north-west

of the Etoliko lagoon generate hydrogen sulfide in the hypolimnion (gypsum dissolves in water to produce sulphate ions $(\text{SO}_4)^{2-}$ which under anaerobic conditions are reduced to elemental sulfur and hydrogen sulfide). The organic and inorganic material rich in sulfur compounds is broken down by bacteria present on the lagoon bottom producing sulfides. The bacteria under anaerobic conditions produce S^{-2} , HS^{-} , and H_2S with the reduction of sulfur. Hydrogen sulfide can either be oxidized to elemental sulfur by sulfur-bacteria or be released into the surrounding water [46,47]. This special condition affects the ratio of sulfur isotopes. According to Winner et. Al., irradiated plants with roots immersed in $(\text{HSO}_3)^{-}$ and $(\text{SO}_4)^{2-}$ solutions at concentrations found in nature emit H_2S from their leaves, and that H_2S emitted contained more ^{32}S and less ^{34}S , namely, lower $\delta^{34}\text{S}_{\text{V-CDT}}$ values, in comparison to irradiation solutions [48]. Therefore, the high H_2S concentration present in the Etoliko and Messolongi lagoon's bottom layer has, as a result, low $\delta^{34}\text{S}_{\text{V-CDT}}$ (‰) values.

The forementioned reason explains the great difference of Messolongi $\delta^{34}\text{S}_{\text{V-CDT}}$ values from the corresponding values of the other areas helping to the authentication of bottarga products originated from Messolongi by using a probabilistic machine learning approach to cope with small number of samples.

Author Contributions: Conceptualization, G.T., A.P. and A.L.; methodology, E.P., E.C.M., K.K. and A.-A.T.; software, G.T., A.-A.T. and A.Z.; validation, E.P., E.C.M. and K.K. formal analysis, G.T. and A.K.; investigation, E.P., E.C.M. and K.K.; resources, A.P. and A.L.; data curation, E.C.M., A.Z., E.P. and G.T.; writing—original draft preparation, G.T., A.-A.T. and A.K.; writing—review and editing G.T., A.P., A.K. and A.L.; visualization, A.K., G.T. and A.Z.; supervision, A.P. and A.L.; project administration, A.L.; funding acquisition, A.P., A.K. and A.L. All authors have read and agreed to the published version of the manuscript.

Funding: Part of this work was funded by the PRIMA program named MEDFOODTTHUBS supported by the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No 1931.



Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. List of different linear combinations that are used in the likelihood function of the proposed generic model.

Formula of Linear Part
$m_{\text{d15NAIR}} (\text{‰}) * \text{d15NAIR} (\text{‰}) + c$
$m_{\text{d13CV-PDB}} \text{d13CV} - \text{PDB} + c$
$m_{\text{d34SV-CDT}} (\text{‰}) \text{d34SV} - \text{CDT} (\text{‰}) + c$
$m_{\text{d15NAIR}} (\text{‰})^2 * \text{d15NAIR} (\text{‰})^2 + c$
$m_{\text{d13CV-PDB}^2} \text{d13CV} - \text{PDB}^2 + c$
$m_{\text{d34SV-CDT}} (\text{‰})^2 \text{d34SV} - \text{CDT} (\text{‰})^2 + c$
$m_{\text{d15NAIR}} (\text{‰}) * \text{d15NAIR} (\text{‰}) + m_{\text{d13CV-PDB}} \text{d13CV} - \text{PDB} + c$

Table A1. *Cont.*[illegible]

Table A1. Cont.

