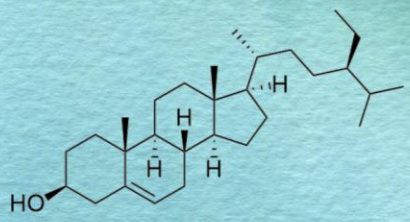
UNIVERSITAT DE BARCELONA

Development of innovative analytical strategies for the geographical and varietal authentication of food products highly susceptible to fraud

Berta Torres Cobos

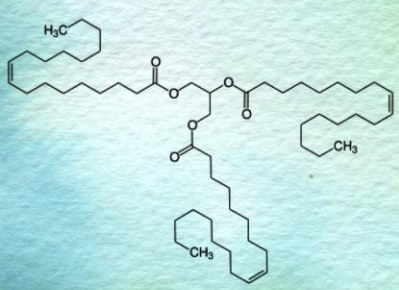
UNIVERSITAT DE BARCELONA

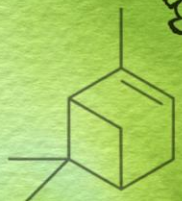
DOCTORAL THESIS
2024



Development of innovative analytical strategies for the geographical and varietal authentication of food products highly susceptible to fraud

Berta Torres Cobos





DEPARTAMENT DE NUTRICIÓ, CIÈNCIES DE L'ALIMENTACIÓ I GASTRONOMIA
FACULTAT DE FARMÀCIA I CIÈNCIES DE L'ALIMENTACIÓ

UNIVERSITAT DE BARCELONA

DEVELOPMENT OF INNOVATIVE ANALYTICAL STRATEGIES FOR THE GEOGRAPHICAL AND VARIETAL AUTHENTICATION OF FOOD PRODUCTS HIGHLY SUSCEPTIBLE TO FRAUD

BERTA TORRES COBOS

2024

DEPARTAMENT DE NUTRICIÓ, CIÈNCIES DE L'ALIMENTACIÓ I GASTRONOMIA
FACULTAT DE FARMÀCIA I CIÈNCIES DE L'ALIMENTACIÓ

UNIVERSITAT DE BARCELONA

PROGRAMA DE DOCTORAT EN ALIMENTACIÓ I NUTRICIÓ

DEVELOPMENT OF INNOVATIVE ANALYTICAL STRATEGIES FOR THE GEOGRAPHICAL AND VARIETAL AUTHENTICATION OF FOOD PRODUCTS HIGHLY SUSCEPTIBLE TO FRAUD

Memòria presentada per **Berta Torres Cobos** per optar al títol de doctora per la Universitat de Barcelona

Dra. Stefania Vichi
Directora i Tutora

Dra. Alba Tres Oliver
Directora

Berta Torres Cobos

BERTA TORRES COBOS
2024

FUNDING SOURCES AND INSTITUTIONAL SUPPORT

Ayuda del Programa de Formación de Profesorado Universitario (FPU20/01454) del Ministerio de Ciencia, Innovación y Universidades.

Proyecto TRACENUTS Desarrollo de herramientas de detección de fraudes en frutos secos españoles con alto riesgo de falsificación (PID2020-117701RB-I00) financiado por MCIN/AEI/https://doi.org/10.13039/501100011033.

Ayudas a la movilidad para estancias breves y traslados temporales para beneficiarios del programa FPU (EST23/00155) del Ministerio de Ciencia, Innovación y Universidades.

OLEUM-Advanced solutions for assuring the overall authenticity and quality of olive oil. This Project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 635690.

AUTENFOOD-Autenticació, traçabilitat i seguretat alimentària. Projecte finançat per la Generalitat de Catalunya a través del Fons Europeu de Desenvolupament regional (2014-2020), emmarcat dins la comunitat INNOÀPAT acreditada per ACCIÓ dins el programa RIS3CAT

Borses de viatge i ajuts de la Facultat de Farmàcia i Ciències de l'Alimentació de la Universitat de Barcelona.

Borses de viatge i d'ajut INSA-ESTADA 2023 de l'Institut de Recerca en Nutrició i Seguretat Alimentària de la Universitat de Barcelona (INSA-UB) reconegut com a Unitat d'Excel·lència Maria de Maeztu (subvenció CEX2021-001234-M) finançada pel MICIU/AEI/FEDER, UE.

Campus de l'Alimentació de Torribera de la Universitat de Barcelona.

Centre wallon de Recherches agronomiques, Département Valorisation des productions.

Institut de Recerca i Tecnologia Agroalimentàries.

Centres Científics i Tecnològics de la Universitat de Barcelona.

Grup de recerca en Lípids i compostos bioactius de la cadena alimentària (LIBIFOOD) 2021-SGR-00854.

Grup de recerca en isótopos estables, mineralogía y geoquímica (MAiMA) 2021-SGR-00308.

"To the rational mind, nothing is inexplicable,

only unexplained."

ACKNOWLEDGEMENTS

Muchas gracias a todos los que me habéis acompañado en esta aventura. No hubiera podido coronar esta cima sin vosotros.

Habéis sido sombra y un soplo de aire fresco, cuando el calor de la frustración y el cansancio me tentaban a abandonar el camino.

Habéis sido estrellas en la noche sin luna, cuando la oscuridad y la duda no me permitían ver, iluminando mi sendero.

Gracias a vosotros he disfrutado de cada paso del viaje. Un viaje inolvidable, que me ha cambiado, que me ha dado grandes alegrías, muy buenos momentos para recordar y la mejor de las compañías.

Antes que todo quiero agradecer a las dos lideres de esta expedición, las capitanas de este navío Stefania y Alba, expertas en sortear todas las tormentas, pasando por los marcadores fantasma, los tornados de muestreo, el abismo del masas estropeado, los tsunamis de la burocracia y la eterna rivalidad con el dinosaurio univariante.

Muchas gracias por todo lo que me habéis enseñado, por tener siempre tanta paciencia, por estar siempre dispuestas, por vuestra chispa y humor, por dejarme explorar, pero jamás dejarme sin guía, por todos esos días, tardes, noches juntas (si si más de una noche cerrando Gaudí), por hacerme sentir una más y acompañada, por hacer de LiBiFOOD una segunda casa a la que volver.

No me olvido, del cartógrafo jefe de esta expedición, Francesc, te sabes todas las rutas, más fiable que cualquier brújula, yo aun no concibo una pregunta que no puedas responder, y haga viento o marea, siempre estas cuando te necesitamos. Gracias por ser tan redondo.

Muchas gracias a todo el LiBi-Team. Gracias Bea por ser una compañera maravillosa, por todo lo que me has transmitido, por orientarme y por brindarme tu apoyo y ayuda. Gracias Soriana, por todos los momentos compartidos en el 36, por siempre tener una sonrisa y estar dispuesta a embarcarte, por tu resiliencia y ganas, contigo de gusto hasta hacer insaponificables. Gracias Laura, por esa energía y buena vibra que llevas

siempre contigo, por estar tan motivada y movernos a todos a nuevos planes, tu alegría se contagia. Gracias Paula, por esos momentos de poyata compartidos, por tus consejos y tu sinceridad sin tapujos. Gracias Anna por todas las buenas vivencias compartidas dentro y fuera del laboratorio, contigo siempre hay algo nuevo que descubrir. Gracias Elisa por ser el alma de la fiesta, por hacernos reír siempre y por contar con tu apoyo, contigo las penas se van cantando. También quiero agradecer a todos los estudiantes que han contribuido al desarrollo de esta tesis. Gracias, Laura, Anna, Raúl, Gerard, Marco, Antonio, Carlo y Celia.

Gracias a las chicas del 36, Judit, Sonia, Irache, y a Oriol (que te tienen desplazado, pero sabemos que tu "cuore" está en el 36), por hacer los días más amenos, por las risas, los desayunos, la compañía, el apoyo, las saliditas, las rutillas (tenemos que hacer más) las conversaciones sin fin, y la calidez, sé que siempre puedo contar con vosotros.

Gracias a todos los Torriberos, por el buen ambiente que generáis, por las comidas juntos, las "calçotadas", las casas rurales y por siempre tener un hombro amigo al que arrimarse.

Gracias a todas las personas que forman parte del Departamento de Nutrición Ciencias de la Alimentación y Gastronomía y del Campus de la Alimentación Torribera.

Gracias a todo el equipo de MAiMA, y en especial a Mònica y Albert por toda su ayuda, conocimiento y paciencia y por descubrirme el fascinante mundo de los isótopos.

Gracias a CCiT-UB y LIRA-PANTHALASSA por su colaboración y apoyo técnico en este proyecto.

Gracias Juan Antonio y Vincent, por vuestra cálida acogida en Bélgica, por hacer de mi estancia un agradable paseo en buena compañía, por darme libertad para probar de todo y orientarme durante todo el proceso, por descubrirme una nueva Bélgica con aires españoles. Gracias a todo el equipo del CRA-W, por permitirme aprender de todos vosotros e integrarme como una más en el equipo.

Gracias a todas las amistades que me habéis apoyado en los buenos y malos momentos, que habéis estado ahí aguantando conmigo el vendaval, hasta ver salir el

sol. En especial gracias a Adri, al que le gustan las hamburguesas y al que le debo un jamón vegano.

Gracias a mi familia, sabéis que sois mi pilar, y no podría haber llegado tan lejos sin vosotros. Gracias Mamá por siempre confiar en mí, incluso más que yo misma, por motivarme y ayudarme a seguir adelante, por todo lo que haces, has hecho y harás.

Gracias Papa, por estar al pie del cañón, por preocuparte y vivirlo conmigo, aunque lo lleves por dentro, por apoyarme en mis decisiones y entenderme, por estar siempre a mi lado de manera incondicional. No podría haberlo hecho sin ti.

Gracias Ferran, por ser como eres, por quererme como me quieres. Por estar ahí siempre, por tu apoyo, por tu comprensión, por tu paciencia, por tu empatía, por tu fortaleza, por tu compañía, por tu cariño, por ti.

ABSTRACT

Due to globalization and the expansion of food supply chains, food fraud has become a widespread practice with global impacts. High-value commercial foods, such as virgin olive oil and nuts, are targets of these fraudulent practices, driven by economic gain. These products are highly valued for their nutritional benefits and sensory properties, which are strongly influenced by their varietal and geographical origin. This makes them particularly susceptible to mislabelling to exploit the premium prices associated with certain regions or cultivars. This situation is exacerbated by the lack or inefficiency of official control methods. In addition to the economic harm to consumers and industry, food fraud also presents a high risk to public health.

Therefore, developing efficient analytical methodologies to verify the geographical and varietal origin of these products is essential for preventing fraud and protecting consumer rights. The main objective of this doctoral thesis is to ensure the varietal and geographical authenticity of high-value food products, such as virgin olive oil, hazelnuts, and pine nuts, through the development of reliable analytical and chemometric techniques based on appropriate markers.

To achieve this goal, methods based on isotopic and metabolic markers have been implemented, using both targeted and untargeted approaches for the three studied foods. Their advantages and limitations have been evaluated to identify the most suitable methods. For virgin olive oil, two methodologies were compared to verify geographical origin by developing classification models: fingerprinting of sesquiterpene hydrocarbons and stable isotope analysis of carbon (C), hydrogen (H), and oxygen (O). Results showed that sesquiterpene fingerprinting is more effective, providing greater sensitivity, specificity, and robustness. This method was also successfully applied to varietal authentication of olive oil, validating it as an effective tool for both geographical and varietal authentication.

For hazelnuts, a method to analyse the isotopic ratio of bulk O, and C and H of fatty acid methyl esters was developed to authenticate geographical origin. Additionally, three untargeted metabolic methods (triacylglycerols, unsaponifiable fraction, and

spectroscopic fingerprinting) were applied to simultaneously verify the cultivar and origin. Although the isotopic method was able to distinguish between regions with different climatic and geological conditions, it was outperformed by the metabolic methods. Triacylglycerol fingerprinting, while offering lower accuracy, was fast and simple, making it a recommended method for initial screening. Conversely, fingerprinting of the unsaponifiable fraction was more laborious but offered higher accuracy as a confirmatory method. Combining both methods enhanced efficiency reducing workload and maintained high accuracy in classification. Lastly, the spectroscopic method, being faster, simpler and more accurate than triacylglycerol analysis, was also a suitable screening method.

Regarding pine nuts, strontium isotopic analysis failed in differentiating Mediterranean pine nuts from Asian ones, highlighting that a single isotopic marker is insufficient to address the complexity of geographical authentication of agri-food products. However, terpene fingerprinting allowed the construction of robust models capable of effectively differentiating Mediterranean pine nuts from Asian ones, proving to be a direct, fast, and efficient method for their authentication.

Overall, metabolic fingerprinting methods have shown superior effectiveness for the geographical and varietal authentication of the three studied foods, positioning them as suitable support tools to official control methods. However, limitations related to their transferability and the lack of standardized validation protocols are their main current challenges.

RESUMEN

Debido a la globalización y la expansión de las cadenas de suministro de alimentos, el fraude alimentario se ha convertido en una práctica extendida, con impactos globales. Alimentos de alto valor comercial, como el aceite de oliva y los frutos secos, son objetivos claros de estas prácticas fraudulentas, motivadas por el beneficio económico. Estos productos son muy apreciados por sus beneficios nutricionales y propiedades sensoriales, los cuales están fuertemente influenciados por su origen geográfico y variedad. Esto los hace especialmente susceptibles a falsificaciones en su etiquetado para aprovechar los precios premium asociados a determinadas regiones o variedades. Esta situación se ve agravada por la falta o ineficiencia de los métodos de control oficiales. Además del perjuicio económico para el consumidor y la industria, el fraude alimentario también representa un riesgo significativo para la salud pública.

Por ello, es necesario desarrollar nuevas metodologías analíticas que permitan verificar el origen geográfico y varietal de estos productos, previniendo así el fraude y protegiendo los derechos de los consumidores. El objetivo principal de esta tesis doctoral es garantizar la autenticidad varietal y geográfica de productos alimentarios de alto valor, como el aceite de oliva virgen, las avellanas y los piñones, mediante el desarrollo de técnicas analíticas y quimiométricas robustas y fiables basadas en marcadores adecuados.

Para alcanzar este objetivo, se han implementado métodos basados en marcadores isotópicos y metabólicos, utilizando enfoques dirigido (*targeted*) y no dirigido (*untargeted*) en los tres alimentos estudiados. Se han evaluado sus ventajas y limitaciones para identificar los métodos más adecuados.

En el caso del aceite de oliva virgen, se compararon dos metodologías para verificar el origen geográfico mediante el desarrollo de modelos de clasificación: el *fingerprinting* de hidrocarburos sesquiterpénicos y el análisis de isótopos estables de carbono (C), hidrógeno (H) y oxígeno (O). Los resultados demostraron que el *fingerprinting* de sesquiterpenos es más efectivo, proporcionando mayor sensibilidad, especificidad y robustez. Este método también se aplicó con éxito a la autenticación varietal del aceite de oliva, validándose como una herramienta eficaz para su autenticación tanto geográfica como varietal.

Para las avellanas, se desarrolló un método para analizar el ratio isotópico de O, y de C y H de los ácidos grasos para autenticar el origen geográfico, así como tres métodos metabólicos *untargeted* (*fingerprinting* de triacilglicéridos, fracción insaponificable y espectroscópico) para verificar simultáneamente la variedad y procedencia. Aunque el método isotópico permitió distinguir entre regiones con diferentes condiciones climatológicas y geológicas, fue superado por los métodos metabólicos. El *fingerprinting* de triacilglicéridos, aunque menos eficiente, por su rapidez y sencillez resultó adecuado para el cribado masivo de muestras. Por otro lado, aunque el *fingerprinting* de la fracción insaponificable era más laborioso, ofreció una gran eficiencia como método de confirmación. La combinación de ambos métodos mejoró la eficiencia y mantuvo una alta exactitud en la clasificación. Por último, el método espectroscópico, al ser más rápido, sencillo y eficiente que el *fingerprinting* de triacilglicéridos también resultó ser un método de cribado adecuado.

En cuanto a los piñones, el análisis isotópico de estroncio no logró diferenciar los piñones mediterráneos de los asiáticos, evidenciando que un solo marcador isotópico no es suficiente para abordar la complejidad de la autenticación geográfica de productos agroalimentarios. Sin embargo, el *fingerprinting* de terpenos permitió construir modelos robustos capaces de diferenciar eficazmente los piñones mediterráneos de los asiáticos, demostrando ser un método directo, rápido y eficiente para su autenticación.

En general, los métodos de *fingerprinting* metabólico han demostrado ser más eficaces para la autenticación geográfica y varietal de los tres alimentos, posicionándose como herramientas de apoyo adecuadas a los métodos de control oficiales. Sin embargo, las limitaciones relacionadas con su transferibilidad y la falta de protocolos de validación estandarizados son sus principales desafíos actuales.

LIST OF ABBREVIATIONS

AAC: Administrative Assistance and Cooperation Network

ACN: Alert and Cooperation Network

ANOVA: One-way analysis of variance

CAL: Calabria

CAT: Catalonia

CHL: Chile

CHN: China

COW: Correlation optimized warping

CRISPR: Clustered regularly interspaced short palindromic repeats

CRM: Certified reference materials

CS: Central Spain

CSIA: Compound-specific isotopic analysis

DI: Diagnostic index

DNA: Deoxyribonucleic acid

EA: Elemental analysis

EC: European Commission

EFSA: European Food Safety Authority

EIC: Extracted ion chromatogram

ESP: Spain

EU: European Union

EVOO: Extra virgin olive oil

FA: Fatty acid

FAME: Fatty acid methyl ester

FIA: Flow injection analysis

FFN: EU Agri-Food Fraud Network

GEO: Georgia

GFSI: Global Food Safety Initiative

GC: Gas chromatography

GI: Geographical indication

GRC: Greece

HESI: Heated electrospray

HL: High linoleic

hNIR: Handheld near-infrared

HPLC: High-performance liquid chromatography

HO: High oleic

HRMS: High resolution mass spectrometry

HS: Headspace

HT: High-temperature

IMSOC: Information Management System for Official Controls

ICP: Inductively coupled plasma

IPR: Intellectual property rights

IRMS : Isotope ratio mass spectrometry

ITA: Italy

JRC: Joint Research Centre

LC: Liquid chromatography

Ln: Linoleic

LV: Latent variables

MALDI: Matrix-assisted laser desorption/ionization

MC-ICPMS: Multi-collector inductively coupled plasma mass spectrometry

MIR: Mid-infrared

MS: Mass spectrometry

MS/MS: Tandem mass spectrometry

NGS: Next-generation sequencing

NIR: Near-infrared

NMR: Nuclear magnetic resonance

O: Oleic

OO: Olive oil

P: Palmitic

PARADISe: Parallel factor analysis 2 deconvolution and identification system

PARAFAC2: Parallel factor analysis 2

PC: Principal component

PCA: Principal component analysis

PCR: Polymerase chain reaction

PDO: Protected designation of origin

PGI: Protected geographical indications

PLS-DA: Partial least squares – discriminant analysis

PMS: Pine mouth syndrome

PoC: Proof-of-concept

POR: Portugal

PUG: Apulia

PV: Predicted value

RAPD: Randomly amplified polymorphic DNA

RASFF: Rapid Alert System for Food and Feed

RMSEcv: Root mean squared error of cross-validation

ROC: Receiver operating characteristics

RUS: Russia

S: Stearic

SEcv: Standard error of cross-validation

SH: Sesquiterpene hydrocarbons

SI: Sicily

SIMCA: Soft independent modelling of class analogy

SNV: Standard normal variate

SPME: Solid-phase microextraction

SSR: Simple sequence repeat

TAG: Triacylglycerol

TG: Tonda di Giffoni

TIMS: Thermal ionization mass spectrometry

TOF: Time-of-flight

TUN: Tunisia

TUR: Turkey

UF: Unsaponifiable fraction

UK: United Kingdom

UPIFA: Unsaturated polymethylene-interrupted fatty acids

VOO: Virgin olive oil

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1.
INTRODUCTION

1.1 Food fraud: evolution, challenges and monitoring actions

1.1.1 Background and definitions

Since ancient times in all cultures and regions, from the antique Babylonian civilization to the Roman Empire widespread trading market, extending into the Middle Ages to the present days, food fraud has been a common practice. Babylonian and Roman laws contained edicts to regulate beer and wine purity, punishing fraudulent activities such as food contamination or adulteration. However, during those periods, the methods available for identifying food tampering were limited and mainly based on organoleptic tests, which propitiated the spreading of the practice [1].

In the recent years, food fraud has come once again under the spotlight due to several notorious cases that resulted in sever health consequences. These include the toxic oil syndrome in Spain (1981), caused by the adulteration of pomace oil with industrial rapeseed oil and sold as olive oil (OO), which resulted in 2,500 deaths, 20,000 victims [2], and enduring health issues for survivors; the adulteration of milk and infant formulas with melamine in China (2008), leading to adverse renal effects on over 294,000 infants, 50,000 hospitalizations, and at least six deaths; and countless methanol poisoning episodes from adulterated alcoholic beverages, such as in Norway (2002-2004), with 51 hospitalizations and 9 deaths, and in the Czech Republic (2012), where 121 individuals were affected, 41 died, and survivors experienced visual or central nervous system sequelae. Other significant cases include the adulteration of alcoholic beverages with methanol in Italy (1986) and the United Kingdom (UK) (2011); the ‘Horse meat scandal’ in UK (2013); the ‘Operation weak flesh’ in Brazil, a corruption that facilitated the distribution of spoiled meat into the food supply chain (2017); and the Belgian meat scandal which involved the falsification of expiration dates and the mislabelling of conventional meat as 'organic’, and diverting meat meant for animal feed into the human food supply chain [3,4].