Formula of Linear Part
$m_{d34SV-CDT}(\%)^2 \cdot d34SV - CDT(\%) + m_{d13CV-PDB^2} \cdot d13CV - PDB^2 + m_{d34SV-CDT}(\%)^2 \cdot d34SV - CDT(\%)^2 + c$
$m_{d15NAIR}(\%)^2 * d15NAIR(\%)^2 + m_{d13CV-PDB^2} \cdot d13CV - PDB^2 + m_{d34SV-CDT}(\%)^2 \cdot d34SV - CDT(\%)^2 + c$
$m_{d15NAIR}(\%) * d15NAIR(\%) + m_{d13CV-PDB} \cdot d13CV - PDB + m_{d34SV-CDT}(\%) \cdot d34SV - CDT(\%) + m_{d15NAIR}(\%)^2 * d15NAIR(\%)^2 + c$
$m_{d15NAIR}(\%) * d15NAIR(\%) + m_{d13CV-PDB} \cdot d13CV - PDB + m_{d34SV-CDT}(\%) \cdot d34SV - CDT(\%) + m_{d13CV-PDB^2} \cdot d13CV - PDB^2 + c$
$m_{d15NAIR}(\%) * d15NAIR(\%) + m_{d13CV-PDB} \cdot d13CV - PDB + m_{d34SV-CDT}(\%) \cdot d34SV - CDT(\%) + m_{d34SV-CDT}(\%)^2 \cdot d34SV - CDT(\%)^2 + c$
$m_{d15NAIR}(\%) * d15NAIR(\%) + m_{d13CV-PDB} \cdot d13CV - PDB + m_{d15NAIR}(\%)^2 * d15NAIR(\%)^2 + m_{d13CV-PDB^2} \cdot d13CV - PDB^2 + c$
$m_{d15NAIR}(\%) * d15NAIR(\%) + m_{d13CV-PDB} \cdot d13CV - PDB + m_{d15NAIR}(\%)^2 * d15NAIR(\%)^2 + m_{d34SV-CDT}(\%)^2 \cdot d34SV - CDT(\%)^2 + c$
$m_{d15NAIR}(\%) * d15NAIR(\%) + m_{d13CV-PDB} \cdot d13CV - PDB + m_{d13CV-PDB^2} \cdot d13CV - PDB^2 + m_{d34SV-CDT}(\%)^2 \cdot d34SV - CDT(\%)^2 + c$
$m_{d15NAIR}(\%) * d15NAIR(\%) + m_{d34SV-CDT}(\%) \cdot d34SV - CDT(\%) + m_{d15NAIR}(\%)^2 * d15NAIR(\%)^2 + m_{d13CV-PDB^2} \cdot d13CV - PDB^2 + c$
$m_{d15NAIR}(\%) * d15NAIR(\%) + m_{d34SV-CDT}(\%) \cdot d34SV - CDT(\%) + m_{d15NAIR}(\%)^2 * d15NAIR(\%)^2 + m_{d34SV-CDT}(\%)^2 \cdot d34SV - CDT(\%)^2 + c$
$m_{d15NAIR}(\%) * d15NAIR(\%) + m_{d34SV-CDT}(\%) \cdot d34SV - CDT(\%) + m_{d13CV-PDB^2} \cdot d13CV - PDB^2 + m_{d34SV-CDT}(\%)^2 \cdot d34SV - CDT(\%)^2 + c$
$m_{d15NAIR}(\%) * d15NAIR(\%) + m_{d15NAIR}(\%)^2 * d15NAIR(\%)^2 + m_{d13CV-PDB^2} \cdot d13CV - PDB^2 + m_{d34SV-CDT}(\%)^2 \cdot d34SV - CDT(\%)^2 + c$
$m_{d13CV-PDB} \cdot d13CV - PDB + m_{d34SV-CDT}(\%) \cdot d34SV - CDT(\%) + m_{d15NAIR}(\%)^2 * d15NAIR(\%)^2 + m_{d13CV-PDB^2} \cdot d13CV - PDB^2 + c$
$m_{d13CV-PDB} \cdot d13CV - PDB + m_{d34SV-CDT}(\%) \cdot d34SV - CDT(\%) + m_{d15NAIR}(\%)^2 * d15NAIR(\%)^2 + m_{d34SV-CDT}(\%)^2 \cdot d34SV - CDT(\%)^2 + c$
$m_{d13CV-PDB} \cdot d13CV - PDB + m_{d34SV-CDT}(\%) \cdot d34SV - CDT(\%) + m_{d13CV-PDB^2} \cdot d13CV - PDB^2 + m_{d34SV-CDT}(\%)^2 \cdot d34SV - CDT(\%)^2 + c$
$m_{d13CV-PDB} \cdot d13CV - PDB + m_{d15NAIR}(\%)^2 * d15NAIR(\%)^2 + m_{d13CV-PDB^2} \cdot d13CV - PDB^2 + m_{d34SV-CDT}(\%)^2 \cdot d34SV - CDT(\%)^2 + c$
$m_{d34SV-CDT}(\%) \cdot d34SV - CDT(\%) + m_{d15NAIR}(\%)^2 * d15NAIR(\%)^2 + m_{d13CV-PDB^2} \cdot d13CV - PDB^2 + m_{d34SV-CDT}(\%)^2 \cdot d34SV - CDT(\%)^2 + c$
$m_{d15NAIR}(\%) * d15NAIR(\%) + m_{d13CV-PDB} \cdot d13CV - PDB + m_{d34SV-CDT}(\%) \cdot d34SV - CDT(\%) + m_{d15NAIR}(\%)^2 * d15NAIR(\%)^2 + m_{d13CV-PDB^2} \cdot d13CV - PDB^2 + c$
$m_{d15NAIR}(\%) * d15NAIR(\%) + m_{d13CV-PDB} \cdot d13CV - PDB + m_{d34SV-CDT}(\%) \cdot d34SV - CDT(\%) + m_{d15NAIR}(\%)^2 * d15NAIR(\%)^2 + m_{d34SV-CDT}(\%)^2 \cdot d34SV - CDT(\%)^2 + c$