Due to globalization and expansion of modern food supply chains and manufacturing infrastructures, food fraud has evolved, scaling its impact to global population. Consequently, specific definitions for food fraud have been developed to properly identify and assess these issues [5]. The food concerns that can happen throughout the food supply chain, compromising its integrity, can be divided into four categories

(**Figure 1**). This classification is known as the **food protection risk matrix** and was firstly presented by Spink & Moyer [5].

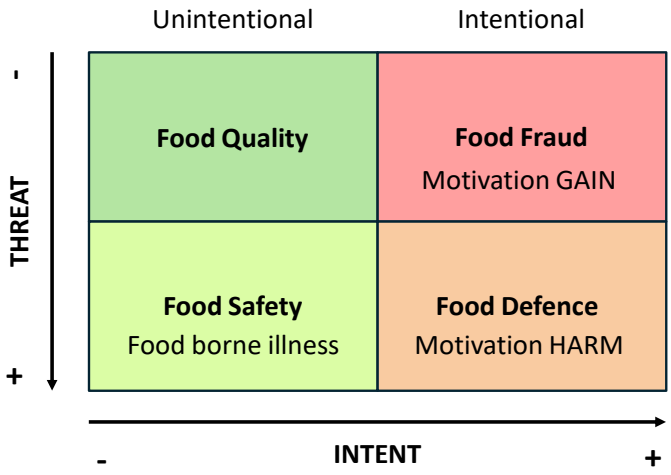


Figure 1. The food protection risk matrix adapted from Spink & Moyer [5].

The main difference between the categories is the intentionality. Food fraud and issues considered by food defence strategies are intentional actions and are therefore regarded as criminal activities. In contrast, food quality and safety issues are typically unintentional or accidental. While subjects addressed by food defence aim to harm or injure consumers, the motivation of food fraud is achieving an economic gain. However, food fraud can also lead to significant food safety issues that endanger public health, as demonstrated by the catastrophic consequences of previously mentioned fraud incidents, which highlight that food fraud poses a significant food safety risk. This is because the substances used for adulteration are often unusual, and existing food protection systems are not yet equipped to detect them. This was the case of the adulteration of infant formulas and milk with melamine, an unexpected adulterant typically used in manufacturing plastic products, which bypassed standard quality tests for protein content [5]. Additionally, new adulterants can introduce allergenic substances, such as nuts, sesame, and soy, among others [6,7]. Therefore, since any type of food fraud inherently leads to the loss of traceability, food safety cannot be ensured, making the risk of consumer harm constant.

According to the Global Food Safety Initiative (GFSI), food fraud is “A collective term encompassing the deliberate and intentional substitution, addition, tampering or

misrepresentation of food, food ingredients, feed, food packaging or labelling, product information or false or misleading statements made about a product for economic gain that could impact consumer health" [8,9].

Therefore, a food concern or issue is considered a food fraud problem when these four criteria are meet: 1) **intention**, 2) **violation of the legislation**, 3) **consumer deception** and 4) **economic gain**. If it does not fulfil one of the characteristics, it cannot be typified as a food fraud criminal activity.

Regarding food fraud as a criminal act committed by fraudsters to achieve a personal gain, criminology and behavioural sciences can offer deeper insights into the issue and promote a more proactive preventive action. **Situational Crime Prevention** is a branch of criminology focused on developing measures that imply a systematic and permanent management, design, or manipulation of the immediate environment to reduce the opportunity of crime. These measures hamper criminal actions by increasing the risk of getting caught and reducing the potential rewards for offenders. The criminology approach of **Situational Crime Prevention** is a suitable way to address food fraud, since its main goal is to prevent fraudulent actions from occurring rather than arresting the criminal once the crime has been committed. Consequently, tools focused on reducing the crime opportunity are more effective in this scenario [10]. On theses bases, a fundamental criminology concept, **the crime triangle**, is also applicable to the field of food fraud (**Figure 2**).

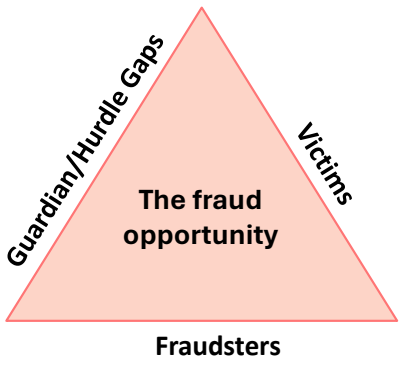


Figure 2. The crime triangle adapted from Spink & Moyer [5].

This concept is used to describe the three elements that generate the crime opportunity, or in this specific case, the opportunity for food fraud. Each side of the triangle represents one of these elements: **victims**, **fraudsters** and **guardian/hurdle**

gaps; and the area of the geometric figure reflects the overall opportunity for fraud. Hence, if one or more sides of the triangle increase, the total area expands, thereby boosting the opportunity for fraud [5].

The number of fraudsters is limitless and tends to rise with the potential economic gain. Similarly, the more popular and consumed a product is, the greater the number of potential victims to scam, particularly those who cannot easily distinguish between genuine and non-authentic products. Additionally, gaps or inefficiency in the systems designed to protect consumers and ensure product authenticity, further facilitate the crime. Entirely removing any element of the triangle is not feasible. However, reducing fraud is possible by increasing the likelihood of detection through efficient control tools or by raising the costs associated with committing the fraud.

Having delineated the problem, the concepts of "food authenticity" and "food integrity" should also be addressed and properly defined. According to the Codex Alimentarius (CX/FICS 18/24/7) [11], **Food authenticity** is "the quality of a food to be genuine and undisputed in its nature, origin, identity, and claims, and to meet expected properties", while **Food integrity** is "the status of a food product where it is authentic and not altered or modified with respect to expected characteristics including, safety, quality, and nutrition."

As previously mentioned by the GFSI, food fraud is a collective term that encompasses multiple deceptive practices. The main types of food fraud, as identified by the GFSI Food Fraud Think Tank initiative and further defined in the FoodIntegrity Handbook from the FoodIntegrity European Union (EU) project, are summarized in **Table 1**. Among these, a specific subgroup defined as adulteration includes the intentional addition of a foreign or inferior substance, typically to prepare a product for sale by replacing more valuable ingredients with less valuable or inert ones [3]. Conversely, other types of fraud do not involve altering the product's composition but focus on falsifying claims or information on the label for economic gain, such as **mislabelling** and counterfeiting, or exploiting unauthorized sales channels, known as the grey market. According to the FoodIntegrity Handbook definitions, the main difference between mislabelling and counterfeiting is that mislabelling refers to any incorrect claims on the

label that could influence consumer choices, while counterfeiting specifically involves a violation of intellectual property rights.

Table 1. Food fraud types and definitions. Adapted from GFSI *Food fraud technical document, tackling food fraud through food safety management systems* [8]

Food fraud type	Definition
Adulteration	
Dilution	Process of mixing a liquid ingredient of high value with a liquid of lower value.
Substitution	Process of replacing an ingredient or part of the product of high value with another ingredient or part of the product of lower value.
Concealment	Process of hiding the low quality of a food ingredients or product.
Unapproved enhancements	Process of adding unknown and undeclared materials to food products to enhance their quality attributes.
Mislabelling	Process of placing false claims on packaging for economic gain.
Grey market production	Production, theft, and diversion involving unauthorised sales channels for products.
Counterfeiting (IPR)	Process of copying the brand name, packaging concept, recipe, processing method etc. of food products for economic gain.

Abbreviations: IPR, intellectual property rights.

1.1.2 Varietal and geographical origin mislabelling

Consumer purchasing behaviour is strongly influenced by food labels, as they are the primary source of information about a product. Labels not only help consumers identify products but also offer key details, such as geographical or varietal origin, which enhance perceived quality and encourage repeat purchases [12,13].

Information about the varietal and geographical origins of food is particularly valuable to consumers, as the characteristics of agrifood products are significantly shaped by environmental and genetic factors. Many consumers associate specific regions or cultivars with high-quality products, reinforcing their purchasing decisions [13].

Products such as Parmesan cheese from Italy, Champagne from France, or ham from Spain are often preferred for their cultural significance and connection to their place of origin, as consumers seek authentic flavours and traditional practices [14]. Similarly, monovarietal virgin olive oils (VOOs) that specify the olive cultivar on the label, along with VOOs that exhibit a geographical indication (GI), which mandates the use of specific traditional olive cultivars from designated regions, increase consumers' perception of quality and result in higher prices [15,16].

Furthermore, consumers have become increasingly aware of environmental and ethical concerns within the food supply chain, leading them to favour products that are sustainably or ethically produced, often tied to specific regions or cultivars [13,17,18].

As a consequence of this consumer preference, food products from specific regions or cultivars that include this information on their labels or fall under a protected denomination of origin (PDO) or protected geographical indication (PGI) tend to command higher prices in the market. This creates an opportunity for economically motivated fraud, where products are mislabelled to exploit the premium prices associated with certain regions or cultivars, or falsely claim to originate from a GI. Hence, high-quality, widely consumed products with significant price differences based on their varietal or geographical origin are particularly vulnerable to fraudulent practices. The risk is even higher for food products that non-expert consumers find difficult to differentiate between genuine and non-authentic. This vulnerability is exacerbated when the control systems designed to verify the declared provenance or cultivar/specie of agrifood products are insufficient, ineffective or even non-existent.

Mislabelling the varietal and geographical origin of foods has several negative impacts on consumers, companies, industries, and even national economies. It causes economic damage on legitimate producers by undercutting prices and reducing market share, while also eroding consumer trust in the food industry. Companies involved in such practices face serious legal consequences, including fines, product recalls, and reputational damage, with long-term repercussions for both the business and the broader industry.

Moreover, the consequences of this type of food fraud extend beyond the economic impact, posing significant risks to consumer health and potentially leading to serious public health hazards [19,20]. Fraudsters introduce food products into the market that

fail to meet safety standards, as they have evaded proper safety controls, leading to the risk of consumer exposure to contaminants, toxins, or allergens not present in genuine products [5,19,21]. Furthermore, because the food traceability chain is compromised, if a health hazard is detected at any point, identifying, detecting, and recalling the affected products becomes significantly more difficult, aggravating the issue.

Authenticating the varietal and geographical origin of food can be more challenging than detecting other fraudulent practices such as adulteration, which involves altering the composition of food by mixing, diluting, or replacing it with different components. Verifying food provenance or cultivar involves detecting subtle compositional differences and identifying specific markers that can distinguish between the products from different regions or cultivars. This process requires sophisticated techniques to detect these nuanced variations, making it inherently more complex than addressing simpler forms of fraud [22,23].

1.1.3 EU monitoring actions

In 2002, the European Commission (EC) developed the Regulation (EC) No. 178/2002, also known as the General Food Law Regulation, which lays down the general principles and requirements of food law and procedures in matters of food safety [24]. It sets the groundwork for the development of food and feed legislation, guiding the development of regulations at both the EU and national levels. This regulation resulted in the establishment of the European Food Safety Authority (EFSA), an independent agency responsible for scientific advice and support to the EC, the European Parliament and all the EU member states. Additionally, it establishes the legal bases of the **Rapid Alert System for Food and Feed (RASFF)**, a tool that enables the fast and constant flow of information between EU countries to ensure a prompt reaction to public health risks associated to food and feed [25,26].

As an outcome of the recent 'Information Management System for Official Controls (IMSOC) Regulation' (Regulation 2019/1715), RASFF has become part of the **Alert and Cooperation Network (ACN)** together with the **Administrative Assistance and Cooperation Network (AAC)** and the **EU Agri-Food Fraud Network (FFN)** [27]. The ACN promotes a forward exchange of administrative information and the cooperation between member states on official controls in the agri-food chain and is governed by

the Official Controls Regulation (Regulation (EU) 2017/625) [28]. Each of these networks addresses distinct aspects of the agri-food chain: the RASFF handles non-compliances that may pose health risks, the AAC deals with non-compliances that do not involve health risks, and the FFN focuses on suspected fraud [29] (**Figure 3**). For a non-compliance to be considered a potential fraud, it must meet the four criteria mentioned in the previous section 1.1.1. Food fraud notifications are confidential and can only be managed by the FFN. However, if they imply any health risk, cooperation with RASFF may be contemplated. In such cases, RASFF notification should not include details of the fraud investigation and should provide only the necessary information to mitigate the risk and enable appropriate actions [30].

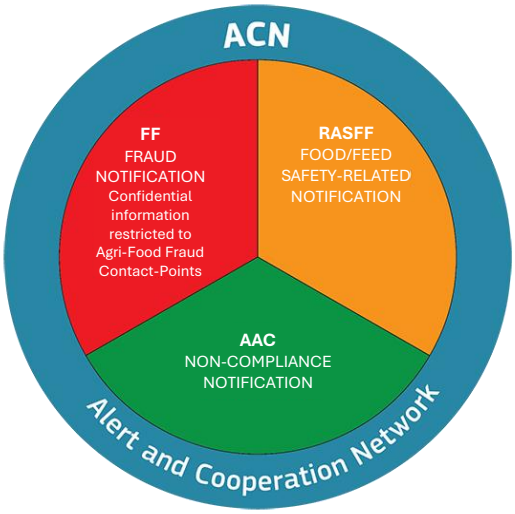


Figure 3. The ACN diagram [29].

Since 2021, information about cross-border non-compliances with EU legislation in the agri-food chain has been shared among ACN members via an online platform called iRASFF, managed by the EC. A report summarizing all the information shared in iRASFF and the ACN activities has been published annually.

The last ACN Annual Report from 2022 shows that the number of notifications associated to suspicions of food fraud affecting more than one country significantly increased compared to 2021. Of these notifications, 274 were associated to food products. Their classification by food product category is shown in **Figure 4** [31].

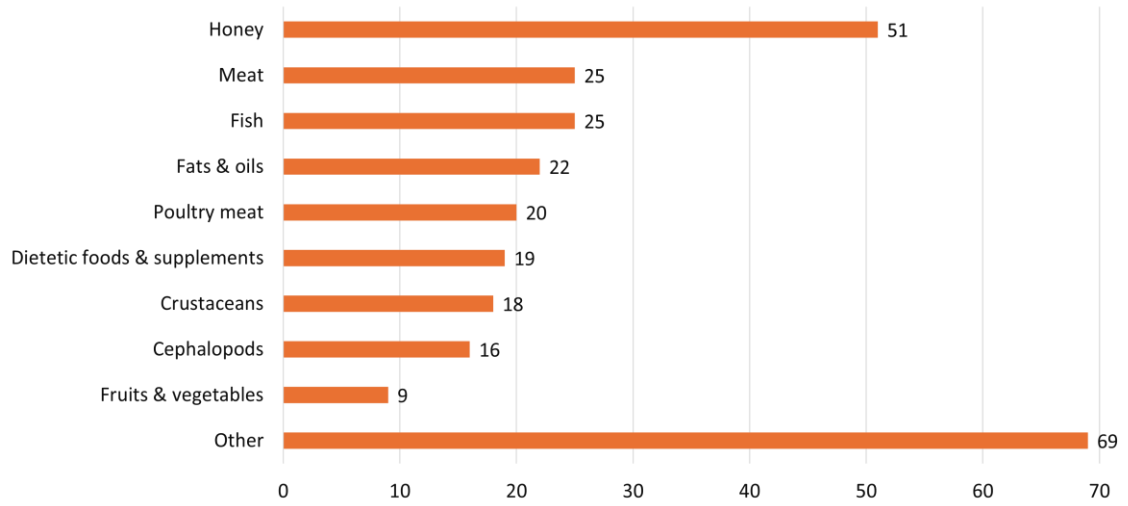


Figure 4. Food fraud notifications by food product category. Adapted from the 2022 ACN Annual Report [31].

The most notified food categories over the past five years were honey, meat, fish, and fats and oils [31–35]. Nonetheless, the ACN only reports cases where more than one country is involved; cases occurring within a single country are not included. Additionally, it should be noted that an increasing number of requests might not represent a growth in cases, but rather an improved reporting. Therefore, food categories that are not extensively checked or controlled, either due to a low number of inspections or a lack of effective authentication methods, are either underrepresented or fall into the "Other" category, yet they may still pose a significant fraud risk.

1.2 Virgin olive oil and nuts: highly susceptible to varietal and geographical mislabelling

1.2.1 Relevance of authenticating virgin olive oil and nuts

The Mediterranean diet is recognized worldwide for its nutritional benefits, and its traditional foods, renowned for their health-promoting properties, are increasing in popularity. As the demand for these products is increasing, the presence of items from numerous geographical and varietal sources is increasing in global markets. High-priced traditional commodities often have significant price differences due to their geographical and varietal origins, making them especially attractive targets for fraud.

Among the plant-based foods that are traditional pillars of the Mediterranean diet, lipid-rich foods such as **VOO** and **nuts** stand out due to their high commercial value and the strong connection between their characteristics, reputation, pricing, and geographical and botanical origins. The qualitative distinctions and market price differences between products from distinct provenances and cultivars make them highly vulnerable to economically motivated fraudulent practices, including the mislabelling of their botanical or geographical origins. The popularity of these high-priced food commodities, combined with the difficulty of consumers in accurately identifying authentic products, further aggravate the susceptibility to fraud. This issue is intensified by inadequate control methods, which enable fraud to yield significant economic gains while maintaining a low risk of detection.

The EU, and Spain in particular, are leading producers of VOO and nuts, significantly contributing to their presence in global markets. Consequently, there is a strong interest in valorising local production to distinguish it from emerging competitors and in ensuring the authenticity of these products.

On one hand, **OO,** particularly **VOO**, holds substantial economic importance due to its high market value, extensive culinary applications, and health benefits. The EU is the world's leading producer, accounting for almost 60% of global production, with Spain as the primary producer country, followed by Italy, Greece, Turkey and Tunisia [36]. VOO prices are influenced by labour and production costs at the farm level in each country, as well as by the demand for VOOs from specific countries or regions. For instance, the production costs of Italian VOOs are, on average, 30% higher than those of Spanish and non-European Mediterranean producers. This, combined with the high demand for Italian VOOs in both international and national markets, is reflected in their higher price [37].

In addition to geographical origin, the composition and qualitative characteristics of VOO are significantly influenced by the olive cultivar [38,39]. Over the centuries, olive cultivars for oil production have been selectively bred based on their adaptability to various climatic conditions and soil types, resistance to pests, productivity, oil yield and sensory characteristics [39,40,41]. Prominent VOO cultivars include ‘Picual’ and ‘Arbequina’ from Spain; ‘Frantoio’ and ‘Leccino’ from Italy; and ‘Koroneiki’ from

Greece. These cultivars produce VOOs with markedly different compositional and sensory characteristics, although these can vary depending on the region of production. For this reason, VOOs within the same commercial category and country of origin can display notably different compositional and sensory attributes depending on the specific cultivar used. Consequently, monovarietal VOOs that specify the olive cultivar can influence consumers' choice and enhance their perception of quality, potentially leading to higher prices [15,16]. Although these cultivars are predominantly grown in Mediterranean countries, their resilience and versatility have led to their cultivation in regions far from their traditional origins, such as Argentina, Chile, California, and Australia [42]. For instance, the 'Arbequina' cultivar is now cultivated worldwide due to its high and consistent productivity, substantial oil content and quality features, and adaptability to different environmental and agricultural conditions [42].

Hazelnuts are one of the top five most consumed nuts worldwide with a global production higher than one million tons per year [43]. They are highly valued for their sensory and nutritional qualities and serve as a key ingredient in a wide range of foodstuff and confectioneries, enriching the flavour and texture of both sweet and savoury products. Hazelnut sensory and nutritional qualities are heavily influenced by their cultivar and geographical origin [44,45], resulting in significant variations in the market price [43]. The main hazelnut producing countries are Turkey (71% of global production), Italy (7%), USA (5%), Azerbaijan (4%), Chile (4%) and Georgia (3%) [46]. These last three emerging producer countries have greatly increased their production areas over the past decade and this trend is expected to continue in the coming years [47-49]. The country of origin is a critical factor influencing hazelnut price in the market, generating notable fluctuations (e.g. the price for in-shell hazelnuts from Italy was 3416 USD/T whereas the ones from Georgia were 1287 USD/T) [43], which in turn incentivizes economically motivated mislabelling. Among the safety risks associated with hazelnut origin mislabelling is the potential for aflatoxin contamination, which is notably prevalent in nuts from certain regions, such as Georgia and Turkey [50].

Regarding hazelnuts botanical origin, each cultivar has unique technological traits, including nut yield, shape, size, and ease of pellicle removal, as well as distinct kernel

flavour and aroma. These characteristics play a crucial role in the industry's selection of cultivars for specific applications. Currently, round-shaped Italian cultivars such as 'Tonda Gentile', 'Tonda Romana', and 'Tonda di Giffoni' (TG), Spanish cultivars such as 'Negret' and 'Pauetet', and Turkish cultivars such as 'Tombul', 'Palaz', and 'Fosa', dominate the market [51]. Although these cultivars are primarily grown in their traditional regions, some, like the TG, have also become popular in new production areas. Similar to hazelnut geographical origin, the qualitative distinctions and market price variations between cultivars make hazelnuts particularly susceptible to economically motivated varietal mislabelling.

The situation with **pine nuts**, often referred to as "white gold" due to their high price, mirrors that of hazelnuts. Pine nuts, despite accounting for only 1% of global tree nut production, are of significant commercial importance, with their supply value exceeding 1.3 billion USD per year [46]. Primarily consumed in EU and the USA, pine nuts and are an essential ingredient in many traditional dishes. Despite Europe being the main consumer, pine nuts leading producers (China, Russia, and North Korea), are located in Asia. The Mediterranean region ranks second in production, with Turkey and the Iberian Peninsula as major contributors [46]. In this context, the organoleptic characteristics and nutritional profiles of pine nuts vary significantly depending on their species and origin [52-54]. These differences are also reflected in their market prices. Iberian pine nuts are highly valued and appreciated by consumers and can fetch prices exceeding 100 EUR/kg, while Asian pine nuts are valued at less than a half that amount [54,55].

Unlike many agricultural products that have cultivars grown worldwide, pine nut species are closely linked to their geographical origin. Among the 20 commercially available species, the most common are *Pinus koraiensis* Siebold & Zucc (*P. koraiensis*) and *Pinus sibirica* (*P. sibirica*), mainly grown in China and Russia, respectively, and *Pinus pinea* L. (*P. pinea*), predominantly found in the Mediterranean basin [52,55].

Their high value, combined with the significant price discrepancy, makes pine nuts highly vulnerable to lucrative fraudulent practices that misrepresent their origin and species to exploit the price disparity between Mediterranean pine nuts (*P. pinea*) and

other species and origins, such as Asian *P. koraiensis* and *P. sibirica*. The situation is further aggravated by the lack of efficient control methods.

Although food fraud generally poses health risks due to the compromised traceability, misrepresenting the origin of pine nuts entails a specific threat, given that Chinese *P. koraiensis* is sometimes marketed mixed with other pine seed species that may include *Pinus armandii* Franch. This species has been linked to the dysgeusia called 'pine mouth syndrome' (PMS), an unpleasant taste disturbance that can last for days or even weeks [56,57].

1.2.2 Legal framework

Even though **geographical origin** it is not strictly regarded as a quality parameter, marketable VOO and nuts can have very different organoleptic and technological qualities based on their provenance and cultivar or species [38,39,44,45,52-54]. This can affect consumers' choices and lead to price variations within the same category [15,16,58]. To prevent market distortion, protect consumers from misleading information and to enable them to make informed decisions at the point of sale, the EU has mandated geographical labelling for several types of food products. Regarding plant-based food products, Commission Delegated Regulation (EU) 2023/2429 supplementing Regulation (EU) No 1308/2013 "on marketing standards for the fruit and vegetables sector, certain processed fruit and vegetable products and the bananas sector", establishes that fruits and vegetables intended to be sold fresh to consumers may only be marketed if the country of origin is indicated [59].

For **hazelnuts** and **pine nuts**, the regulation specifies that shelled hazelnuts and pine nuts "shall not be required to conform to the marketing standard except regarding the indication of the country of origin" [59], thereby making the declaration of the country of origin mandatory for these nuts.

The mandatory declaration of origin is more detailed for **VOOs**, as stipulated by Commission Delegated Regulation (EU) 2022/2104 which specifically addresses marketing standards for OO packaging and labelling [60]. This regulation requires specifying whether the VOOs come from the EU, a specific EU member state, or a third country. The place of origin must refer to the geographical area where the olives were

harvested and where the mill that extracted the oil is located. If the oil extraction mill is situated in a different Member State or third country from where the olives were harvested, this shall also be declared. Additionally, if the VOO is produced in multiple EU or non-EU countries, or is a blend of both EU and non-EU origins, this blend should be clearly stated [60].

For products under the EU's geographical quality schemes, the geographical origin is denoted by specific designations such as **PDO** and **PGI** according to the EU regulation No 2024/1143. These GIs were established to promote and protect regional foods and traditional agricultural and industrial practices [61]. They are IPRs that identify products originating in a specific place, region or country, whose quality, reputation or characteristics are essentially or exclusively attributable to a particular geographical origin [61,62]. The main difference between PDO and PGI is that the first one requires that all the production steps take place in the defined geographical area, whereas PGI only demands that at least one of the production steps takes place in the defined geographical area [61]. Currently, there are over 140 registered PDO and PGI designations for VOO [63], while for hazelnuts there are three PGIs (*Nocciola del Piemonte* and *Nocciola di Giffoni*, Italy; *Noisette de Cervione*, France) and three PDOs (*Nocciola Romana*, Italy; *Giresun Tombul Fındığı*, Turkey; *Avellana de Reus*, Spain). In contrast, no GIs are currently registered for pine nuts.

In contrast to geographical origin, specifying the **cultivar or species** on VOO and nut labels is not mandatory. Nevertheless, as non-expert consumers often find it difficult to distinguish between nuts from different cultivars or species, international standards for hazelnuts [64] and pine nuts [65] recommend declaring the cultivar or species on the label to ensure detailed quality criteria. In addition, many producers voluntarily include this information on the label, because it positively influences consumers' perception of quality and can result in higher sale prices. On the other hand, products under GIs are often required to use specific traditional cultivars that are characteristic of a particular region, meaning that although cultivar labelling is not mandatory, it is implicitly indicated by the GI designation [38,39,62].

1.2.3 Current fraud incidence and prevention measures in virgin olive oil and nuts

Fraudulent activities involving **OO** have been reported for a long time, due to the substantial financial profits it offers and to the reported vulnerability to fraud in the VOO supply chain [66]. Currently, OO continues to be one of the most frequently notified food products associated with suspected fraud [31]. According to the Food Fraud Monthly Reports published by Joint Research Centre (JRC, Brussels, Belgium) for the first semester of the 2024, which summarise media articles on food fraud published by the EU Commission [67], a significant percentage of fraud cases were related to mislabelling issues, including false declarations of origin. This agrees with the outcomes of previous years [68] and shows an increasing tendency. A study on conformity checks in the OO sector across the EU [69], reports that VOOs with fraudulent claims about cultivar, as well as omissions and irregular or misleading use of origin indications, are among the main infringements detected.

Nuts, including **hazelnuts** and **pine nuts**, are high-value products whose economic significance is greatly dependent on their provenance and botanical origin, making them highly susceptible to fraudulent practices [3,25]. Their relabelling with old or expired stocks, their replacement with cheaper cultivars or provenances, or their mixing with different species, pose serious risks for consumers with allergies and intolerances [71]. Between 2000 and 2015, RASFF documented 1,035 notifications of "adulteration/fraud," with nuts comprising 16.3% of these cases [72]. In 2016, nuts and seeds accounted for 4.8% of the 147 EU fraud reports [73]. Nuts are among the most reported foods by RASFF and one of the predominantly linked to fraud hazard [74-76]. The main fraud in hazelnuts and pine nuts regards basically the misrepresentation of their origin and cultivar as the market prices fluctuate greatly accordingly [70,57].

Geographical and botanical label claims need to be verified to safeguard consumer rights and prevent fraud. Official controls to verify VOO, hazelnuts and pine nuts geographical and varietal origin, as well as GI quality schemes, currently rely on documented traceability of production, as there is no overall accepted analytical method for verification yet. The lack of an official analytical method to verify the geographical and botanical origin, despite legal requirements or recommendation on labels, is a **critical gap** in preventing fraud.

In the case of VOO, it has been reported as an issue that significantly concerns stakeholders and needs urgent attention [48,58]. For nuts, these documental controls can be supported by phenotypic observations of physical traits such as kernel morphology [70,45,77-80]. However, their reliability is often compromised by susceptibility to external factors and their limitation to whole kernels. Therefore, there is a critical **need to develop analytical tools** that can effectively **authenticate the geographical and botanical origins of VOO and nuts**.

Appropriate markers, methodologies and strategies to authenticate VOO, hazelnuts and pine nuts will be discussed in depth in the following sections 1.3 and 1.4.

1.3 Food geographical and varietal authenticity markers

1.3.1 General authenticity marker requirements and marker classes

A food authenticity marker is a distinctive identifier that can be used to directly or indirectly verify that a product's features match the claims made about it. This marker must enable, either individually or in combination with other markers, to distinguish between a genuine product and a non-authentic one [81].

The selection of a marker is the initial step in analytical food authentication. Therefore, a suitable marker must:

1) **Fit the specific authentication purpose** by reliably distinguishing between authentic and non-authentic samples.
2) Be **directly related to the property of interest**, with its presence, distribution, or concentration in the sample depending on this property.
3) Be **minimally influenced by factors other than the property** of interest.
4) Additionally, the methods used to analyse these markers should be reliable, robust, cost-effective, fast, and automatable to be **suitable for routine analysis**.

In this context, varietal and geographical food authentication markers must:

1) Be able to **distinguish a food** product from a **specific species or cultivar** or a **specific provenance,** from similar products of different origins, species and cultivars.

2) Be **directly related to the varietal and geographical origin of the food** product, hence its **presence or concentration** must be **driven by genetic and pedoclimatic factors**.
3) Be **minimally influenced by** other factors such as the **processing of the product or the storage conditions**. Hence must be **robust** and **stable, not easily degraded with time and storage conditions** (i.e. light, temperature, humidity, time).

The main marker classes and their related analytical techniques, applications and limitations are summarized in **Table 2** [81,82].

The most widely recognized methods for assessing the geographical origin and biological identity of agricultural food products are isotopic and genetic markers, respectively. **Isotopic markers** establish a strong connection between agri-food products and the environmental factors of their specific zones of origin, while **genetic markers** offer clear evidence of the product's genetic heritage. In contrast, **metabolic markers**, directly or indirectly driven by both genetic and environmental factors, have the potential to provide insights into both the biological identity and provenance of agri-food products.

In the following sections, the principles behind these key marker classes for assessing the biological identity and geographical origin of food will be examined. Their application in addressing key authenticity challenges in VOO, hazelnuts, and pine nuts will be also discussed in subsequent sections (sections 1.4.1-1.4.5).

Table 2. Main marker classes, analytes, analytical techniques, applications and limitations adapted from Bayen et al. [81].

Marker class	Analytes	Techniques	Applications	Limitations
Elements/Isotopes	Common Ratios: $^{2}H/^{1}H$, $^{13}C/^{12}C$, $^{15}N/^{14}N$, $^{18}O/^{16}O$, $^{34}S/^{32}S$, $^{87}Sr/^{86}Sr$	Isotope ratio mass spectrometry IRMS or inductively coupled plasma mass spectrometry ICP-MS	Verification of geographical origin Verification of agricultural practices Verification of declared processing Detection of fraudulent practices	Variations in stable isotopes abundance Metabolic turnover
Metabolites	Small molecules: sugars, lipids, amino acids, fatty acids, phenolic compounds, alkaloids. Mass range: 50–1500 Da	Chromatography + MS, vibrational spectroscopy, nuclear magnetic resonance NMR	Verification of geographical origin Differentiation of species, varieties Verification of production methods Detection of fraudulent practices	Confidence of identification results
Peptides/Proteins	Whole proteome, subset of proteins, characteristic peptides, structural modification of peptides	Gel-based: 2D-electrophoreses, lab-on-chip microelectrophoresis MS-based: MALDI-TOF-MS, LC-MS/MS, LC-MS	Differentiation of species, varieties Verification of processing and storage conditions Detection of fraudulent practices	High complexity of peptides/proteins in food samples
DNA/RNA	Whole genome, highly conserved target genes, nuclear DNA, mitochondrial DNA, plastid DNA	Sanger sequencing, NGS, PCR-techniques, RAPD, CRISPR-based nucleic acid assays	Verification of biological identity. Detection of fraudulent practices.	DNA degradation due to processing

Abbreviations: DNA, deoxyribonucleic acid; IRMS, isotope ratio mass spectrometry; ICP-MS, inductively coupled plasma mass spectrometry; MS, mass spectrometry; NMR, nuclear magnetic resonance; MALDI-TOF-MS, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry; LC-MS/MS, liquid chromatography tandem mass spectrometry; LC-MS, liquid chromatography-mass spectrometry; NGS, next-generation sequencing; PCR, polymerase chain reaction; RAPD, randomly amplified polymorphic DNA; CRISPR, clustered regularly interspaced short palindromic repeats.

1.3.2 Isotopic markers

Stable isotope analysis is one of the most recognized methods for establishing the geographical origin of food, leveraging the influence of factors such as geology, hydrogeology, and agricultural practices on the isotopic composition of agricultural products [23,83-87]. Isotope ratios of specific elements correlate directly with the climate and soil geology of the production area. Over recent decades, this method has become a benchmark for food geographical authentication due to its ability to provide robust and reliable results that are easily comparable and reproducible across different laboratories, enhancing its transferability. This reliability has resulted in the application of stable isotope analysis in legal cases as part of enforcement actions [84].

The most common isotopes for food authentication include light bioelements such as carbon (C), hydrogen (H), nitrogen (N), oxygen (O) and sulphur (S), as well as heavy geoelements, such as strontium (Sr), which have contributed to significant advances in food traceability [23, 88-91]. Isotopic ratios of light bioelements in vegetable products are greatly influenced by geographical factors such as altitude, latitude, proximity to the coast, climate and precipitations [85-88]. In contrast, the isotopic ratios of heavy geoelements, such as Sr, are primarily related to soil characteristics, including age and geological evolution, which vary by soil type. This is because the primary source of Sr in plants is the labile fraction of Sr in the soil, which, once absorbed by the plant, is neither metabolised nor subjected to isotopic fractionation, thereby preserving the Sr isotopic ratio from soil to plant [92-94].

Isotopic variations in nature are usually small, often on the order of a few per mil. Consequently, analysing isotopic ratios demand highly precise measurements. This precision is achieved by comparing the ratio of heavy to light stable isotopes in a sample with that of a reference material with a known isotope ratio. This allows for the accurate detection of even very small variations in isotopic composition. Therefore, for light bioelements, such as H, C, O, N or S, isotopic composition is usually expressed using delta notation (δ) (Eq.1) [86].

$$\delta^{i}(E_{sample/standard}) = \frac{R(\ {}^{i}E/\ {}^{j}E)_{sample}}{R(\ {}^{i}E/\ {}^{j}E)_{standard}} - 1 \qquad \text{(Eq.1)}$$

where R_{sample} and $R_{standard}$ are the isotopic ratios of the sample and the reference material, respectively, and *i* denotes the mass number of the heavier isotope of the element *E*. The result is expressed in per mil (‰), which highlights small differences more effectively than raw isotopic ratios.

In contrast, Sr isotope composition in a sample is generally expressed as the ratio of radiogenic ^{87}Sr to the stable ^{86}Sr, as there is no consensus on a universally accepted reference material, unlike with other isotopes. The $^{87}Sr/^{86}Sr$ ratio is typically normalized against the $^{86}Sr/^{88}Sr$ ratio (0.1194), to correct instrumental mass-dependent fractionation [86].

Regarding the potential of specific **light bioelements** as authenticity markers, the $\delta^{13}C$ ratio is used to distinguish between the C3 and C4 plants, which follow different metabolic pathways to fixate and assimilate atmospheric CO_2 during photosynthesis. This ratio has been applied to detect food adulteration, such as the fraudulent sweetening of honey by adding sugar of C4 plants [95]. Additionally, the $\delta^{13}C$ values can be influenced by the plant's developmental stage and several environmental conditions such as humidity, precipitation, ambient CO_2 concentration and temperature [90,96-99]. For olive trees, $\delta^{13}C$ measurement has been used to assess olive ripening index and to track climatic changes during the fruit-ripening season [100].

The $\delta^{2}H$ and $\delta^{18}O$ values are linked to the water dynamics of the cultivation area. These isotopic ratios in water result from fractionation processes occurring during water phase changes (solid-liquid-vapour) [85,86]. Hence, they are influenced by geographical factors such as latitude, longitude, elevation, and distance from the sea, as well as climatic conditions including temperature, precipitation and evaporation rates. Additionally, plant type can influence evaporation rates through transpiration [93]. $\delta^{18}O$ ratios can also be affected by several other sources such as atmospheric oxygen and photosynthesized CO_2 [85,101].

Stable isotopes of light elements are usually analysed in bulk matrices by elemental analysis-isotope ratio mass spectrometry (EA-IRMS), a method that requires minimal sample manipulation and short analysis time. However, recent advancements are

shifting beyond bulk analysis toward compound-specific isotopic analysis (CSIA), which provides more detailed insights and enhances the ability to discriminate between samples from different geographical origins [88,101,102]. This approach involves analysing the isotopic ratios of elements in specific food components, such as proteins, lipids, and carbohydrates [88]. It typically requires the integration of separation techniques, such as gas chromatography (GC) or high-performance liquid chromatography (HPLC), with IRMS. This approach has already been tested in a wide range of foods showing promising results for geographical authentication [101, 103-106].

Regarding trace **heavy geoelements** such as Sr, their composition in plants and derived foods directly reflects the bioavailable minerals present in the soils where they are cultivated. Alkali and alkaline earth metals, such as Sr and rubidium (Rb), are easily mobilised from the soil into plants, being sufficiently abundant to be used as indicators of a plant's geographical origin [86,93]. Sr isotopic analysis relies on the ratio of radiogenic ^{87}Sr, formed through the slow radioactive decay of ^{87}Rb (with a half-life of 49 billion years), to stable ^{86}Sr. Moreover, as previously mentioned, for heavy elements like Sr, isotope effects and mass fractionations are minimal due to the small relative differences in isotopic mass. Bioavailable Sr, the water-soluble fraction, moves from soil to water, eventually entering the food chain by being absorbed by plants and distributed throughout the plant, including leaves and fruits, with negligible fractionation. Consequently, the Sr isotopic ratios in plants depends on the concentrations of Rb and Sr and the geological age of the underlying bedrock. For instance, young rocks with low Rb/Sr such as basaltic and carbonate rocks, tend to have lower $^{87}Sr/^{86}Sr$ values, while older acidic crystalline rocks (e.g. granites) with high Rb/Sr ratios have higher $^{87}Sr/^{86}Sr$ values [86,92,93, 107].

The Sr isotopic analysis is typically conducted using thermal ionization mass spectrometry (TIMS) or multi-collector inductively coupled plasma mass spectrometry (MC-ICPMS) [107]. Sample digestion and separation of Sr from matrix elements are essential due to the isobaric interference between ^{87}Sr and ^{87}Rb [107]. While TIMS was the original method for precise measurement of heavy elements like Sr, it has drawbacks such as limited ionization efficiency for elements with high ionization

energy, extensive sample preparation, and long measurement times. MC-ICPMS offers advantages over TIMS, including higher ionization yield, versatile sample introduction systems, and simultaneous isotopic determinations, which eliminate time-dependent mass fractionation. However, TIMS is less affected by matrix effects and has lower instrumental isotopic fractionation. Both techniques are highly precise but come with high costs and require skilled operators [93,108,109].

It is important to note that determining food geographical origin is a complex issue that cannot typically be resolved by analysing a single isotopic marker. An accurate determination of food geographical origin usually requires analysing multiple isotopic parameters [83,84,87,101,110]. The accuracy of this process is further enhanced when multi-isotopic ratio analysis is combined with molecular-specific isotopic data [101]. Finally, it is important to mention that among the main limitations of light element isotopic markers are their inter-annual variations in abundance and potential changes due to metabolic turnover, which can impact the accuracy of distinguishing between products from different origins.

1.3.3 Genetic markers

Genetic markers provide a powerful tool for addressing varietal authenticity and traceability challenges in the agri-food sector. DNA-based methods exploit these markers to detect and identify specific variations in the nucleotide sequences within a genome enabling the distinction between and within species [111-113].

Among the various DNA-based techniques, PCR is particularly prominent. This method is based on the amplification of highly specific DNA fragments. In the context of food species and cultivar authentication, one of the most commonly used DNA markers are simple sequence repeat (SSR), also known as microsatellites. SSR are short sequences of DNA, typically 2 to 6 base pairs long, that are repeated in tandem multiple times in a genome. SSRs are highly polymorphic, meaning that the number of repeat units can vary significantly among individuals, and these differences can be easily detected by PCR. The high reproducibility and polymorphism degree of SSR, make them a fit for purpose marker for various applications including varietal identification and adulteration detection [111,112,114].

Overall, DNA-based methods are highly sensitive and effective for food varietal authentication, however, their application can be limited by technical complexity, DNA degradation due to processing, cost, and the need for extensive reference databases [81,111,114-117]. For instance, vegetable oils and their derived products contain low amounts of DNA, which can be significantly degraded during extraction and refining processes, making it challenging to obtain sufficient high-quality genetic material for reliable authentication [116,117]. Despite these challenges, DNA-based methods remain the reference method for verifying food botanical origin.

1.3.4 Metabolic markers

Metabolic markers hold significant potential in both geographical and varietal authentication because plant metabolic profiles are heavily influenced by both genetic traits and environmental conditions. Hence, metabolic profiles can be effectively linked to specific cultivars and geographical regions, enabling reliable authentication.

Secondary metabolites are small organic molecules produced by plants, fungi, and bacteria that, although not essential for growth, development, or reproduction, are crucial for an organism's interaction with its environment. Unlike primary metabolites such as carbohydrates, proteins, and nucleic acids, secondary metabolites are not required for basic survival but play a key role in helping organisms adapt and thrive in their ecological niches [118,119]. These compounds are structurally diverse and are typically synthesized in response to environmental stressors, both biotic (e.g., pathogens, herbivores, competition) and abiotic factors (e.g., ultraviolet radiation, extreme temperatures, nutrient scarcity). They play a crucial role in defence mechanisms, protecting against environmental stresses, and facilitating interactions with other organisms, such as attracting pollinators or deterring herbivores [118,119].

The production of secondary metabolites is influenced by both environmental and genetic factors. Environmental stress can trigger their biosynthesis, leading to increased concentrations of these compounds [119,120]. For example, plants may upregulate specific secondary metabolites in response to herbivory or pathogen attacks to defend themselves. Similarly, abiotic stresses like drought or heavy metal exposure can induce the synthesis of protective secondary metabolites, helping the

plant to cope with the adverse conditions. Genetics also play a significant role in the production and variability of secondary metabolites. The composition and concentration of these compounds can vary widely among species and even among individuals of the same species, depending on their genotype [121]. This variability is often exploited in chemotaxonomy, where secondary metabolite profiles serve as markers for botanical classification and differentiation between closely related species [119].