Table A1. Cont.

Formula of Linear Part
$m_{d15NAIR} (\text{‰}) * d15NAIR (\text{‰}) + m_{d13CV-PDB} d13CV - PDB + m_{d34SV-CDT} (\text{‰}) d34SV - CDT (\text{‰}) + m_{d13CV-PDB^2} d13CV - PDB^2 + m_{d34SV-CDT} (\text{‰})^2 d34SV - CDT (\text{‰})^2 + c$
$m_{d15NAIR} (\text{‰}) * d15NAIR (\text{‰}) + m_{d13CV-PDB} d13CV - PDB + m_{d15NAIR} (\text{‰})^2 * d15NAIR (\text{‰})^2 + m_{d13CV-PDB^2} d13CV - PDB^2 + m_{d34SV-CDT} (\text{‰}) d34SV - CDT (\text{‰})^2 + c,$
$m_{d15NAIR} (\text{‰}) * d15NAIR (\text{‰}) + m_{d34SV-CDT} (\text{‰}) d34SV - CDT (\text{‰}) + m_{d15NAIR} (\text{‰})^2 * d15NAIR (\text{‰})^2 + m_{d13CV-PDB^2} d13CV - PDB^2 + m_{d34SV-CDT} (\text{‰})^2 d34SV - CDT (\text{‰})^2 + c,$
$m_{d13CV-PDB} d13CV - PDB + m_{d34SV-CDT} (\text{‰}) d34SV - CDT (\text{‰}) + m_{d15NAIR} (\text{‰})^2 * d15NAIR (\text{‰})^2 + m_{d13CV-PDB^2} d13CV - PDB^2 + m_{d34SV-CDT} (\text{‰})^2 d34SV - CDT (\text{‰})^2 + c,$
$m_{d15NAIR} (\text{‰}) * d15NAIR (\text{‰}) + m_{d13CV-PDB} d13CV - PDB + m_{d34SV-CDT} (\text{‰}) d34SV - CDT (\text{‰}) + m_{d15NAIR} (\text{‰})^2 * d15NAIR (\text{‰})^2 + m_{d13CV-PDB^2} d13CV - PDB^2 + m_{d34SV-CDT} (\text{‰})^2 d34SV - CDT (\text{‰})^2 + c$

Appendix B

Table A2. Decision table.