Consequently, in the context of food authentication, secondary metabolites have emerged as powerful markers to verify the provenance, varietal and botanical origin and cultivation practices [111,118,122-124]. In addition to secondary metabolites, primary metabolites can also be affected by these same factors, although their primary roles are related to growth and basic physiological functions.

Some specific metabolic fractions hold great potential as effective geographical and varietal markers, as they contain particular classes of metabolites that are significantly influenced by environmental and genetic factors. The protein or phenol fraction of various food products have been shown to be highly influenced by the cultivar and pedoclimatic conditions [82,113,125-129]. However, they do not fully meet the criteria of stability and minimal dependence on technological and storage conditions [126,130,131]. Their lack of stability under some processing conditions or storage may call into question their suitability as authenticity markers. In contrast, methods based on lipid metabolites such as fatty acids (FAs), triacylglycerols (TAGs) and sterols have been tested to check for the authenticity of lipid-rich foods regarding the cultivar and provenance [132,133], even in highly processed foods [134,135].

Previous research has demonstrated that mono- and sesquiterpenes could be reliable markers of varietal and geographical origin of different plant species and vegetable-derived products such as spices, alcoholic beverages and oils [136-141]. In particular, sesquiterpene hydrocarbons (SH) are secondary semi-volatile metabolites presenting a structure with three isoprene units, following the general formula $C_{15}H_{24}$. Despite sharing the same molecular formula, SH exhibit remarkable structural diversity due to the action of various sesquiterpene synthases which results in multiple products, usually derived from common intermediates [142,143]. Their presence, types and

concentration have been shown to vary significantly in VOOs from different origins and cultivars due to the effect of genetic factors and environmental conditions of the olive trees' growing area, in the SH profile [143]. On top of this, what makes SH efficient authenticity markers is not solely their significant reliance on genetic and pedoclimatic factors, but also its stability to oxidation and minimal influence of other technological and processing factors [143]. Different techniques have been applied for the analysis of semi-volatile minor compounds, such as SH, in food matrices [136-138,141,143-146]. Due to their low concentration, most of them require an extraction technique, coupled to a separation method, usually GC, and a sensitive detection system, such as MS [143,147,148]. Among the extraction techniques, headspace solid-phase microextraction coupled with GC-MS (HS-SPME-GC-MS) stands out as an effective approach for SH analysis. This technique is particularly well-suited for routine analysis due to its simplicity, automation potential, and ability to perform direct, solvent-free sample analysis [139,140,143]. Moreover, HS-SPME-GC-MS is widely available in food analysis laboratories, requiring minimal additional investment in instrumentation. These features make it an ideal choice for routine monitoring, offering a cost-efficient, rapid, environmentally friendly, selective, and sensitive solution for analysing semi-volatile compounds.

Since the volatile terpenoid fraction may not be prominent in all vegetable-based foods, other fractions containing secondary metabolites that are responsive to relevant variables could be explored and utilized for authentication purposes. The unsaponifiable fraction (UF) of oils and lipid-rich vegetable foods includes several secondary metabolites (hydrocarbons, linear and terpene alcohols, sterols, methylsterols, dimethylsterols) [149-152], whose presence, composition and concentration are driven by both environmental [150,153,154] and genetic factors [44,154,155]. Moreover, most of these metabolites are rather stable to oxidation, storage and processing conditions such as time, light, temperature. Therefore, the UF fulfils the key requirements of an authenticity marker and shows to be a promising tool for varietal and geographical authentication of plant-based foods. The UF can be analysed by GC–MS, which offers high sensitivity, detailed molecular-level information, and is available in many routine laboratories. However, this method requires a previous

sample preparation in order to isolate this fraction from the food matrix, which can be laborious and time-consuming, limiting its applicability to a reduced number of samples. Then, alternative markers that can be analysed using simpler and faster procedures can enhance analytical efficiency.

In this regard, TAGs are the major components of the lipid fraction. Their concentration, composition and the arrangement of the FAs that conform them, have been reported to be determined by genetic and environmental conditions [44,133]. Traditionally, official methods for the analysis of major lipid components are based on the conversion of TAGs to FA methyl esters (FAMEs) and analysed through GC [156,157]. However, these approaches result in the loss of the information of the stereospecific distribution of FAs on the TAG, making direct TAG analysis the preferred method to retain this information [158]. This is usually performed by HPLC separation coupled to refraction index detector [158-163]. In this context, high-temperature–gas chromatography (HT-GC) has shown promising results for TAG separation in numerous food matrices [158], and its combination with spectrometric techniques such as MS may become a powerful and sensitive tool for straightforward TAG analysis involving minimal sample preparation and suitable for routine analysis.

On the other hand, following a more general approach that enables the analysis of multiple chemical compounds without extensive analytical preparation could be a suitable solution. In this regard, spectroscopic techniques such as mid-infrared (MIR) and near-infrared (NIR) spectroscopy, offer several analytical advantages, being faster, simpler and non-destructive techniques that can be applied on-site. These techniques relay on the unique infrared absorption patterns generated by the vibrational and rotational transitions of the molecular chemical bonds when the sample is irradiated by infrared light, leading to changes in the dipole moments of the molecules [164]. The infrared spectral range is divided in three regions from more to less energetic: near-infrared (800 to 2500 nm), mid-infrared (2500 to 25000 nm), and far-infrared (25000 to 500000 nm). In NIR, the spectral bands come from overtones and combinations of fundamental vibrations, providing information about functional groups and chemical bonds. Whereas MIR bands are the result of the fundamental vibrations of specific

functional groups, providing a better resolved spectra containing highly specific molecular information that can be used for structural identification [164].

1.4 From established to emerging authentication approaches

In addition to selecting the appropriate authentication markers and analytical techniques, an essential aspect of the authentication strategy is the data analysis approach. The following section discusses the main data analysis approaches, along with their advantages and limitations.

1.4.1 Definitions, pros and cons of authentication approaches

The traditional analytical approach for food authentication is **targeted analysis**. This approach involves the **identification** and **quantification**, or semi-quantification of one or more predefined analytical targets by comparison to a known analyte [165,166]. By this approach, a food product is considered authentic if the values of the targeted markers comply with the reference limits. This methodology is robust and easily transferable, as its performance can be validated according to international standards [167]. These standards employ well-established validation parameters such as limits of detection and quantification, accuracy, measurement of uncertainty, selectivity, and robustness, which account for variations in laboratory conditions, instruments, technicians. These validation procedures together with the application of univariate data analysis facilitate the application of targeted approaches in official control. Targeted methods can be classified according to the number of compounds considered: **single/dual targeted analysis,** which focuses on one or maximum two analytes as markers, and **targeted profiling analysis**, where multiple compounds are identified and quantified as analytical markers [Publication 1].

Nonetheless, both of these targeted approaches share the significant **drawback** of offering only **limited information**. In fact, targeted approaches only consider a portion of the analytical output, leading to the loss of valuable information [165]. In addition, the difficulties in comprehensively identifying and quantifying the target compounds in the samples due to the presence of minor species, not well-resolved peaks, coelutions, complex identification or lack of standards, further limit the information they can extract [168]. Moreover, these approaches require labour-intensive identification,

integration and quantification procedures, which are often performed manually. Therefore, as the number of selected target compound increases, these processes become more cumbersome, making the methodology more time-consuming and less cost-efficient.

In contrast, **untargeted approaches** are based on the simultaneous analysis of **numerous non-predefined compounds** (> 100), in the case of **untargeted profiling** [165], or on the **data points of the raw analytical signal**, for the **fingerprinting approach** [81,169,Publication 1].

Untargeted profiling analysis allows to automatically detect and semi-quantify a high number of compounds, **providing** a great amount of **chemical information**. It often requires **deconvolution tools** to **unmix the complex analytical signal into the individual compound signals within a spectrum and relate them to the relative concentrations of the compounds that constitute the sample** [170,171,Publication 1]

The **fingerprinting approach** involves analysing the **data points of the raw analytical signal** to **identify specific patterns**, known as **fingerprints**, which are **unique to a certain characteristic of the sample** [81,168,169,Publication 1].

Untargeted approaches overcome the main drawbacks of the targeted methods by considering a larger portion or even the whole available information, and by eliminating the need for time-consuming identification and quantification steps. However, the transferability and applicability of these methods as standard control references are limited by the absence of reliable validation procedures [81,172-174]. This challenge is particularly pronounced for fingerprinting methods used in classification or categorical prediction, where traditional validation metrics—such as linearity, standard deviation, and limit of detection—are not applicable.

Untargeted methods can also be classified according to the techniques used through the analytical determination, mainly distinguishing between **spectroscopic** and **chromatographic untargeted methods**.

1.4.2 Chromatographic untargeted methods

Untargeted methods usually require signal preprocessing for correcting instrument-derived offsets that can introduce analytical noise into the data. Preprocessing techniques applied to spectroscopic untargeted methods have been extensively studied, and strategies to overcome scattering effects, such as Multiplicative Scatter Correction, Standard Normal Variate (SNV), are well-known and usually applied. But untargeted methods applied to chromatographic data face several intrinsic challenges related to the separation technique and to the management of the particular chromatographic offsets. In this regard, signal preprocessing is crucial for the interpretation of chromatographic fingerprints. This preprocessing of the chromatographic signals needs to address distinct sources of analytical variability. First, baseline correction methods may be required to solve baseline drift generated by column bleeding [175-179]. Then, normalization techniques are used to correct intensity differences between batches and to balance weights among samples (**Figure 5b**). These techniques consist in dividing the chromatographic profiles by a scalar, such as the Euclidean norm, the maximum value of the chromatogram or the total sum of each sample [180,181]. Finally, correcting retention time shifts among samples is especially critical for chromatographic fingerprints and is achieved by applying alignment methods (**Figure 5c**) [Publication 1]. Alignment is defined as "a mathematical operation where similar chemical features are repositioned so that they appear at the same elution time in different runs" [168]. There are different alignment techniques, but correlation optimized warping (COW) stands out among them [182-184]. The COW algorithm is a flexible method based on piecewise linear stretching or compression to align chromatograms towards a reference, without altering peak positions [182,183,185]. **Figure 5** shows a global scheme of signal preprocessing in chromatographic fingerprints.

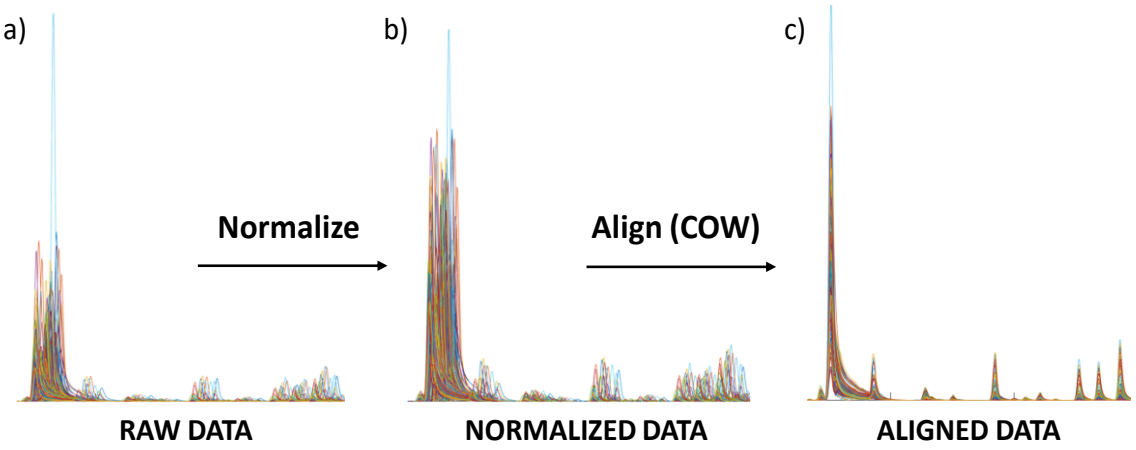


Figure 5. Representation of the signal preprocessing in chromatographic fingerprints. a) raw data, b) normalized data, c) aligned data. COW: correlation optimized wrapping algorithm.

In addition, when chromatographic separation is coupled to techniques such as MS, the data analysis may involve handling three-way data (retention time x intensity x *m/z*).

Detailed information regarding the challenges that arise from untargeted approaches applied to chromatographic techniques and strategies to address them are thoroughly discussed in Publication 1.

1.4.3 Publication 1

The volatile metabolome - GC-MS approaches in the context of food fraud

Beatriz Quintanilla-Casas, Berta Torres-Cobos, Rasmus Bro, Francesc Guardiola, Stefania Vichi, Alba Tres

Accepted for publication in Current Opinion in Food Science 2024

The volatile metabolome - GC-MS approaches in the context of food fraud

Beatriz Quintanilla-Casas[a], Berta Torres-Cobos[b,c], Rasmus Bro[a], Francesc Guardiola[b,c], Stefania Vichi[b,c,*], Alba Tres[b,c].

[a] Department of Food Science, University of Copenhagen, Rolighedsvej 30, DK-1958, Frederiksberg C, Denmark

[b] Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona. Av Prat de la Riba, 171. 08921 Santa Coloma de Gramenet, Spain

[c] Institut de Recerca en Nutrició i Seguretat Alimentària (INSA-UB), Universitat de Barcelona. Av Prat de la Riba, 171. 08921 Santa Coloma de Gramenet, Spain

*Corresponding author: Stefania Vichi, stefaniavichi@ub.eduDepartament de Nutrició, Ciències de l'Alimentació i Gastronomia, Campus de l'Alimentació Torribera, Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona. Av Prat de la Riba, 171. Edifici Gaudí; 08921 Santa Coloma de Gramenet, Spain.

ABSTRACT

This work discusses food authentication tools based on the volatile metabolome assessed by gas chromatography-mass spectrometry (GC-MS). Volatile organic compounds (VOCs) serve as markers for food flavour and aroma and, in some cases, they hold potential for food authentication. On top of this, advanced data analysis approaches enhance analytical data extraction and understanding. Untargeted methods provide comprehensive authenticity insights, surpassing traditional targeted approaches, and automated processing improves robustness and reduces user dependency. Chemometric tools, particularly classification methods, are extensively used for authentication based on both GC-MS fingerprinting and untargeted profiling of VOCs. Despite challenges in transferability, untargeted methods perform well in authentication models, making them valuable for internal quality control and official control guidance for fraud detection. This work highlights the need for further research to establish untargeted authentication tools as official methods in the near future.

Keywords: volatile organic compounds; VOCs; GC-MS; authentication; untargeted; fingerprinting; chemometrics.

Food fraud: definition, present state, and current gaps

Food fraud is a complex problem, prompting many definitions to identify its elements and, more importantly, to determine what is needed to solve it in the current context [1]. Despite slight differences, key criteria are present in all definitions of food fraud: i) Violation of the corresponding food law; ii) full intentionality; iii) economic gain; and iv) customer deception. For a case to be classified as food fraud, all four elements must be met. While economic motivation is the primary driver behind food fraud practices, these acts pose a public health threat of uncertain magnitude. Food safety concerns are generally associated with adulteration, especially because of emerging adulterants that evade the current food protection system [2]. Nevertheless, the lack of traceability in any fraudulent food product – including counterfeiting and mislabelling – can also present a health risk to consumers.

According to the criminalistic approach of food fraud defined by Spink & Moyer [3], fraud opportunities can be significantly reduced by implementing efficient control measures. The 2022 annual report of the Alert and Cooperation Network within the European Commission highlights gaps in food fraud control by showing the food categories with high number of notifications, as well as the fraudulent practices most frequently encountered, affecting more than one country [4]. This is the closest approximation for quantifying food fraud cases, as many instances go unnoticed. The top five notified food categories are “Honey & royal jelly”, “Live animals”, “Meat & meat products”, “Fish and products thereof”, and “Fats & oils”. The most notified agri-food fraud categories and subcategories are shown in **Figure 1**, standing out “Substitution (Adulteration/product tampering)”, followed by “Mislabelling”, which includes false quality terms and geographical claims. Lastly, related to counterfeit, some notified cases involved Protected Designations of Origin and Protected Geographical Indications.

Current official authentication methods rely on analytical techniques where one or few compounds are investigated to distinguish authentic from non-authentic products. Unfortunately, fraudsters often stay a step ahead of control methods by exploiting the low efficiency of certain methods, the lack of sensitivity of some techniques e.g., limits

of quantification; or the fact that methods focus on a single compound or compound family. Additionally, for certain commodities like virgin olive oil, verifying the declared commercial category involves assessing sensory quality through a panel test procedure, which may present drawbacks such as limited sample assessments per day and inconsistencies among panels [5]. Therefore, being up to date about the state-of-the-art fraudulent practices is key to develop fit-for-purpose authentication methods. Developing optimal authentication methods requires several elements: selecting reliable chemical markers for the intended purpose, measuring them through a high-throughput yet affordable analytical technique, and applying suitable chemometric methods to build effective authentication tools.

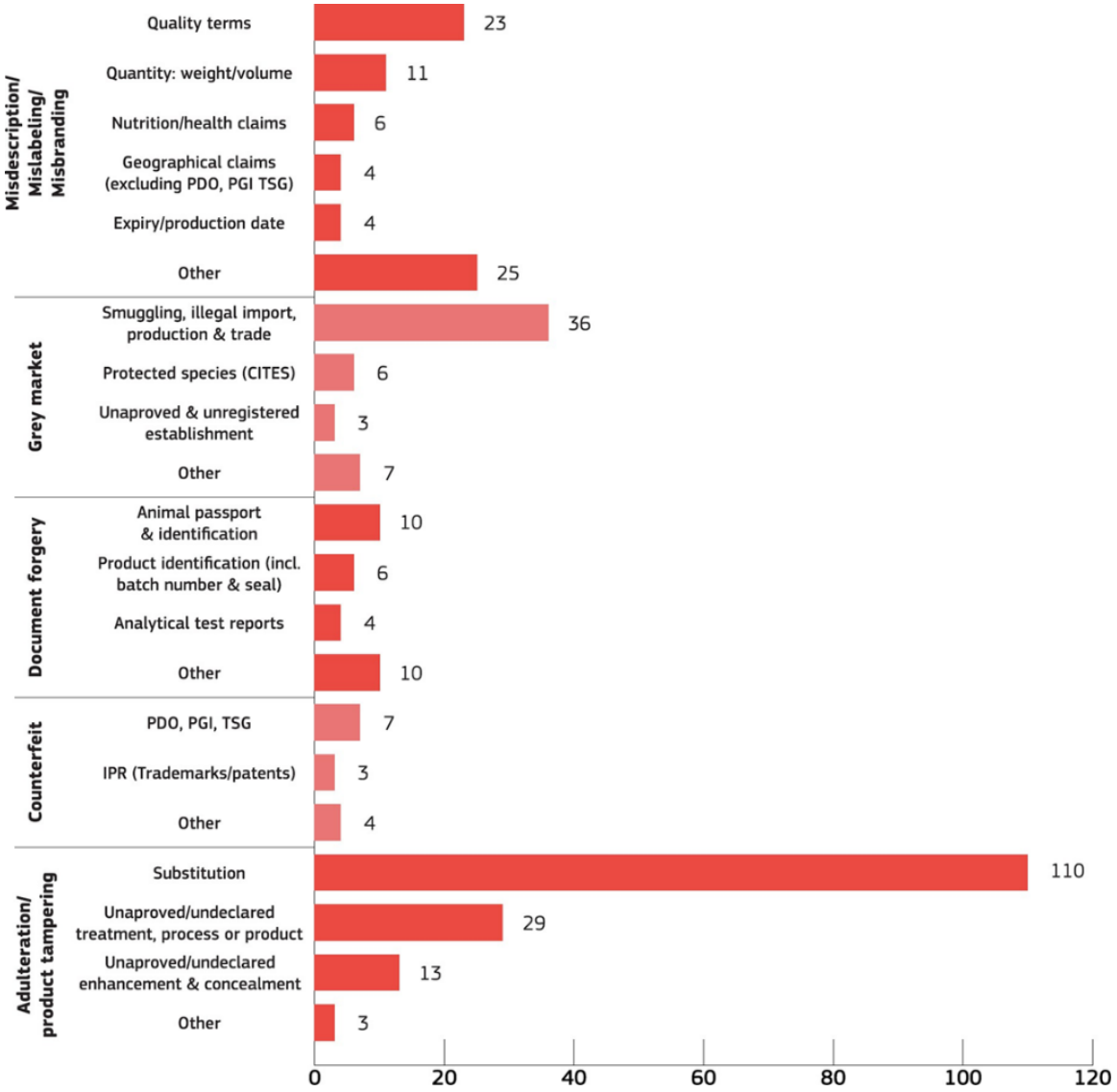


Figure 1. Notifications by agri-food fraud categories and subcategories [4].

Food volatile metabolome: definition and implications in food quality and authenticity

The volatile metabolome refers to the array of VOCs produced in foods through various chemical or metabolic pathways. These compounds may originate from raw materials, influenced by factors such as species, geographical origin, and other variables, or they may form during processing and storage via chemical or enzymatic reactions, or through microbial metabolism. Consequently, they serve as diagnostic markers for various aspects of the food product, including the type, quality, and origin of raw materials, as well as the processing and storage conditions, factors often considered critical for food quality and authenticity. These VOCs, characterized by their small molecular size and propensity to vaporize at room temperature, significantly contribute to the aroma profiles of foods [6]. Consequently, they serve as ideal markers for instrumentally assessing or verifying sensory attributes and overall food quality. For this reason, the volatile metabolome of a various food products has been extensively studied in recent years, particularly those favoured by consumers for their hedonic value, such as wine, olive oil, and coffee [7].

Detecting food fraud through volatile metabolome analysis

Volatile metabolome as authentication marker

The gold standard for food authentication markers is their direct relation with the feature being authenticated. Certain VOCs are responsible for both quality sensory attributes, such as fruitiness in virgin olive oils [8], and undesirable off-flavours, such as off-odours in fish [9]. This is why VOCs and the volatile metabolome have been reported as useful for assessing sensory quality. Assessing VOCs has proven valuable in global sensory evaluation, such as verifying the commercial categories of virgin olive oil [10,11], and in predicting specific attributes derived from endogenous VOCs or those generated during processing and storage, as recently reviewed by Díez-Simon et al. [12].

Besides, the VOC metabolome is influenced by various factors such as seasonal variation, geographical origin, feeding regimes, and storage conditions [13] which is a

blessing and a curse. In the one hand, this apparently offered possibilities of using VOC to diagnose various food characteristics such as the geographical or varietal origin. In fact, promising results were reported in works that tested VOCs as geographical markers [11,14–18]. However, as mentioned above, some VOC may be greatly affected by processing and storage conditions [5], compromising the long-term stability of models authenticating quality aspects unrelated to sensory quality. Consequently, in these cases, it would be advisable to base models only on VOC stable compounds that are less prone to undesired profile modifications during processing and storage. Recent studies have revealed that semi-volatile sesquiterpene hydrocarbons (SHs), secondary plant metabolites, are stable and valid markers for botanical and geographical authentication of certain food products [19–22].

Instrumental assessment of food volatilome: sample preparation and analytical technique

Sample preparation is a critical step when measuring VOCs; therefore, several sampling techniques have been reported. Non-extraction techniques, such as static headspace, might not be ideal for food samples given the presence of key VOCs at trace levels. Among the sampling techniques involving VOC extraction, such as dynamic headspace, purge and trap, or stir-bar sorptive extraction, solid phase microextraction (SPME) has emerged as the predominant method in several food matrices [23]. Proven robust in recent collaborative studies on virgin olive oil [24], this solvent-free, automatable technique requires minimal investment and is well-suited for control purposes.

An important aspect to consider when evaluating the applicability of an authentication tool as a control method is the availability and affordability of the instrumental technique. Gas chromatography coupled to mass spectrometry (GC-MS) is among the main analytical techniques to measure the VOC profile of different kinds of samples, including food products. In addition to its high sensitivity and widespread availability in routine laboratories, GC-MS offers comprehensive molecular-level information through three-way data (an array of intensity × retention time × m/z for each sample). This allows for more reliable compound identification in targeted approaches and enhances the efficiency of untargeted-based modelling compared to other detection techniques, such as flame ionization detectors. Using high-resolution MS instead of conventional

low-resolution quadrupolar and ion trap mass analysers offers improved sensitivity and more reliable identification through exact mass measurements. This enables the detection of trace VOCs, which is crucial for differentiating authentic from non-authentic food products. Despite its potential for food authentication through VOC assessment [25], high resolution MS remains financially inaccessible for many control laboratories.

Data analysis approaches

From targeted to untargeted methods

Conventional methods for food analysis, including VOC assessment, are based on targeted methods. These involve the identification and quantification, or semi-quantification by an internal standard, of one or few pre-defined chemicals (**Figure 2a**).

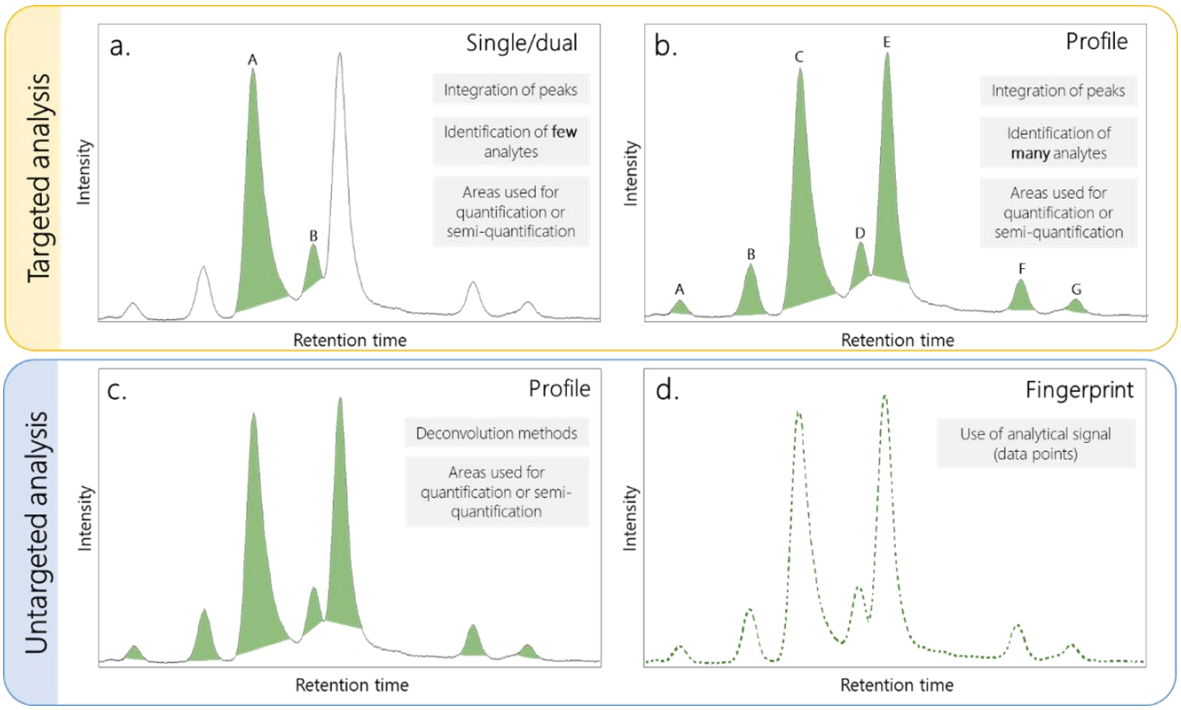


Figure 2. Summary of targeted and untargeted methods applied to chromatography: a) single/dual targeted approach, b) targeted profiling, c) untargeted profiling, and d) fingerprinting. (B Quintanilla-Casas, PhD thesis, University of Barcelona, 2022).

The conformity of the product to a given standard depends on whether the resulting values are above or below an established limit. This strategy is known to be robust, straightforward to transfer and follows well-established validation schemes [26]. The time-consuming step of identification, quantification and manual data extraction

involved in target approach can be facilitated by in silico tools for GC-MS data analysis (**Figure 3**). These include proprietary toolboxes such as ChemStation® (Agilent™), ChromaTOF® (LECO™) MassLynx® (Waters™ corporation), or MassHunter® (Agilent™), as well as open toolboxes such as AMDIS, pyMS and XCMS [27,28]. After peak identification, the mass spectra of identified peaks are usually matched with commercial or in-house spectral libraries to find the corresponding chemical structures. However, beyond workflow optimization, valuable information is lost in targeted approaches since only a small part of the analytical signal is used [26,28]. This fact triggers fraud opportunity because adulteration is way more difficult to detect, as long as the targeted compound stays within the boundaries.

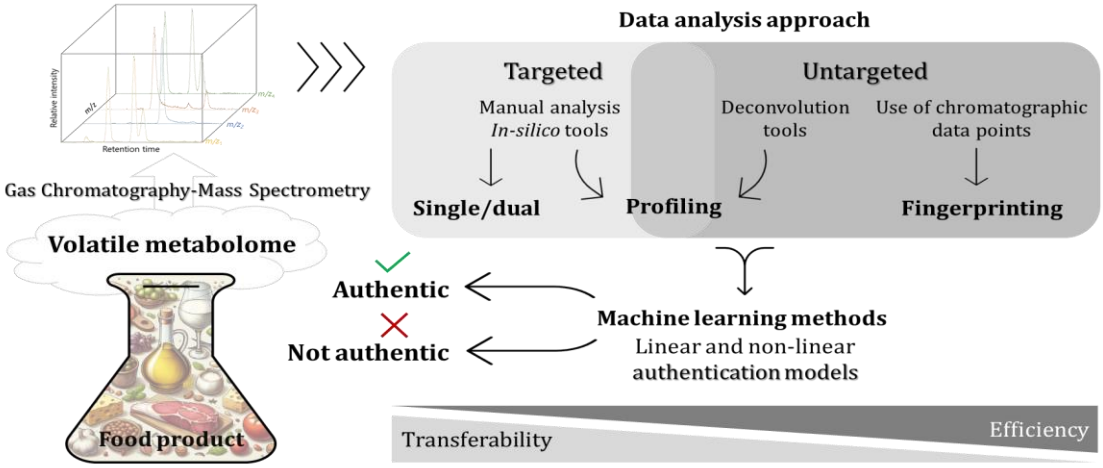


Figure 3. Overview of GC-MS volatile metabolome approaches in food authentication.

The so-called targeted profiling enhances the efficiency of control methods, by selecting a priori multiple target compounds for identification and quantification (**Figure 2b**) [26]. In the context of using the volatile metabolome to detect food fraud, the targeted approach has been successfully applied to assess the sensory quality of virgin olive oil, aimed at complementing the official panel test method [11,24]. Yet, the inconveniences related to the cumbersome identification and integration of analytes, related to detection limits or chromatographic coelution, become more pronounced as the number of compounds under consideration increases. Additionally, when measuring authentication markers such as SHs, the targeted identification is hindered by the similarity of their mass spectra, which causes that many SHs have not been consistently identified yet [29]. As reviewed by Xu et al. [30] despite targeted analysis

might be more sensitive for fruit juice authentication, untargeted methods would be more effective when the fraudulent practice we face remains unclear.

For these reasons, untargeted (or non-targeted) methods are gaining importance in food authentication. On the one hand, the untargeted profiling analysis simultaneously semi-quantifies a high number of not predefined compounds - often >100 [26] - (**Figure 2c**). As previously mentioned, GC-MS data corresponds to a three-way array (**Figure 4a**); therefore, this approach could benefit from advanced data analysis tools to translate the analytical signal into peak tables with the corresponding chemical information. Unlike the above-mentioned toolboxes for GC-MS data analysis, which can also be applied for untargeted approaches, PARAllel FACtor analysis 2 (PARAFAC2) models stand out due to the ability to automatically deconvolute peaks across all samples, including those with retention time shifts, low signal-to-noise ratios, or co-eluted compounds. [27]. Moreover, its integration and automatization through the software PARADISe (PARAllel factor analysis 2 Deconvolution and Identification System) increases the robustness of the results due to the lack of user-interaction, while optimizing the analysis time [31]. This software has been successfully applied on VOC data to assess the quality and authenticity of different foods [8,32,33].

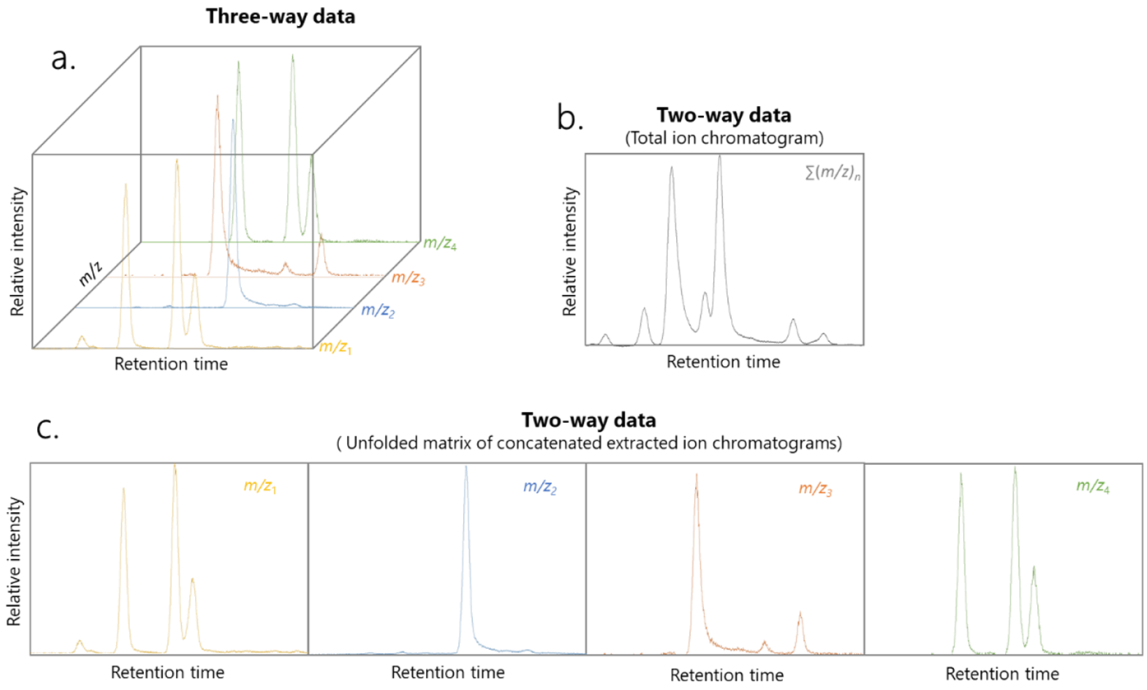


Figure 4. From three-way to two-way data in chromatography coupled to MS (one sample) *m/z*: mass/charge ratio. (B Quintanilla-Casas, PhD thesis, University of Barcelona, 2022).

On the other hand, the state-of-the-art strategy in food authentication involves finding specific patterns in the analytical fingerprint, thus introducing the untargeted fingerprinting approach [34] (**Figure 3**). Some authors consider fingerprints as the compilation of multiple analytical parameters [26] or molecular markers [34], while others define it as the analytical outcome of a single technique [35]. The former definition agrees better with the untargeted profiling approach; therefore, the term fingerprint should refer to the set of data points composing the raw analytical signal (**Figure 2d**) (B Quintanilla-Casas, PhD thesis, University of Barcelona, 2022). Three-way data, such as the GC-MS output, typically require complex multi-way chemometric algorithms. Nevertheless, they can be transformed into two-way data to enable the application of bilinear methods. This can be achieved either by using the total ion chromatogram (TIC) (**Figure 4b**) [10,36], or by unfolding the original array through the extraction and concatenation of all or some of the acquired m/z (**Figure 4c**), as proposed in recent studies on food volatile metabolome [20,29]. In any case, chromatographic fingerprints need alignment to correct retention time shifts among samples, an inherent issue in chromatography. This can be done through different techniques, ranging from the most rigid ones, such as correlation optimized shifting (coshift), or less rigid approaches such as extension by intervals (icoshift), to more flexible methods that shift, stretch and/or compress profiles without swapping peaks, such as correlation optimized warping (COW) [37]. Classification models constructed with aligned chromatographic fingerprints have provided successful results for several food authentication problems [10,19,20,29,36,38,39]. Specifically, the fingerprinting approach has proven significantly more efficient than profiling when assessing the origin of virgin olive oil based on SHs data, thanks to its comprehensive use of analytical information [44]. Nevertheless, it is important to note that models developed from fingerprinting approaches may be more susceptible to overfitting than other methods if proper modelling practices and validation techniques are not adhered to.

Chemometrics in food authentication

Despite resulting in more efficient authentication models [29,30], moving from conventional targeted to novel untargeted methods requires multivariate chemometric methods to effectively process and extract the information from complex chromatographic data.

Classification methods are the most prevalent supervised machine learning techniques in food authentication [13], particularly when the goal is to develop screening tools to detect potential fraudulent samples (**Figure 3**). Among these, discriminant and class-modelling techniques are applied for authentication purposes based on the volatile metabolome [36,38]. In discriminant techniques, samples are classified into one of the predefined classes, whereas class-modelling techniques focus on a single class and determine whether samples belong to that class. The main representative of the former group is partial least squares – discriminant analysis (PLS-DA). PLS-DA seeks for maximizing the covariance between the latent variables resulting from variable reduction of a given data matrix and the response variable, which is a dummy variable identifying the different classes – regardless of being a binary or multi-class authentication problem [40,41]. PLS regression is also commonly applied when aiming at a quantitative prediction of a numeric response variable for authentication purposes [13]. Among class-modelling techniques, soft independent modelling of class analogy (SIMCA) is one of the most applied methods [42]. SIMCA models are based on principal component analysis (PCA) for each of the categories. They are often called one-class methods, since they focus on a well-represented category of samples rather than multiple classes. PLS, instead, is generally applied to build multiclass models through PLS-DA. However, the extension one-class PLS could be used for a similar aim than SIMCA, although its use on VOCs has not been reported for authentication so far. On the other hand, supervised non-linear classification methods based on e.g., artificial neural networks (ANN), support vector machine (SVM) or random forest (RF) have also been explored for authentication purposes using the volatile metabolome [43,44]. Model interpretability still remains a challenge for non-linear methods due to their complexity, non-linear relationships, and intricate model structures, which also complicates visualization. Contrarily, linear methods make it easier to identify relevant

variables, even in untargeted approaches. This enables the identification of specific VOCs relevant to a particular class, ensuring the model does not function as a "black box" [10,29].

When developing authentication tools, achieving a high performance in classification becomes a priority. Threshold optimization strategies such as the application of Receiver Operating Curves (ROC) curves might reveal the thresholds leading to optimal sensitivity and specificity, but the definition of the decision criteria may depend on the intended use of the authentication method. For instance, VOC fingerprinting was suggested as a screening method to detect boundary samples in olive oil sensory quality evaluation whose identity would be further confirmed by a sensory panel test [36].

Besides supervised models, previous works also focused on unsupervised machine learning tools such as PCA, usually to explore trends within the analytical data, as well as clustering methods [13]. Nevertheless, the reduced matrix consisting of few principal components is sometimes not enough to explain the variance within different categories in authentication problems, especially when chromatographic fingerprints are used.

Transferability of untargeted authentication tools: a challenge

As previously mentioned, targeted methods are recognised for being robust and straightforward to transfer due to their adherence to well-stablished validation procedures which is essential for any official method [45]. This is the main reason why current control methods rely on targeted approaches. In this context, recent research aimed at validating a targeted method based on VOCs by GC-MS for quality and authenticity assessment of olive oils [24]. Contrarily, untargeted approaches hinder reproducibility among operators and instruments, and validation procedures are still unclear [34]. Some authors proposed method validation strategies for untargeted metabolomic analysis [46], by adapting the target validation procedure to the detected chromatographic peaks in quality control samples. The main challenge in validating untargeted methods emerges in the context classification models using fingerprinting methods, in which the raw analytical signal is considered rather than selecting specific

chromatographic peaks. On top of that, when untargeted methods are based on VOCs by GC-MS, additional challenges may arise due to issues inherent to the analytical technique or the characteristics of the VOCs themselves, which can further complicate the transferability of authentication models. For instance, chromatographic fingerprints might vary because of the GC column or SPME fiber brand and shelf-life. As Riedl et al. suggested [47], system challenge validations are needed to ensure the long-term stability of models including the exchange of measurement data between laboratories. A preliminary in-house validation of fingerprinting for virgin olive oil commercial category verification has yielded promising results [36], but significant work remains to be done to successfully advance towards interlaboratory validations.

Moreover, authentication models should be regularly updated including samples from new batches [47,48]. Continuous statistical modelling strategies have been proposed for some food products using spectroscopic methods [49], but more research is needed to extrapolate these approaches to GC-MS methods and to define validation protocols for untargeted methods to become reference methods. In the meantime, untargeted authentication tools built with in-house databases are promising as guidance for official inspections and conformity checks by the control bodies. In fact, EU member states shall ensure that conformity checks are performed selectively based on a risk analysis, where the untargeted screening tools could easily be included.

Conclusions

The analysis of volatile metabolome using GC-MS in the context of food fraud detection offers a promising tool for ensuring the authenticity and quality of food products (**Figure 3**). Untargeted profiling and fingerprinting emerge as the most comprehensive approaches for understanding food volatilome, showing significant potential as markers for both sensory quality assessment and authentication. Conversely, targeted VOC analysis represents a more readily transferable approach, with higher likelihood of application in official control procedures in the near future. Ongoing research is crucial to fully unlock the potential of these markers and improve the transferability of the methods.

Acknowledgements

INSA-UB Maria de Maeztu Unit of Excellence (Grant CEX2021-001234-M) funded by MICIU/AEI/FEDER, UE. B. Quintanilla-Casas acknowledges the Danish Data Science Academy Postdoctoral fellowship, funded by the Novo Nordisk Foundation (NNF21SA0069429). B. Torres-Cobos thanks the Spanish Ministry of Science, Innovation and Universities for the predoctoral fellowship FPU20/01454.

References

1. European Committee for Standardization (CEN): *Authenticity in the feed and food chain - General principles and basic requirements (CEN/WS 086)*. 2019.

2. The food Integrity project: *FoodIntegrity Handbook - A guide to food authenticity issues and analytical solutions*. Eurofins Analytics France; 2018.

3. Spink J, Moyer DC: **Defining the public health threat of food fraud**. *J Food Sci* 2011, **76**:157–163.

4. The Alert and Cooperation Network: *2022 Annual Report*. Available at: https://food.ec.europa.eu/document/download/499ffcf1-6c99-43ec-8905-5ff3e812eeb2_en?filename=acn_annual-report_2022.pdf

   ** This annual report covers all the information shared in 2022 between members of the Alert and Cooperation Network (ACN) that includes the Rapid Alert System for Food and Feed network (RASFF), the Administrative Assistance and Cooperation network (AAC) or the Agri-Food Fraud Network (FFN). The reported cross-border violations of the European Union agri-food chain legislation concerning non-compliance as well as fraudulent notifications are reported.*

5. Conte L, Bendini A, Valli E, Lucci P, Moret S, Maquet A, Lacoste F, Brereton P, García-González DL, Moreda W, et al.: **Olive oil quality and authenticity: A review of current EU legislation, standards, relevant methods of analyses, their drawbacks and recommendations for the future**. *Trends Food Sci Technol* 2020, **105**:483–493.

6. Ni Q, Amalfitano N, Biasioli F, Gallo L, Tagliapietra F, Bittante G: **Bibliometric review on the volatile organic compounds in meat**. *Foods* 2022, **11**: 3574.