d34SV-CDT (‰)	Confidence % (Based on the Mean of Multiple Measurements of a Sample)	Lower Limit of Confidence % (Due to Measurement Error)	Upper Limit of Confidence % (Due to Measurement Error)
4.00	100.00	100.00	100.00
4.20	100.00	100.00	100.00
4.40	100.00	100.00	100.00
4.60	99.54	99.05	100.00
4.80	98.00	97.52	98.50
5.00	96.40	95.84	96.94
5.20	94.75	94.15	95.30
5.40	92.83	92.33	93.48
5.60	91.01	90.39	91.61
5.80	89.13	88.60	89.72
6.00	87.25	86.63	87.90
6.20	85.13	84.52	85.89
6.40	83.06	82.45	83.80
6.60	80.89	80.22	81.65
6.80	78.61	77.93	79.44
7.00	76.39	75.67	77.16
7.20	73.99	73.28	74.86
7.40	71.65	70.86	72.38
7.60	69.13	68.38	69.91
7.80	66.76	66.00	67.50
8.00	64.37	63.55	65.22
8.20	61.68	60.78	62.58
8.40	58.92	58.10	59.77
8.60	56.21	55.29	57.05

Table A2. *Cont.*

d34SV—CDT (‰)	Confidence % (Based on the Mean of Multiple Measurements of a Sample)	Lower Limit of Confidence % (Due to Measurement Error)	Upper Limit of Confidence % (Due to Measurement Error)
8.80	53.24	52.41	54.24
9.00	50.54	49.68	51.46
9.20	47.54	46.58	48.55
9.40	44.44	43.43	45.49
9.60	41.31	40.33	42.31
9.80	38.16	37.14	39.19
10.00	34.89	33.88	35.97
10.20	31.65	30.59	32.74
10.40	28.32	27.33	29.35
10.60	24.95	23.93	26.08
10.80	21.69	20.62	22.75
11.00	18.15	16.96	19.33
11.20	14.49	13.49	15.64
11.40	10.87	9.87	12.10
11.60	7.10	5.95	8.36
11.80	3.32	2.18	4.50
12.00	0.00	0.00	0.85
12.20	0.00	0.00	0.00
12.40	0.00	0.00	0.00
12.60	0.00	0.00	0.00
12.80	0.00	0.00	0.00
13.00	0.00	0.00	0.00
13.20	0.00	0.00	0.00
13.40	0.00	0.00	0.00
13.60	0.00	0.00	0.00
13.80	0.00	0.00	0.00
14.00	0.00	0.00	0.00
14.20	0.00	0.00	0.00
14.40	0.00	0.00	0.00
14.60	0.00	0.00	0.00
14.80	0.00	0.00	0.00
15.00	0.00	0.00	0.00
15.20	0.00	0.00	0.00
15.40	0.00	0.00	0.00
15.60	0.00	0.00	0.00
15.80	0.00	0.00	0.00

Table A2. Cont.

d34SV—CDT (‰)	Confidence % (Based on the Mean of Multiple Measurements of a Sample)	Lower Limit of Confidence % (Due to Measurement Error)	Upper Limit of Confidence % (Due to Measurement Error)
16.00	0.00	0.00	0.00
16.20	0.00	0.00	0.00
16.40	0.00	0.00	0.00
16.60	0.00	0.00	0.00
16.80	0.00	0.00	0.00
17.00	0.00	0.00	0.00