7. Lytou AE, Panagou EZ, Nychas GJE: **Volatilomics for food quality and authentication**. *Curr Opin Food Sci* 2019, **28**:88–95.

   ** This review focus on the significant role of volatilomics in food quality and authenticity assessment, highlighting the technical advantages of the analytical method involved in its measurement, as well as the relevance of multivariate analysis for the success of the volatilomics application in this field.*

8. Ríos-Reina R, Aparicio-Ruiz R, Morales MT, García-González DL: **Contribution of specific volatile markers to green and ripe fruity attributes in extra virgin olive oils studied with three analytical methods**. *Food Chem* 2022, **399**: 133942.

9. Noguera PM, Egiddi M, Södergren J, da Silva MR, Beauchamp J, Petersen MA, Buettner A, Jørgensen NOG: **More than just geosmin and 2-methylisoborneol? Off-flavours associated with recirculating aquaculture systems**. *Rev Aquac* 2024, doi:10.1111/raq.12949.

10. Quintanilla-Casas B, Bustamante J, Guardiola F, García-González DL, Barbieri S, Bendini A, Toschi TG, Vichi S, Tres A: **Virgin olive oil volatile fingerprint and chemometrics: Towards an instrumental screening tool to grade the sensory quality**. *LWT* 2020, **121**:108936.

    ** This study proposes a screening tool based on the volatile untargeted fingerprint of virgin olive oils to assess their sensory quality according to the commercial categories. By reducing the panel's workload, this tool improved the classification accuracy.*

11. Cecchi L, Migliorini M, Mulinacci N: **Virgin olive oil volatile compounds: Composition, sensory characteristics, analytical approaches, quality control, and authentication**. *J Agric Food Chem* 2021, **69**:2013–2040.

    ** This review provides comprehensive information on the use of the volatile metabolome of virgin olive oil for various authentication and quality control purposes, covering different approaches in VOCs assessment, data treatment, and practical application.*

12. Diez-Simon C, Mumm R, Hall RD: **Mass spectrometry-based metabolomics of volatiles as a new tool for understanding aroma and flavour chemistry in processed food products**. *Metabolomics* 2019, **15**, 41.

13. Noshad M, Behbahani BA, Karabagias IK: **Volatilomic with chemometrics: a toward authentication approach for food authenticity control**. *Eur Food Res Technol* 2023, **249**:2215–2226.

    ** Volatilomics systems play a crucial role in evaluating food quality, and chemometrics reveal hidden information in complex instrumental datasets. This review contributes to efforts in authenticating foods and protecting consumers from fraud*

14. Ch R, Chevallier O, McCarron P, McGrath TF, Wu D, Nguyen Doan Duy L, Kapil AP, McBride M, Elliott CT: **Metabolomic fingerprinting of volatile organic compounds for the geographical discrimination of rice samples from China, Vietnam and India**. *Food Chem* 2021, **334**, 127553.

15. Fang S, Ning J, Huang WJ, Zhang G, Deng WW, Zhang Z: **Identification of geographical origin of Keemun black tea based on its volatile composition coupled with multivariate statistical analyses**. *J Sci Food Agric* 2019, **99**:4344–4352.

16. Sousa A, Vareda J, Pereira R, Silva C, Câmara JS, Perestrelo R: **Geographical differentiation of apple ciders based on volatile fingerprint**. *Food Res Int* 2020, **137**: 109550.

17. Tsagkaris AS, Koulis GA, Danezis GP, Martakos I, Dasenaki M, Georgiou CA, Thomaidis NS: **Honey authenticity: analytical techniques, state of the art and challenges**. *RSC Adv* 2021, **11**:11273–11294.

18. Giménez-Campillo C, Arroyo-Manzanares N, Pastor-Belda M, Campillo N, Viñas P: **Discrimination of the geographical origin of peaches by the monitoring of volatile organic compounds by gas chromatography with mass spectrometry and chemometric tools**. *J Food Compos Anal* 2024, **129**, 106125.

19. Quintanilla-Casas B, Torres-Cobos B, Guardiola F, Romero A, Tres A, Vichi S: **Geographical authentication of virgin olive oil by GC-MS sesquiterpene hydrocarbon fingerprint: Scaling down to the verification of PDO compliance**. *Food Contr* 2022, **139**:109055.

20. Quintanilla-Casas B, Torres-Cobos B, Guardiola F, Servili M, Alonso-Salces RM, Valli E, Bendini A, Toschi TG, Vichi S, Tres A: **Geographical authentication of virgin olive oil by GC–MS sesquiterpene hydrocarbon fingerprint: Verifying EU and single country label-declaration**. *Food Chem* 2022, **378**:132104.

21. Ugolini T, Mattagli F, Melani F, Zanoni B, Migliorini M, Trapani S, Giambanelli E, Parenti A, Mulinacci N, Cecchi L: **HS-SPME-GC-MS and chemometrics for the quality control and clustering of monovarietal extra virgin olive oil: A 3-year study on terpenes and pentene dimers of Italian cultivars**. *J Agric Food Chem* 2024, **72**: 11124−11139.

22. May B, Wüst M: **Temporal development of sesquiterpene hydrocarbon profiles of different grape varieties during ripening**. *Flavour Fragr J* 2012, **27**:280–285.

23. Yue Li: **Analytical methods for the analysis of volatile natural products**. *Natural Product Report* 2023, **40**:922–956.

24. García-González DL, Casadei E, Aparicio-Ruiz R, Ortiz Romero C, Valli E, Brereton P, Koidis A, Korytkowska M, Servili M, Selvaggini R, et al.: **Multianalyte analysis of volatile compounds in virgin olive oils using SPME-GC with FID or MS detection: results of an international interlaboratory validation**. *Eur J Lipid Sci Tech* 2023, doi:10.1002/ejlt.202300079.

25. Rivera-Pérez A, Romero-González R, Garrido Frenich A: **Fingerprinting based on gas chromatography-Orbitrap high-resolution mass spectrometry and chemometrics to reveal geographical origin, processing, and volatile markers for thyme authentication**. *Food Chem* 2022, **393**: 1333377

26. Ballin NZ, Laursen KH: **To target or not to target? Definitions and nomenclature for targeted versus non-targeted analytical food authentication**. *Trends Food Sci Technol* 2019, **86**:537–543.

27. Baccolo G, Quintanilla-Casas B, Vichi S, Augustijn D, Bro R: **From untargeted chemical profiling to peak tables – A fully automated AI driven approach to untargeted GC-MS**. *TrAC TrendsiAnal Chem* 2021, **145**:116451.

28. Jimenez-Carvelo AM, Cuadros-Rodríguez L: **Data mining/machine learning methods in foodomics**. *Curr Opin Food Sci* 2021, **37**:76–82.

    *** This paper aims to provide novel definitions and nomenclature for targeted and non-targeted analytical approaches as a first step towards a harmonization, that is particularly necessary for non-targeted approaches together with the validation schemes*

29. Quintanilla-Casas B, Bertin S, Leik K, Bustamante J, Guardiola F, Valli E, Bendini A, Gallina Toschi T, Tres A, Vichi S: **Profiling versus fingerprinting analysis of sesquiterpene hydrocarbons for the geographical authentication of extra virgin olive oils**. *Food Chem* 2020, **307**:125556.

    * *This work proved the usefulness of sesquiterpene hydrocarbons as geographical authentication markers of extra virgin olive oils, measured by HS-SPME-GC-MS. Here,*

targeted profiling and untargeted fingerprinting approaches are compared, resulting the classification efficiency of the model based on the fingerprint significantly higher.

30. Xu L, Xu Z, Liao X: **A review of fruit juice authenticity assessments: Targeted and untargeted analyses**. *Crit Rev Food Sci Nutr* 2022, **62**:6081–6102.

31. Quintanilla-Casas B, Bro R, Hinrich JL, Davie-Martin CL: **Tutorial on PARADISe: PARAFAC2-based Deconvolution and Identication System for processing GC-MS data.** 19 January 2023, PROTOCOL (Version 1) available at Protocol Exchange [https://doi.org/10.21203/rs.3.pex-2143/v1]

32. Quelal-Vásconez MA, Macchioni R, Livi G, Pérez-Esteve É, Lerma-García MJ, Talens P, Barat JM, Petersen MA, Bro R: **Automatic and non-targeted analysis of the volatile profile of natural and alkalized cocoa powders using SBSE-GC-MS and chemometrics**. *Food Chem* 2022, **389**:133074.

33. Lacalle-Bergeron L, Portolés T, Sales C, Carmen Corell M, Domínguez F, Beltrán J, Vicente Sancho J, Hernández F: **Gas chromatography-mass spectrometry based untargeted volatolomics for smoked seafood classification**. *Food Res Int* 2020, **137**, 109698.

34. Medina S, Pereira JA, Silva P, Perestrelo R, Câmara JS: **Food fingerprints - A valuable tool to monitor food authenticity and safety**. *Food Chem* 2019, **278**:144–162.

35. Cuadros-Rodríguez L, Ruiz-Samblás C, Valverde-Som L, Pérez-Castaño E, González-Casado A: **Chromatographic fingerprinting: An innovative approach for food "identitation" and food authentication - A tutorial**. *Anal Chim Acta* 2016, **909**:9–23.

36. Quintanilla-Casas B, Marin M, Guardiola F, García-González DL, Barbieri S, Bendini A, Toschi TG, Vichi S, Tres A: **Supporting the sensory panel to grade virgin olive oils: an in-house-validated screening tool by volatile fingerprinting and chemometrics**. *Foods* 2020, **9**:1509.

37. Bloemberg TG, Gerretzen J, Lunshof A, Wehrens R, Buydens LMC: **Tutorial Warping methods for spectroscopic and chromatographic signal alignment: A tutorial**. *Anal Chim Acta* 2013, **781**:14–32.

38. Vera DN, Jiménez-Carvelo AM, Cuadros-Rodríguez L, Ruisánchez I, Callao MP: **Authentication of the geographical origin of extra-virgin olive oil of the Arbequina cultivar by chromatographic fingerprinting and chemometrics**. *Talanta* 2019, **203**:194–202.

39. Stilo F, Jiménez-Carvelo AM, Liberto E, Bicchi C, Reichenbach SE, Cuadros-Rodríguez L, Cordero C: **Chromatographic fingerprinting enables effective discrimination and identitation of high-quality Italian extra-virgin olive oils**. *J Agric Food Chem* 2021, **69**:8889.

40. Tres A, Ruiz-Samblas C, Van Der Veer G, Van Ruth SM: **Geographical provenance of palm oil by fatty acid and volatile compound fingerprinting techniques**. *Food Chem* 2013, **137**:142–150.

41. Pranata AW, Yuliana ND, Amalia L, Darmawan N: **Volatilomics for halal and non-halal meatball authentication using solid-phase microextraction–gas chromatography–mass spectrometry**. *Arab J Chem* 2021, **14**: 103146.

42. Oliveri P: **Class-modelling in food analytical chemistry: Development, sampling, optimisation and validation issues - A tutorial**. *Anal Chim Acta* 2017, **982**:9–19.

43. Deng X, Liu Z, Zhan Y, Ni K, Zhang Y, Ma W, Shao S, Lv X, Yuan Y, Rogers KM: **Predictive geographical authentication of green tea with protected designation of origin using a random forest model**. *Food Contr* 2020, **107**:106807.

44. Aurum FS, Imaizumi T, Thammawong M, Suhandy D, Zaman MZ, Purwanto E, Praseptiangga D, Nakano K: **Predicting Indonesian coffee origins using untargeted SPME – GCMS - based volatile compounds fingerprinting and machine learning approaches**. *Eur Food Res Technol* 2023, **249**:2137–2149.

45. International Organization for Standardization (ISO): *Accuracy (Trueness and Precision) of Measurement Methods and Results (ISO 5725)*. 2019.

46. Sangster T, Major H, Plumb R, Wilson AJ, Wilson ID: **A pragmatic and readily implemented quality control strategy for HPLC-MS and GC-MS-based metabonomic analysis**. *Analyst* 2006, **131**:1075–1078.

47. Riedl J, Esslinger S, Fauhl-hassek C: **Review of validation and reporting of non-targeted fingerprinting approaches for food authentication**. *Anal Chim Acta* 2015, **885**:17–32.

48. McGrath TF, Haughey SA, Patterson J, Fauhl-Hassek C, Donarski J, Alewijn M, van Ruth S, Elliott CT: **What are the scientific challenges in moving from targeted to non-targeted methods for food fraud testing and how can they be addressed? – Spectroscopy case study**. *Trends Food Sci Technol* 2018, **76**:38–55.

49. Georgouli K, Martinez Del Rincon J, Koidis A: **Continuous statistical modelling for rapid detection of adulteration of extra virgin olive oil using mid infrared and Raman spectroscopic data**. *Food Chem* 2017, **217**:735–742.

1.4.4 Data treatment and chemometric methods

While data obtained from targeted approaches, such as the isotopic ratios of specific elements, are usually treated applying univariate statistic (t-test, one-way analysis of variance (ANOVA), among others), the great amount of data obtained from untargeted approaches requires multivariate statistical tools. Contrarily to univariate methods that focus solely on a single variable, multivariate methods use all variables simultaneously and deal with the relationship between them [186], allowing to efficiently extract the maximum information from data [187].

As summarized in Publication 1, the state-of-the-art in food authentication is the application of chemometric multivariate tools to reduce the data dimensionality keeping all the relevant information [188,Publication 1]. These chemometric tools can be divided in two categories: supervised and unsupervised methods [189].

Unsupervised methods are primarily employed for exploratory analysis since **no prior information is used to guide the model**. These methods help visualize the data and highlight differences and similarities among samples. The most common unsupervised method in food authentication is principal component analysis (PCA) [188]. PCA reduces the dimensionality of the data by generating a smaller number of new variables, called principal components (PCs), which are linear combinations of the original variables. The PCs are orthogonal to each other to avoid redundant information, thereby maximizing the explained variance of the original data [190-192].

Supervised methods use the information of a response variable (y) to build models. This response variable can be quantitative, in regression models, or qualitative/categorical in classification models. Among supervised approaches, classification methods are the most widely employed in food authentication [188,193], especially in the case of screening tools for food fraud detection. From classification approaches, two main techniques arise, discriminant analysis and class-modelling [187,189,193,Publication 1]. In discriminant techniques, all the categories are concerned in the discrimination models, hence, samples are classified into these predefined categories, whereas class-modelling techniques focus solely on a single class to ascertain whether samples belong to that class or not. The primary examples

of each group are **partial least squares – discriminant analysis (PLS-DA)** and soft independent modelling of class analogy (SIMCA) [187,188,193,194,Publication 1].

PLS-DA aims to **maximize the covariance between** the **latent variables (LVs)**, derived from the variable reduction of the original data matrix (X), **and the response variable** (y), which is a dummy variable identifying the different categories [187,189]. PLS-DA finds the maximum separation between classes by minimising the influence of common features, and increasing the influence of the, even subtle, differences [195,Publication 1].

SIMCA models are based on PCA. PCs, which are defined as the directions of maximum variance in the multivariate data space, are computed for each of the categories. Each SIMCA model focus solely on a well-represented class of the samples, hence the significant PCs define the SIMCA inner space of that class and samples to be predicted are projected on that space, obtaining the corresponding score values. Depending on the score value obtained the sample is classified into that category or not [194].

1.4.5 Challenges of the analytical approaches for food authentication

As previously stated in section 1.4.1 and in Publication 1, targeted methods are acknowledged for their robustness and straightforward transferability, making them the predominant approach in current control methods, as validation is essential for an analytical method to become official. Isotopic methods exemplify this, providing high precision and reproducibility and enabling bias correction through the use of matrix-matched certified reference materials (CRMs). This ensures highly accurate results and facilitates their transferability across different laboratories, making them suitable for use even in legal cases, as enforcement of official control methods [84]. However, because they rely on only a few targeted values, the sample information is limited and influenced by a restricted number of climatic and geological factors. Therefore, their applicability is constrained when samples originate from geographical regions with similar climatic conditions or soil characteristics [86,90,197,198]. Additionally, they require large databases that need to be annually updated due to the high variability in isotope ratios from year to year [84,97,196].

In untargeted methods, the transferability is hindered by the absence of defined validation procedures and by the challenges of reproducibility across different operators and instruments. While some performance parameters of the standard validation methods can be applied to untargeted methods, others are either unavailable or inapplicable [81,172-174]. Traditional validation assessment techniques, such as linearity, standard deviation, limit of detection, and limit of quantification rely on univariate statistics and cannot be applied to pattern recognition methods designed for developing classification models for categorical predictions. This issue is particularly evident in in approaches that use class discrimination fingerprinting where the entire instrumental output is analysed instead of selecting specific chromatographic peaks [199,Publication 1]. In a recent publication of Bayen et al. [81], the authors propose that the sensitivity and specificity rates of the chemometric model (based detection probabilities) should be evaluated to validate untargeted methods. Additionally, they highlight the importance of transparency in databases, encouraging their disclosure, when possible, to allow for careful examination. On top of that, a crucial aspect of validating untargeted methods is the proper assessment of chemometric models applying system challenge strategies. These strategies involve testing the method's performances under different instrumental and analytical conditions or sample characteristics from those used in model development to ensure its long-term stability [200]. Finally, models should be regularly updated including new samples representative of the natural variability, to guarantee the robustness of models for future analytical requirements [47,174,Publication 1].

1.5 Geographical and varietal authentication of virgin olive oil and nuts: state-of-the-art and current gaps

Currently, there are no reference analytical methods to authenticate the botanical and geographical origin of foods, including VOO and nuts. As previously discussed, documental traceability and morphological examination of nuts have several limitations that reduce the effectiveness of controls. Literature reports various approaches to address these issues, applying a wide range of techniques and analysing different classes of markers by distinct approaches. Despite several significant

advancements, various gaps continue to hinder the effective resolution of geographical and varietal authentication for VOO and nuts.

1.5.1 Virgin olive oil

A substantial amount of research has been dedicated to develop efficient and reliable methods for authenticating the provenance and cultivar of VOO. These methods employ various marker classes, analytical techniques, and approaches, as recently reviewed [58,201-203].

Geographical authentication of VOO has been pursued by applying **isotopic methods** [90,204], focusing primarily on the light bioelements C, H, and O, which have been the most extensively studied for this purpose [90]. Bulk isotope analysis has been used to characterise Italian, Greek, and Turkish VOO production [97], revealing significant regional differences within these countries. However, in the majority of cases, bulk isotope analysis was combined with CSIA or used in combination with other techniques and markers such as mineral profiles or metabolite data [101,104,205,206]. This suggests that bulk isotope analysis alone may not always be sufficient to accurately verify the geographical origin of VOO. Moreover, VOOs from different geographical regions may exhibit indistinguishable isotopic ratios if the climate conditions of those regions are similar, which can hinder their differentiation [197].

In Italy, an isotopic databank for the Italian PDO and PGI extra virgin olive oils (EVOOs), has been created by the Ministry of Agriculture, Food and Forestry, and is annually updated with samples from several GI. With this system, C, H, O bulk isotope ratios of the collected VOO samples are measured, and authenticity limits are set for different Italian regions. When a test sample from the market fails to meet reference limits, traceability and verification procedures are conducted. If the paperwork audit confirms the analytical suspicions, a legal procedure is then initiated [84].

Among methods based on **metabolic markers**, a VOO geographical authentication method utilizing the SH fingerprint stands out for its exceptional performance, as recently reported [139,140]. SH fingerprint obtained by HS-SPME-GC-MS and combined with PLS-DA allowed to i) successfully discriminate between EU and non-EU VOOs and classify them according to their country of origin; and ii) efficiently distinguish between

VOOs of closely located PDOs and from other VOOs not belonging to PDOs. These results were obtained with PLS-DA classification models constructed using large sample sets (n = 350-400) including multiple producing regions, olive cultivars and harvest years, covering a high natural variability and being representative of the real market scenario. The methodology yielded a high classification accuracy in external validation, with >89% and >93% of correct classification for the EU and PDO models respectively [139,140]. The results demonstrated the effectiveness of the methodology as a tool for verifying EU/non-EU and country of origin label declarations in VOO, as well as for ensuring its compliance with a specific GI quality scheme. However, the reproducibility and transferability of this method is limited due to the absence of validation procedures and strategies for untargeted fingerprinting-based methods [200] as reported in previous sections. It is crucial to assess whether the expected high efficiency of fingerprinting compensates for its potentially lower reproducibility compared to isotopic analysis.

So far, **no systematic comparison has been made between targeted isotopic method and emerging metabolic fingerprinting for VOO geographical origin authentication** (**Gap 1**). The results obtained by both approaches are not easily comparable, underscoring the need for a systematic comparison under standardized conditions. This is needed to objectively evaluate and contrast their performance in authenticating VOO origin and provide insight into their strengths and weaknesses as well as to identify areas for further advancement in these approaches.

Despite the great influence of **olive cultivar** in the qualitative and sensory characteristics of VOO, the required use of specific traditional olive cultivars in GIs [61], and the fact that VOOs with cultivar label claims are highly appreciated by consumers [15,16], up to date, there are no reference methods to verify VOO cultivar. The varietal authentication of VOO has also been widely addressed by the research community by focusing on several major and minor compounds and by applying multiple analytical techniques and chemometric approaches [38,160,201]. However, except for DNA-based methods [207-211], which are too costly for routine analysis [201], reliable **metabolic markers** for VOO cultivar authentication remain unavailable. This is mainly due to the fact that, although VOO composition is significantly influenced

environmental conditions, most studies focus on monovarietal VOOs from the same specific region. Likewise, the effect of other external factors, such as maturation, phytosanitary status, harvest year, and processing and storage conditions [38,130,143], is seldom considered. While these studies emphasise the impact of genetic factors on VOO characteristics, they often tend to overlook broader environmental and technological variables, making their findings less representative of real VOO production. Consequently, at present, **there are no official analytical methods to authenticate OO cultivars**, and existing studies in this area still need further evaluation to accurately reflect the real-world olive oil production [38,49-130,160,201]. As commented in section 1.3.4, SHs in VOO are not only affected by the geographical origin, but also by the olive cultivar [212-214]. However, **SH fingerprinting has not yet been evaluated for VOO varietal authentication beyond geographical, agronomic and processing conditions** (**Gap 2**).

1.5.2 Hazelnut

As previously exposed, direct visual inspection and morphological evaluations based on physical characteristics are the standard techniques to verify hazelnut authenticity in addition to documentary controls [70]. Nonetheless, these methods are limited in efficiency as they are susceptible to external influences and can only be applied to whole kernels [45,77]. Consequently, numerous studies have been conducted over the past years to explore more suitable markers for hazelnut authentication [77,215-223].

In this sense, although **isotopic markers** represent a widely recognised tool to determine the **geographical origin** of food products [23,83-87], and they have been explored for the authentication of pistachios [224], walnuts [219] and peanuts [225], they have been scarcely applied for hazelnut geographical authentication. To date, only Sammarco et al. (2023) used bulk $\delta^{13}C$ and $\delta^{18}O$ values to authenticate hazelnut origin [226]. Despite observing significant mean differences in these stable isotopes among samples from different provenances, their study found that combining isotopic values with mineral profile was necessary to achieve acceptable discrimination of hazelnuts according to their origin.

As previously described, the isotope ratio of various elements is associated to distinct environmental and agricultural factors, each of which can play a crucial role in distinguishing hazelnuts from different origins. For this reason, to fully exploit the authentication potential of isotopic markers, it is necessary to comprehensively explore light bioelements both in bulk samples and in specific compounds. CSIA in proteins, lipids or carbohydrates may offer deeper insights into isotopic data, as demonstrated for several food products [101,88,103,105,106]. In the case of hazelnuts, in which lipids account for the main percentage of the kernel [44,227], the study of the isotopic composition of the FAs could provide additional and complementary information to the bulk isotopic markers. Additionally, the ability of heavy geoelements to discriminate between origins should be investigated. Bearing all this information in mind, **a gap exists in exploring and properly assessing both bulk and compound-specific isotopic markers, as well as heavy geoelements, for authenticating hazelnut provenance**, evaluating its efficacy in comparison with other methods (**Gap 3**).

Although isotopes could be suitable markers for geographical origin authentication, they cannot be used for varietal discrimination. Regarding DNA-based methods, despite the high accuracy provided [220], their complex and costly procedures make them unsuitable for routine analysis. In turn, both phenotypic and DNA-based methods are limited to varietal authentication only and cannot be explored to assess hazelnut provenance. In this context, an ideal solution to address the main hazelnut authentication issues would be a marker capable of simultaneously verifying both the **origin and cultivar** of hazelnuts in a single analysis.

This could be achieved through methods based on **metabolic markers**. Different analytical methods based on metabolic markers have been tested to authenticate hazelnuts [77,153,215,216,218,220-226]. However, the majority of metabolomic studies to date have primarily focused on distinguishing either the cultivar or the geographical provenance of hazelnuts [77,215,216,218,222,228-230]. Only a limited number of works have attempted to authenticate both the botanical and geographical origins simultaneously [153], and none has successfully demonstrated the ability to verify them independently of one another. Moreover, many existing studies have significant limitations; some of them only differentiate between groups of cultivars

[221, 231], and others focus exclusively on a single PDO or country of origin [216,221,222]. Additionally, some are tested on small and limited sample sets [222,229] or lack proper validation of their classification models [77,153,222,228-230].

This gap in research highlights the insufficiency of current metabolic methods in providing comprehensive varietal and geographical authentication, which is crucial for ensuring the integrity and authenticity of hazelnuts in the market. To address this gap, metabolic markers influenced by both environmental conditions and genetic factors, while minimally affected by storage or processing, should be explored. Their efficiency should be tested in challenging scenarios, such as verifying the geographical origin of the same cultivar grown in multiple geographical areas and determining the varietal identity of hazelnuts from different cultivars produced in the same region.

In this regard, previous studies have shown that terpenoids in plants are influenced by genetic and pedoclimatic conditions [143,168,214]. These compounds have successfully been used to authenticate several food products [113,136-141,232]. Additionally, volatile terpene compounds, including monoterpenes and SH have been identified in hazelnut oil and in crude and roasted hazelnuts [233-236]. However, no studies have yet explored their potential as hazelnut botanical and geographical markers. Therefore, analysing and evaluating these markers in hazelnuts could provide valuable insights and potentially result in an effective tool for authenticating hazelnuts.

Another promising approach to addressing this challenge is the analysis of the UF of hazelnuts. As previously mentioned (section 1.3.4), the UF is a lipidic fraction rich in secondary metabolites that satisfies all key requirements for a reliable marker of both varietal and geographical authenticity. Moreover, its analysis can be efficiently performed through GC–MS, which is a highly sensitive and widely accessible technique commonly available in control laboratories.

In addition to UF analysis, the TAG profile of hazelnuts offers another promising method for hazelnut authentication. As primary components of the hazelnut lipid fraction, TAGs vary in response to both genetic and environmental factors [44,133]. The analysis of TAGs provides advantages such as simplicity, speed, and minimal sample preparation, making it an ideal option for routine analysis or as a preliminary screening

tool. However, despite its potential, TAG fingerprinting remains underexplored, with only a few studies applying this method to develop geographical classification models [228]. To date, no studies have applied TAG analysis to differentiate hazelnut cultivars, further underscoring the need for its exploration.

Regarding spectroscopic techniques such as MIR and NIR spectroscopy, that respond to multiple chemical compounds [164,237], they have been previously applied to verify the botanical and geographical origin of various nuts with excellent results [221,223,229,230,238-242] and have shown promise in distinguishing Italian hazelnuts from those of other origins [230], their potential to simultaneously authenticate hazelnut cultivar and provenance remains unexplored.

Therefore, given the **absence of effective tools for verifying the cultivar and provenance of hazelnuts,** there is a pressing need for efficient, fast and robust methods to authenticate hazelnuts and protect consumers interests. In this context, although **metabolic markers** such as **terpenes, UF, TAGs, and spectroscopic fingerprints**, show great potential due to their **dependence on pedoclimatic conditions and genetics**, **they have not yet been fully assessed as tools for simultaneously authenticating the varietal and geographical origin of hazelnuts** (**Gap 4**)**.**

1.5.3 Pine nut

In addition to conventional methods for verifying pine nut species through phenotypic observations of physical characteristics, the limitations of which are discussed in section 1.2.3, alternative methods based on various markers and analytical approaches have been investigated [53,55,56,78,243-245]. Yet, these approaches require further validation using diverse sample sets that encompass the breadth of pine nut diversity, including different producers, harvest years, origins, and species. A comprehensive evaluation using large, representative sample sets reflecting the actual market scenario is still required. Therefore, there remains a critical need to develop efficient and robust pine nut authentication methods suitable for routine control applications.

Although some **isotopic data** have been reported to describe the heterogeneity of North American pine nuts [246], **no studies to date have applied isotopic markers to**

authenticate the origin of pine nuts (**Gap 5**). Given the extensive use of these markers in geographical traceability of food products [23,88-91,247] and their strong correlation with the pedoclimatic and geological conditions of the growing region, exploring their efficacy in distinguishing among pine nuts from different provenances could enhance both pine nut geographical and botanical discrimination. In fact, pine nut species are closely linked to their geographical origins. In this regard, bearing in mind that pine nuts are not a farmed crop but collected from natural forests, where no fertilizers are used, and considering previous research on other pine nut species showing that their spatial distribution is mainly driven by edaphic factors [246], isotopic ratios closely linked to geological parameters, such as Sr, [92-94,247] are a promising option for pine nut authentication. Consequently, **a gap exists in thoroughly investigating the potential of Sr isotopes as an authenticity marker for determining the geographical origin of pine nuts to develop reliable tools for accurately verifying pine nut label declared provenance.** This geological marker should be assessed using a representative sample set that reflects the actual conditions of the pine nut production market (**Gap 5**).

Regarding the assessment of pine nut species, some authors have proposed genetic markers. However, these efforts have primarily focused on identifying the inedible *Pinus armandii* species in food products, which is responsible for PMS [248-251]. While these studies have been successful in identifying various pine nut species, they often include very few, if any, samples of *P. pinea*, which is the species more susceptible to fraud. Additionally, certain species could not be accurately distinguished due to their genetic similarities, while in some cases, the same species exhibited genomic differences, leading to its classification into two separate groups [248,249]. This inconsistency may be due to the insufficient characterization of genomic differences among certain pine nut species in the existing literature [248]. Although DNA-based methods are reliable and accurate, with the ability to identify various pine nut species even in mixtures, they are costly, labour-intensive, destructive, and require extensive databases and specialized technicians [78,245].

With regard to methods based on **metabolic markers**, FAs have been extensively explored in the literature for the authentication of pine nut species [53,56,78,80,243].

Research by Wolff et al. [252] demonstrated that the FA profiles of various *Pinus* species exhibit species-specific differences that can be used for differentiation. Specifically, seven species-specific FAs have been identified that enable the calculation of a diagnostic index (DI), which distinguishes between different *Pinus* species [80]. This DI is derived from the ratio between the main Δ5-unsaturated polymethylene-interrupted FAs (Δ5-UPIFAs) and their metabolic precursors [78,80,243]. However, while the DI based on FA profiles shows potential for species discrimination, its effectiveness diminishes when dealing with pine nut species with similar FA composition, being unable to clearly differentiate between them. Some studies have found that DI values can be inconclusive when analysing commercial pine nuts limiting the identification of species [78,80].

Furthermore, most studies evaluating these metabolic markers have relied on small sample sets, often fewer than 50 samples, which limits their robustness and generalizability [53,56,78,80,243]. Consequently, the ability of these methods to build accurate classification models for pine nut species and origin discrimination remains unproven. The overlapping DI values between species underscore the need for additional diagnostic parameters to support the FA profile data.

As commented above, the suitability of volatile and semi-volatile terpene compounds, such as monoterpenes and SH, as authenticity markers for plants has been demonstrated in several food products [113,136-141,232]. Moreover, conifers and their fruits present considerable amounts of volatile and semi-volatile terpene metabolites [253,254]. Products derived from pine trees, including wood, needles, cones, and essential oils, as well as the volatile fraction of raw pine nut kernels, predominantly contain monoterpenes as their main constituents [253-255]. Furthermore, previous studies have reported variations in the monoterpene and SH composition among different pine species and origins [253-255], suggesting their potential as valuable markers for authenticating pine nut species and their geographical origins.

Despite their proven success as authentication markers in numerous plant-based food commodities, monoterpene and SH compounds have still not been evaluated as potential tools for authenticating pine nuts species and geographical origin (**Gap 6**).

CHAPTER 2.

HYPOTHESES AND AIMS

The primary objective of this doctoral thesis is to contribute to **ensure the varietal and geographical authenticity** of **high-value food** products, specifically **VOO**, **hazelnuts**, and **pine nuts**, which are highly **susceptible to fraudulent practices**, through:

- The development of robust and reliable analytical techniques and their application to a large number of samples to generate highly informative data on food authenticity.
- The implementation of chemometric techniques to identify new markers and develop efficient, robust authentication models.

To achieve this overarching goal, the main gaps in the authenticity of OO and nuts have been identified, hypotheses have been formulated based on existing literature, and specific objectives have been set to address these gaps.

2.1 Virgin olive oil geographical and varietal authenticity

Gap 1: At present, no systematic comparison has been made between targeted isotopic method and emerging metabolic fingerprinting for VOO geographical origin authentication.

Hypothesis 1: The metabolic SH fingerprinting method will outperform isotopic methods in terms of prediction accuracy for authenticating VOO provenance, as it will provide more relevant analytical information without the need of cumbersome identification and quantification procedures, potentially emerging as a more suitable screening tool to support and improve official controls.

***Objective 1**: To **systematically compare** and **assess isotopic analysis and SH fingerprinting** to **develop classification models to discriminate VOO**, using the same sample set with high natural variability including VOO samples from different harvest years, countries of origin, cultivars and regions.*

Gap 2: Currently, no official analytical methods exist for authenticating VOO cultivars. While SH fingerprinting shows promise for verifying the olive cultivar of VOO, it has not yet been evaluated for VOO varietal authentication beyond geographical, agronomic and processing conditions.

Hypothesis 2: Given the successful application of SH fingerprint for authenticating VOO geographical origin beyond technological and storage factors [139,140,168],

and considering that SH composition in VOO is also driven by genetics, it is hypothesized that if VOO cultivar is chosen as the supervising variable in the discriminant analysis, the resulting model will be able to identify cultivar-specific features within the SH fingerprint, thereby accurately predicting the botanical origin of VOO regardless of provenance, and agronomic and processing conditions.

***Objective 2**: To develop and validate **VOO varietal authentication models** based on the **SH fingerprint** obtained by HS-SPME-GC–MS and discriminant techniques, using a large sample set representative of the actual VOO production market.*

2.2 Hazelnut geographical and varietal authenticity

Gap 3: Isotope ratios have been suggested as reliable markers for verifying the geographical origin of various foods. However, a comprehensive exploration and thorough assessment of both bulk and compound-specific isotopic markers, along with heavy geoelements, for authenticating hazelnut provenance is still lacking.

Hypothesis 3: Since bulk stable isotope ratios (δ^{2}H, δ^{13}C, δ^{15}N, δ^{18}O, δ^{34}S, ^{87}Sr/^{86}Sr) and compound-specific ratios (δ^{2}H and δ^{13}C of FAMEs) reflect the pedoclimatic and geological conditions of hazelnut growing areas, they will be suitable markers for hazelnut geographical authentication. Integrating these isotopic data with chemometric tools will enable the identification of the most promising markers and the construction of robust discriminant models for hazelnut provenance.

***Objective 3:** To develop a **hazelnut geographical authentication method** based on the analysis of **stable isotope ratios** (both bulk and compound-specific) to create and validate **discriminant models that differentiate hazelnuts from various producing regions**.*

Gap 4: There is an absence of effective tools for verifying the cultivar and provenance of hazelnuts. Although metabolic markers (terpenes, UF, TAGs, and spectroscopic fingerprints), show great potential due to their dependence on pedoclimatic conditions and plant genetics, they have not yet been fully assessed as tools for simultaneously authenticating the varietal and geographical origin of hazelnuts.

Hypothesis 4.1: As certain hazelnut metabolites are predominantly shaped by genetic and environmental factors, with minimal influence from other variables, their GC-MS untargeted analysis will provide comprehensive information to enable the development of robust classification models for simultaneously authenticating both the botanical and geographical origin of hazelnuts On the other hand, spectroscopic techniques will be a rapid and efficient way to capture variations in the hazelnut metabolome related to the cultivar and geographical origin.

***Objective 4.1**: To develop **hazelnut varietal and geographical authentication methods** based on **metabolic markers (SH, UF, TAG** and **spectroscopic fingerprints**) that are efficient, robust, and fast **to distinguish among hazelnuts from different cultivars and various producing countries**.*

Hypothesis 4.2: Metabolic markers of the hazelnut lipidic fraction assessed by untargeted fingerprinting methods will outperform isotopic targeted methods as authentication tools for geographical origin as they will provide more detailed information for the verification of hazelnut origin.

***Objective 4.2**: **To systematically compare metabolic and isotopic hazelnut authentication methods** in terms of prediction accuracy and analytical efficiency, using a large sample set with high natural variability.*

2.3 Pine nut geographical and varietal authenticity

Gap 5: Although stable isotopes in plants have been linked to the soil type and climate conditions, no studies to date have applied isotopic markers to authenticate the origin of pine nuts.

Hypothesis 5: As pine nuts are harvested from natural forests without the influence of fertilizers on geological markers such as Sr ratios, this isotopic parameter will be effective to authenticate pine nut origin by leveraging its strong correlation with geological conditions of the growing region.

***Objective 5**: To develop a **pine nut geographical authentication method** based on the analysis of the **$^{87}Sr/^{86}Sr$ ratio** to **differentiate Spanish pine nuts from** those originating in major **Asian** producing countries.*

Gap 6: Despite the proven success of monoterpene and SH compounds as authentication markers in numerous plant-based food commodities, these compounds have still not been evaluated as potential tools for authenticating pine nuts species and geographical origin.

Hypothesis 6: Given the successful application of SH fingerprinting for authenticating the geographical and botanical origins of other foods, and the reported influence of pedoclimatic and genetic factors on monoterpenes and SH in conifer species and their fruits, these volatile and semi-volatile terpenes will be reliably used for authentication when analysed using an untargeted approach combined with chemometric tools. They will offer comprehensive information and overcome challenges associated to terpene identification and quantification.

***Objective 6**: To develop and validate **geographical and varietal authentication models** based on the **SH fingerprint** obtained through HS-SPME-GC–MS, as a screening tool to **authenticate Mediterranean P. pinea** pine nuts and **distinguish** them **from pine nuts of other species and countries**.*

CHAPTER 3.

EXPERIMENTAL DESIGN AND METHODOLOGY

To achieve the primary and specific objectives of the present thesis, independent studies with their specific experimental designs were developed for each food product: VOO, hazelnuts and pine nuts. This chapter outlines the overall experimental design (**Figure 6**), while **chapters 4**, **5**, and **6** provide detailed information about the design, methodology, and results of the studies conducted to achieve each specific objective.

Each of these independent studies required sample sets tailored to their specific objectives and detailed in the corresponding chapters. However, all the experimental designs in this thesis fulfil **common requirements** to ensure that sample sets are representative and valid for food authentication purposes [174].

The **samples** used for model development and validation were both **authentic** and **traceable**. These samples were sourced from reliable suppliers: VOOs were obtained in the framework of the OLEUM European project, the AUTENFOOD regional project, and the scientific collaboration with the University of Perugia (Italy), while hazelnuts and pine nuts were obtained in the frame of the TRACENUTS national project.

The **number of samples** included in each sample set was established **to adequately cover the variability** of the food products studied. In food authentication, especially with fingerprinting approaches, accounting for samples' natural variability is critical to generate robust models. However, this requirement considerably increases the workload. To manage this, a model enrichment approach was followed during model development. Initially, a smaller sample set was collected to build a proof-of-concept (PoC) model. If the PoC model yielded promising results, the sample set was further expanded to include more samples, allowing the model to be developed and tested against a greater natural variability. The **broader** the **natural variability within the sample set**, the **more reliable** the **model** becomes.

All sample sets were designed to capture the natural variability of each food products as accurately as possible. This was achieved by incorporating samples from multiple harvest seasons, geographical regions, producers, and cultivars or species.

Regarding the methodology, some studies applied preexisting analytical methods. However, when no previous analytical methods were suitable for the specific purpose, the development of the authentication models was preceded by an initial phase

dedicated to develop new methodologies. This included optimizing the sample treatment and analytical conditions, as well as the data analysis approach and process.

The first part of the experimental design, addresses the **geographical and varietal authentication of VOO**, employing two distinct approaches and experimental designs to achieve specific **objectives 1** and **2** (**Chapter 4**).

The first study (Publication 2) systematically compared and evaluated SH fingerprint and $\delta^{13}C$, $\delta^{2}H$, $\delta^{18}O$ isotopic analysis to verify Italian VOO **geographical authenticity**. To enable a direct comparison between these two methodologies, a same sample set (n = 393) of traceable VOOs from four different campaigns (2016 - 2020), six countries of origin, and multiple cultivars, was analysed by both methods (**Figure 6**). In both cases, preexisting analytical methods were applied [84,97,139,140]. Two discrimination approaches were followed: one aimed at distinguishing Italian VOOs from those of other countries, while another focused on differentiating olive oils from three regions within Italy (Apulia, Calabria, and Sicily). Individual PLS-DA classification models were built and externally validated for each method (Publication 2).

The second study (Publication 3) involved developing and validating authentication models to verify **VOO cultivar** based on the SH fingerprint obtained through HS-SPME-GC–MS [256]. Models were built and externally validated using a large sample set (n = 404) of traceable VOOs from different cultivars (**Figure 6**). The samples were produced under real processing conditions across four harvest seasons (2015-2019) in various EU and non-EU countries and regions. The primary focus of these models was to distinguish 'Arbequina' VOOs from other VOOs (Publication 3).

Regarding the **objectives 3 and 4**, which focus on **developing** and **comparing authentication methods for hazelnut varietal and geographical origins** (**Chapter 5**), the experimental design consisted of a same set of 309 traceable raw hazelnut samples, obtained directly from producers between 2019 and 2022, that was used to allow direct comparison between different isotopic and metabolic approaches. Among these samples, 207 belonged to the ‘Tonda di Giffoni’ (TG) cultivar from various origins, while 102 samples were from different cultivars (non-TG) produced in Spain.