References

- Chaguri, M.P.; Maulvault, A.L.; Nunes, M.L.; Santiago, D.A.; Denadai, J.C.; Fogaça, F.H.; Sant'Ana, L.S.; Ducatti, C.; Bandarra, N.; Carvalho, M.L.; et al. Different tools to trace geographic origin and seasonality of croaker (*Micropogonias furnieri*). *LWT—Food Sci. Technol.* **2015**, *61*, 194–200. [\[CrossRef\]](#)
- Pascoal, A.; Barros-Velazquez, J.; Cepeda, A.; Gallardo, J.M.; Calo-Mata, P. Survey of the authenticity of prawn and shrimp species in commercial food products by PCR-RFLP analysis of a 16S rRNA/tRNAVal mitochondrial region. *Food Chem.* **2008**, *109*, 638–646. [\[CrossRef\]](#)
- Spink, J.; Moyer, D.C. Defining the public health threat of food fraud. *J. Food Sci.* **2011**, *76*, 157–163. [\[CrossRef\]](#)
- Tacon, A.G.J.; Metian, M. Fish matters: Importance of aquatic foods in human nutrition and global food supply. *Rev. Fish. Sci.* **2013**, *21*, 22–38. [\[CrossRef\]](#)
- Mazzeo, M.F.; De Giulio, B.; Guerriero, G.; Ciarcia, G.; Malorni, A.; Russo, G.L.; Siciliano, R.A. Fish authentication by MALDI-TOF mass spectrometry. *Agric. Food Chem.* **2008**, *56*, 11071–11076. [\[CrossRef\]](#)
- Mansfield, B. Is fish health food or poison? Farmed fish and the material production of un/healthy nature. *Antipode* **2011**, *43*, 413–434. [\[CrossRef\]](#)
- Ortea, I.; Gallardo, J.M. Investigation of production method, geographical origin and species authentication in commercially relevant shrimps using stable isotope ratio and/or multi-element analyses combined with chemometrics: An exploratory analysis. *Food Chem.* **2015**, *170*, 145–153. [\[CrossRef\]](#)
- Camin, F.; Bontempo, L.; Perini, M.; Piasentier, E. Stable isotope ratio analysis for assessing the authenticity of food of animal origin. *Compr. Rev. Food Sci. Food Saf.* **2016**, *15*, 868–877. [\[CrossRef\]](#) [\[PubMed\]](#)
- Gopi, K.; Mazumder, D.; Sammut, J.; Saintilan, N.; Crawford, J.; Gadd, P. Isotopic and elemental profiling to trace the geographic origins of farmed and wild-caught Asian seabass (*Lates calcarifer*). *Aquaculture* **2019**, *502*, 56–62. [\[CrossRef\]](#)
- Piras, C.; Scano, P.; Locci, E.; Sanna, R.; Marincola, F.C. Analysing the effects of frozen storage and processing on the metabolite profile of raw mullet roes using ¹H NMR spectroscopy. *Food Chem.* **2014**, *159*, 71–79. [\[CrossRef\]](#) [\[PubMed\]](#)
- Katerinopoulou, K.; Kontogeorgos, A.; Salmas, C.E.; Patakas, A.; Ladavos, A. Geographical origin authentication of agri-food products: A review. *Foods* **2020**, *9*, 489. [\[CrossRef\]](#)
- Kim, H.; Kumar, K.S.; Shin, K.H. Applicability of stable C and N isotope analysis in inferring the geographical origin and authentication of commercial fish (Mackerel, Yellow Croaker and Pollock). *Food Chem.* **2015**, *172*, 523–527. [\[CrossRef\]](#)
- Kelly, S.; Heaton, K.; Hoogewerff, J. Tracing the geographical origin of food: The application of multi-element and multi-isotope analysis. *Trends Food Sci. Technol.* **2005**, *16*, 555–567. [\[CrossRef\]](#)
- Gonzalvez, A.; Armenta, S.; De La Guardia, M. Trace-element composition and stable-isotope ratio for discrimination of foods with Protected Designation of Origin. *Trends Anal. Chem.* **2005**, *28*, 1295–1311. [\[CrossRef\]](#)
- Guo, B.L.; Wei, Y.M.; Pan, J.R.; Li, Y. Stable C and N isotope ratio analysis for regional geographical traceability of cattle in China. *Food Chem.* **2010**, *118*, 915–920. [\[CrossRef\]](#)
- Yanagi, Y.; Hirooka, H.; Oishi, K.; Choumei, Y.; Hata, H.; Arai, M.; Kitagawa, M.; Gotoh, T.; Inada, S.; Kumagai, H. Stable carbon and nitrogen isotope analysis as a tool for inferring beef cattle feeding systems in Japan. *Food Chem.* **2012**, *134*, 502–506. [\[CrossRef\]](#)
- Aursand, M.; Mabon, F.; Martin, G.J. Characterization of farmed and wild salmon (*Salmo salar*) by a combined use of compositional and isotopic analyses. *J. Am. Oil Chem. Soc.* **2000**, *77*, 659–666. [\[CrossRef\]](#)
- Dempson, J.B.; Power, M. Use of stable isotopes to distinguish farmed from wild Atlantic salmon, *Salmo salar*. *Ecol. Freshw. Fish* **2004**, *13*, 176–184. [\[CrossRef\]](#)
- Bell, J.G.; Preston, T.; Henderson, R.J.; Strachan, F.; Bron, J.E.; Cooper, K.; Morrison, D.J. Discrimination of wild and cultured European sea bass (*Dicentrarchus labrax*) using chemical and isotopic analyses. *J. Agric. Food Chem.* **2007**, *55*, 5934–5941. [\[CrossRef\]](#)
- Morrison, D.J.; Preston, T.; Bron, J.E.; Hemderson, R.J.; Cooper, K.; Strachan, F.; Bell, J.G. Authenticating production origin of gilthead sea bream (*Sparus aurata*) by chemical and isotopic fingerprinting. *Lipids* **2007**, *42*, 537–545. [\[CrossRef\]](#)

21. Serrano, R.; Blanes, M.A.; Orero, L. Stable isotope determination in wild and farmed gilthead sea bream (*Sparus aurata*) tissues from the western Mediterranean. *Chemosphere* **2007**, *69*, 1075–1080. [\[CrossRef\]](#)
22. Busetto, M.L.; Moretti, V.M.; Moreno-Rojas, J.M.; Caprino, F.; Giani, I.; Malandra, R.; Bellagamba, F.; Guillou, C. Authentication of farmed and wild turbot (*Psetta maxima*) by fatty acid and isotopic analyses combined with chemometrics. *J. Agric. Food Chem.* **2008**, *56*, 2742–2750. [\[CrossRef\]](#)
23. Moreno-Rojas, J.M.; Tulli, F.; Messina, M.; Tibaldi, E.; Guillou, C. Stable isotope ratio analysis as a tool to discriminate between rainbow trout (*O. mykiss*) fed diets based on plant or fish-meal proteins. *Rapid Commun. Mass Spec.* **2008**, *22*, 3706–3710. [\[CrossRef\]](#)
24. Thomas, F.; Jamin, E.; Wietzerbin, K.; Guérin, R.; Lees, M.; Morvan, E.; Billault, I.; Derrien, S.; Moreno-Rojas, J.M.; Serra, F.; et al. Determination of origin of Atlantic salmon (*Salmo salar*): The use of multiprobe and multielement isotopic analyses in combination with fatty acid composition to assess wild or farmed origin. *J. Agric. Food Chem.* **2008**, *56*, 989–997. [\[CrossRef\]](#)
25. Fasolato, L.; Novelli, E.; Salmaso, L.; Corain, L.; Camin, F.; Perini, M.; Balzan, S. Application of nonparametric multivariate analyses to the authentication of wild and farmed European sea bass (*Dicentrarchus labrax*). Results of a survey on fish sampled in the retail trade. *J. Agric. Food Chem.* **2010**, *58*, 10979–10988. [\[CrossRef\]](#)
26. Moretti, V.M.; Turchini, G.M.; Bellagamba, F.; Caprino, F. Traceability issues in fishery and aquaculture products. *Vet. Res. Commun.* **2003**, *27*, 497–505. [\[CrossRef\]](#)
27. Turchini, G.M.; Quinn, G.P.; Jones, P.L.; Palmeri, G.; Gooley, G. Traceability and discrimination among differently farmed fish: A case study on Australian Murray cod. *J. Agric. Food Chem.* **2009**, *57*, 274–281. [\[CrossRef\]](#)
28. Carter, J.; Tinggi, U.; Yang, X.; Fry, B. Stable isotope and trace metal compositions of Australian prawns as a guide to authenticity and wholesomeness. *Food Chem.* **2015**, *170*, 241–248. [\[CrossRef\]](#)
29. Dinca, O.R.; Ionete, R.E.; Popescu, R.; Costinel, D.; Radu, G.L. Geographical and botanical origin discrimination of Romanian honey using complex stable isotope data and chemometrics. *Food Anal. Methods* **2015**, *8*, 401–412. [\[CrossRef\]](#)
30. Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 2nd ed.; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2019.
31. Burkov, A. *The Hundred-Page Machine Learning Book*; Andriy Burkov: Quebec City, QC, Canada, 2019.
32. Murphy, K.P. *Probabilistic Machine Learning: An Introduction*; The MIT Press: Cambridge, MA, USA, 2022.
33. Gelman, A. *Bayesian Data Analysis*, 3rd ed.; CRC Press: Boca Raton, FL, USA, 2014.
34. McElreath, R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, 2nd ed.; Taylor and Francis: Abingdon, UK; CRC Press: Boca Raton, FL, USA, 2020.
35. Martin, O. *Bayesian Modeling and Computation in Python*; CRC Press: Boca Raton, FL, USA, 2022.
36. Salvatier, J.; Wiecki, T.V.; Fonnesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Comput. Sci.* **2016**, *2*, 55. [\[CrossRef\]](#)
37. Carpenter, B.; Gelman, A.; Hoffman, M.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M.; Guo, J.; Li, P.; Riddell, A. Stan: A probabilistic programming language. *J. Stat. Softw.* **2017**, *76*, 1. [\[CrossRef\]](#)
38. Homan, M.D.; Gelman, A. The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **2014**, *15*, 1593–1623.
39. James, G.M.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning: With Applications in R*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2021.
40. Watanabe, S. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **2010**, *11*, 3571–3594.
41. Bruce, P.; Bruce, A.; Gedeck, P. *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*, 2nd ed.; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2020.
42. Yokoyama, H.; Tamaki, A.; Harada, K.; Shimoda, K.; Koyama, K.; Ishihi, Y. Variability of diet-tissue isotopic fractionation in estuarine macrobenthos. *Mar. Ecol. Prog. Ser.* **2005**, *296*, 115–128. [\[CrossRef\]](#)
43. Arcagni, M.; Campbell, L.M.; Arribure, M.A.; Kyser, K.; Klassen, K.; Casaux, R.; Miserendino, M.L.; Guevara, S.R. Food web structure in a double-basin ultra-oligotrophic lake in Northwest Patagonia, Argentina, using carbon and nitrogen stable isotopes. *Limnologia* **2013**, *43*, 131–142. [\[CrossRef\]](#)
44. Suh, Y.J.; Shin, K.H. Size-related and seasonal diet of the manila clam (*Ruditapes philippinarum*), as determined using dual stable isotopes. *Estuar. Coast. Shelf Sci.* **2013**, *135*, 94–105.
45. Sayle, K.L.; Cook, G.T.; Ascough, P.L.; Hastie, H.R.; Einarsson, T.; McGovern, T.H.; Hicks, M.T.; Edwald, A.; Friðriksson, A. Application of ³⁴S analysis for elucidating terrestrial, marine and freshwater ecosystems: Evidence of animal movement/husbandry practices in an early Viking community around Lake Mývatn, Iceland. *Geochim. Cosmochim. Acta* **2013**, *120*, 531–544. [\[CrossRef\]](#)
46. Koutsodendris, A.; Brauer, A.; Zacharias, I.; Putyrskaya, V.; Klemm, E.; Sangiorgi, F.; Pross, J. Ecosystem response to human- and climate-induced environmental stress on an anoxic coastal lagoon (Etoliko, Greece) since 1930 AD. *J. Paleolimnol.* **2015**, *53*, 255–270. [\[CrossRef\]](#)
47. Leonardos, I.; Sinis, A. Fish mass mortality in the Etolikon lagoon, Greece: The role of local geology. *Cybiuim* **1997**, *21*, 201–206.
48. Winner, W.E.; Smith, C.L.; Koch, G.W.; Mooney, H.A.; Bewley, J.D.; Krouse, H.R. Rates of emission of H₂S from plants and patterns of stable Sulphur isotope fractionation. *Nature* **1981**, *289*, 672–673. [\[CrossRef\]](#)