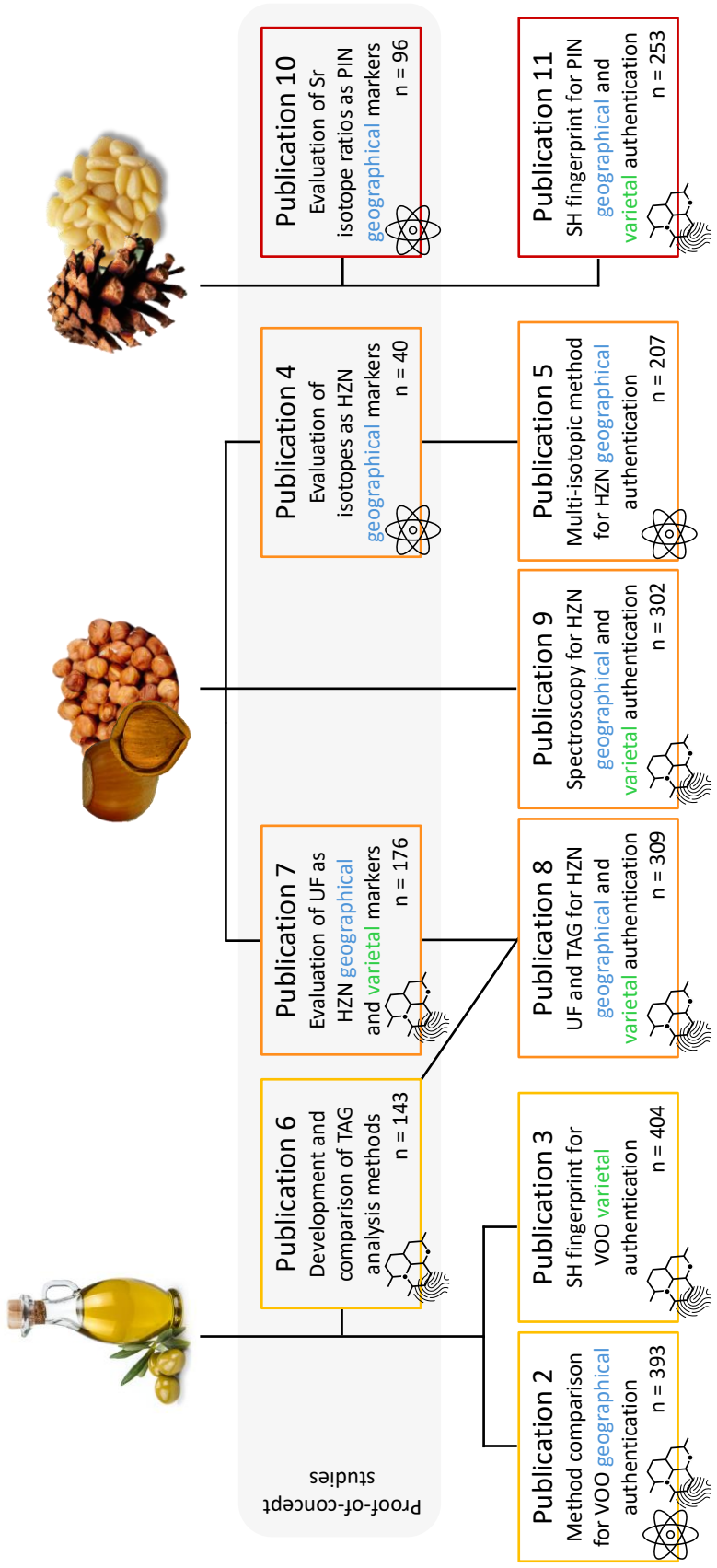


Figure 6. Scheme of the experimental design to fulfil the different authentication objectives. VOO: virgin olive oil, SH: sesquiterpene hydrocarbon, HZN: hazelnut, PIN: pine nut, UF: unsaponifiable fraction, TAG: triacylglycerols

To achieve **objective 3**, which focuses on developing and validating a geographical authentication method for hazelnuts using **stable isotope ratio analysis**, an initial PoC study was conducted (**Figure 6**) to optimize the analytical procedures and to identify the most suitable isotopic markers in verifying hazelnut provenance (Publication 4) for their further implementation in a multi-isotopic approach (Publication 5). First, an analytical method was developed for the hazelnut isotopic analysis, whose details are included in Publication 4. Then, the elemental isotopic profile ($\delta^{13}C$, $\delta^{2}H$, $\delta^{18}O$, $\delta^{15}N$, $\delta^{34}S$, and $^{86}Sr/^{87}Sr$), and the $\delta^{13}C$ and $\delta^{2}H$ of the FAMEs were analysed respectively by EA-IRMS and GC-IRMS in a reduced sample set (n = 40) of raw hazelnuts from four different origins. Chemometric tools were used to build classification models to differentiate hazelnuts according to their region of origin (Publication 4).

Based on the findings of the PoC study (Publication 4), the most relevant isotopic markers for discriminating hazelnut samples from different origins (bulk $\delta^{18}O$, and $\delta^{13}C$ and $\delta^{2}H$ of the FAMEs), were analysed in a larger sample set which included hazelnut samples of the same cultivar (TG) and different geographical origins (n = 207) (**Figure 6**). Discriminant models were built and externally validated to distinguish between hazelnuts from different regions of origin (Publication 5).

To develop the hazelnut authentication models based on **TAG and UF fingerprinting** related to **objective 4**, PoC studies were conducted to assess their feasibility (Publications 6 and 7). Firstly, to develop and validate an efficient and robust TAG fingerprinting method, a preliminary study (**Figure 6**) was carried out to compare four **TAG analysis methods** and to select the most efficient analytical approach for distinguishing TAG compositions with even minor differences (Publication 6). For this method selection, VOOs were selected as samples as VOO and hazelnuts share a very similar lipid profile [44,150]. This saved workload because oil samples could be directly analysed while nuts would have required extraction of the lipid fraction. Moreover, there are already established methods for analysing TAG in VOO, which enabled a direct comparison with our findings. The sample set consisted of 143 traceable genuine VOOs and their blends with adulterant oils of various botanical origins (Publication 6). The selected TAG method was then extrapolated to analyse TAGs in hazelnuts.

Moreover, a PoC study was conducted on a reduced sample set of traceable hazelnuts collected over two consecutive harvest years, 2019 and 2020 (n = 176) **to evaluate the potential of UF** for hazelnut cultivar and geographical authentication (**Figure 6,** Publication 7). This study aimed to optimize sample preparation and instrumental conditions for effective UF data acquisition and to assess the most suitable data processing technique (fingerprinting or untargeted profiling) to identify the optimal method for developing efficient authentication models (Publication 7).

Based on the results of the PoC studies (Publications 6 and 7), the most suitable methods for analysing TAG and UF were applied to the large sample set of 309 hazelnut samples from different origins and cultivars (**Figure 6**). These methods were used to develop and externally validate hazelnut varietal and geographical classification models (Publication 8). The models were built in a challenging scenario in where origin was discriminated among samples of the same cultivar (TG), ensuring that only pedoclimatic factors influenced the markers without genetic interference. In turn, the cultivar (TG/non-TG) was discriminated among samples from the same origin (Spain), guaranteeing that only genetic factors drove the discrimination. In Publication 8, the performance of both methodologies was compared and an integrated TAG and UF fingerprinting method was developed and externally validated.

In line with **objective 4.1**, faster alternatives such as **direct spectroscopic methods** were explored (**Figure 6,** Publication 9). It is well known that NIR and MIR spectra are influenced by components of the lipid fraction. Therefore, given that genetic and pedoclimatic conditions affect the composition of this lipid fraction [44,133,150,153-155], NIR and MIR could be suitable candidates for achieving geographical and varietal authentication tools. To this end, the spectroscopic fingerprint of hazelnuts obtained using MIR and NIR spectroscopy was evaluated on a sample set of 302 hazelnuts from various origins and cultivars (Publication 9). The aim was to develop a fast and accurate method to simultaneously authenticate the geographical and varietal origins of hazelnuts. Independent classification models for origin and cultivar were built for each technique. As with previous methods, cultivar discrimination was performed on samples from the same origin (Spain), and origin discrimination was performed on samples of the same cultivar (TG) produced in different countries. The classification

models were externally validated, and their performances were assessed and compared (Publication 9).

In regards to **pine nut authenticity**, to fulfil **objective 5**, a PoC study was conducted to assess the suitability of the Sr isotope ratio as a marker for the geographical authenticity of pine nuts (**Figure 6,** Publication 10). To achieve this purpose, the $^{86}Sr/^{87}Sr$ ratio of 96 pine nut samples obtained from 2020 to 2023, from different geographical regions and species was analysed (Publication 10). The Sr isotopic analysis method was adapted from Publication 4. The sample set included commercial samples from China, Russia and Turkey, and *P. pinea* samples from Spain obtained directly from producers. The method aimed at discriminating Mediterranean *P. pinea* from pine nuts of other species and origins (**Chapter 6**). Differences in the Sr and Rb content and Sr isotopic ratios across various origins (individual countries, or Spanish regions) and crop years were studied to explore the potential of these specific elements to discriminate pine nuts according to their geographical area of origin

Finally, according to **objective 6**, discriminant models were developed and externally validated based on the SH fingerprint of 253 pine nut samples from various origins and species, analysed by adapting the HS-SPME-GC-MS analytical methodology optimized in Publication 3 [139,140,256], were developed and externally validated (**Chapter 6**). Different models were built to discriminate between Spanish *P. pinea* pine nuts and other species sourced from the main Asian producing regions (China and Russia), and to differentiate between Spanish *P. pinea* produced in two different regions (Central Spain and Catalonia) (Publication 11).

CHAPTER 4.

VIRGIN OLIVE OIL GEOGRAPHICAL AND VARIETAL AUTHENTICATION

Chapter 4 presents the studies aimed at enhancing the **geographical and varietal authentication of VOO** (**Figure 6**). Section 4.1 addresses the comparison of two established methodologies—multi-isotopic analysis and SH fingerprinting—to verify the provenance of VOO (**objective 1**). Section 4.2 details the development and validation of reliable discriminant models based on SH fingerprinting to authenticate VOO 'Arbequina' cultivar (**objective 2**).

4.1 Virgin olive oil geographical authentication

4.1.1 Comparison between multi-isotopic and sesquiterpene fingerprinting methods

To enable a direct systematic comparison between the two methodologies—SH fingerprinting and multi-isotopic analysis— the same sample set of 393 VOOs was analysed by both approaches. This sample set included:

- 393 VOO samples from four consecutive campaigns (2016-2020), six countries of origin (Italy ITA, n= 242; Spain ESP, n = 51; Greece GRC, n = 39; Portugal POR, n = 23; Turkey TUR, n =21; Tunisia TUN, n =17) and various cultivars usually grown in each of these geographical areas.

- Among the 242 Italian samples, 171 came from the three main producing regions (Apulia - PUG, n = 73; Calabria - CAL, n = 58; Sicily - SIC, n = 40).

The bulk isotope values ($\delta^{13}C$, $\delta^{2}H$, $\delta^{18}O$) of VOO samples were determined at the conditions reported by Camin et al. [97].

The same samples were analysed through HS-SPME-GC-MS at the conditions reported by Vichi et al. [256] and seven ions specific to the SH were selected [168,214], yielding the corresponding Extracted ion chromatogram (EIC). EIC matrices generated for each ion were normalized, aligned and concatenated to obtain the unfolded matrix.

The SH unfolded matrix and the isotopic data were used to build independent PLS-DA models (**Figure 7**):

1) A binary PLS-DA model to discriminate Italian (ITA) VOOs from those of other countries (non-ITA).

2) A multiclass PLS-DA model with the Italian VOOs to distinguish among the three main producing regions (CAL, PUG, SIC).

Models were built with 80% of the total sample set (ITA/no-ITA, n =313; CAL/PUG/SI, n = 136), by randomly selecting 80% of the samples from each category, and externally validated by predicting the 20% of samples not included in the model.

Methods' performance was assessed in terms of prediction accuracy, robustness and transferability (Publication 2).

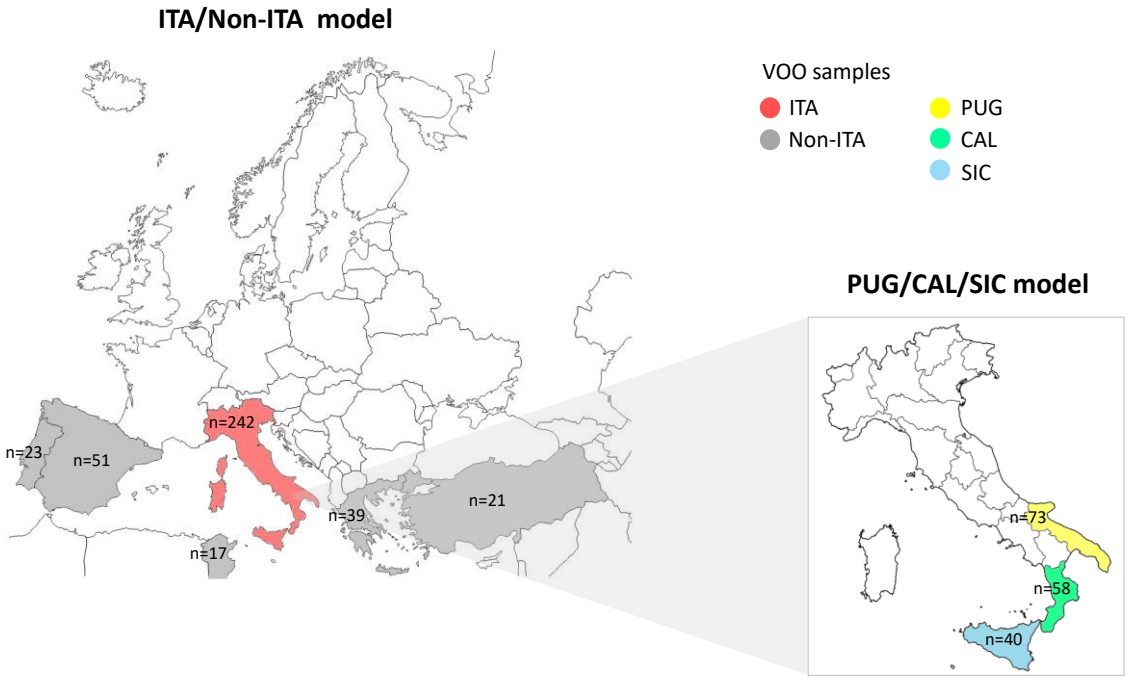


Figure 7. VOO samples corresponding to the comparative study of multi-isotopic and sesquiterpene fingerprinting methods to authenticate VOO geographical origin (Publication 2). VOO: virgin olive oil, ITA: Italy, PUG: Apulia, CAL: Calabria, SIC: Sicilia.

4.1.2 Publication 2

Ground-breaking comparison of target stable isotope ratios vs. emerging sesquiterpene fingerprinting for authenticating VOO origin

Berta Torres-Cobos, Luana Bontempo, Alberto Roncone, Beatriz Quintanilla-Casas, Maurizio Servili, Francesc Guardiola, Stefania Vichi, Alba Tres

Under review in Food Chemistry 2024

Supplementary material available in **Annex 1**

Ground-breaking comparison of target stable isotope ratios vs. emerging sesquiterpene fingerprinting for authenticating VOO origin

Berta Torres-Cobos[1,2], Luana Bontempo[3], Alberto Roncone[3], Beatriz Quintanilla-Casas[1,a], Maurizio Servili[4], Francesc Guardiola[1,2], Stefania Vichi[1,2*], Alba Tres[1,2].

[1] *Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Universitat de Barcelona. Av Prat de La Riba, 171, 08921, Santa Coloma de Gramenet, Spain*

[2] *Institut de Recerca en Nutrició i Seguretat Alimentària (INSA-UB), Universitat de Barcelona. Av Prat de La Riba, 171, 08921, Santa Coloma de Gramenet, Spain*

[3] *Research and Innovation Centre, Fondazione Edmund Mach, Via E. Mach 1, 38098 San Michele all'Adige, Trento, Italy*

[4] *Dipartimento di Scienze Agrarie, Alimentari ed Ambientali, Università di Perugia, Via San Costanzo S.n.c., 06126 Perugia, Italy*

[a] Present address: *Department of Food Science, University of Copenhagen, Rolighedsvej 30, DK-1958, Frederiksberg C, Denmark*

*Corresponding author: Stefania Vichi, PhD, Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Campus de l'Alimentació Torribera, Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona. Av Prat de la Riba, 171. Edifici Gaudí, 08921 Santa Coloma de Gramenet, Spain.

ABSTRACT

This study presents a pioneering comparison of stable isotope ratios analysis and sesquiterpene (SH) fingerprinting for authenticating virgin olive oil (VOO) geographical origin. Both methods were selected for being among the most promising targeted and untargeted approaches, respectively. These methods were applied to the same sample set of nearly 400 VOO samples, covering diverse harvest years, cultivars and producers. PLS-DA classification models were developed to differentiate between Italian and non-Italian VOOs, as well as between VOOs from three closely located Italian regions. Isotopic models based on bulk $\delta^{13}C$, $\delta^{18}O$ and $\delta^{2}H$ achieved over 75% classification accuracy in distinguishing Italian from non-Italian VOOs, while SH fingerprinting outperformed with over 90% accuracy and greater sensitivity to regional differences, as assessed in external validation. This systematic comparison provides insights into the strengths and weaknesses of each method, and the results will guide future research to enhance their effectiveness and reliability in VOO geographical authentication.

Keywords: stable isotope ratios, fingerprinting, sesquiterpene, virgin olive oil, geographical authentication, chemometrics, food fraud.

1. Introduction

Food fraud has gained increasing concern over the years and currently remains a critical issue undermining food chain integrity (Bannor et al., 2023; Everstine et al., 2024). Due to the reported vulnerability to fraud in the olive oil supply chain, particularly for extra virgin olive oil (EVOO) (Yan et al., 2020), this commodity has persistently ranked among the foods with the highest fraud incidence (The EU Food Fraud Network and the System for Administrative Assistance & Food Fraud. Annual Reports 2014-2020; Joint Research Center of the European Commission). A substantial percentage of fraud cases in official reports were related to mislabelling issues, including false declarations of origin. Mandatory declarations of origin require specifying whether virgin olive oils (VOOs) come from a European Union (EU) country, a non-EU country, a blend of EU and non-EU oils, or a combination of both (Regulation EU No 2022/2104), or indicating the Protected Designation of Origin (PDO) or Protected Geographical Indication (PGI) for oils under geographical quality schemes (Regulation (EU) 2024/1143). These EU regulations aim to help consumers make informed purchasing decisions and to protect them from misleading information, given the significant influence that the geographical origin of EVOO has on consumer choices (Conte et al., 2020). Indeed, the country of origin significantly influences consumer preferences, a trend especially notable for oils from certain Mediterranean EU countries such as Italy (Kavallari et al., 2014), which serves as an exemplary case. Italy is the second-largest olive oil producer in the EU after Spain, accounting for around 15% of EU production over the past decade while contributing more than 30% to the EU's olive oil exports (IOC, 2023). This phenomenon is closely tied to the internationally renowned "Made in Italy" label, which positions Italy as a leader in the global food market (Cappelli et al., 2017). The high-quality reputation of Italian agri-food products enhances demand, prompting consumers to willingly pay premium prices across various sectors, including VOOs (Cappelli et al., 2017; Carbone & Henke, 2023). This trend is also evident in the domestic Italian market, where Italian-declared EVOO commands a premium price 35% higher than other European EVOOs and over 45% higher than EVOOs from non-European origins (Bimbo et al., 2020). Combined with higher production costs compared to other EU and non-EU countries, this results in

Italian EVOO reaching the highest prices in both international and domestic markets (IOC, 2023; Bimbo et al., 2020).

According to Spink and Moyer (2011), this situation fosters fraud due to consumers' willingness to pay premium prices for Italian EVOOs, to strong incentives to counterfeit with cheaper alternatives, and to inadequate control methods. In this sense, the lack of an official analytical method to verify the geographical origin of EVOOs and VOOs, despite legal requirements on labels, is a critical gap in preventing fraud that significantly concerns stakeholders and needs urgent attention (Conte et al., 2020; Casadei et al., 2021). In response, researchers from various fields have intensified their efforts on developing reliable methods for VOO geographical authentication (Maléchaux et al., 2020), leading to significant advancements in the state of the art (Bajoub et al., 2016; Conte et al., 2020; Zaroual et al., 2021). This research has focused on developing efficient, cost-effective, and relatively rapid screening methods capable of effectively detecting fraud, while also addressing challenges such as reproducibility and transferability, crucial for adoption by regulatory authorities.

One of the most recognized methods for establishing the geographical origin of food is stable isotope analysis, which relies on the strong influence of production zone factors, such as geology and hydrogeology, on the isotopic composition of agricultural products (Laursen et al., 2016). Isotopic analysis of light bio-elements (C, H, O, N, S) has been widely and successfully applied to verify the geographical origin of various foodstuffs and, in some cases, has even been proposed for legal verification (Camin et al, 2017). Stable isotopes of light elements are usually analysed in bulk matrices by Elemental Analysis-Isotope Ratio Mass Spectrometry (EA-IRMS), which requires minimal sample manipulation and short analysis time. Regarding its application for tracing the origin of VOO, the dependence of bulk $\delta^{13}C$, $\delta^{18}O$ and $\delta^{2}H$ on latitude and altitude, distance from the sea, and environmental and climatic conditions enabled characterising Italian VOO production (Camin et al., 2010b; Portarena at al., 2014; Chiocchini et al., 2016), differentiating part of Italian VOOs from Tunisian oils (Camin et al., 2016) and from other EU oils (Camin et al., 2010a), and establishing significant differences among Italian macro-regions (Bontempo et al., 2009; Camin et al., 2010a; Portarena at al., 2014).

The main strengths of stable isotope analysis are its high precision under repeatable measurement conditions (<0.05% RSD) and the use of certified reference materials for bias correction. Moreover, isotope ratio analysis can be implemented by accredited food testing laboratories according to ISO and AOAC standards (Bayen et al., 2024). These factors are crucial for achieving highly accurate results, unifying isotope measurements, and ensuring the comparability of isotope data worldwide. On this basis, the Italian Ministry of Agricultural, Food and Forestry Policies (MIPAAF, Italy) established the Italian National Database of PDO/PGI EVOOs for Italian olive oil traceability, which has been recording olive oil isotopic measurements since 2000. However, isotopes are highly dependent on the harvest season (Camin et al., 2010b), which can hinder the detection of fraud in oils from different harvests. Moreover, significant differences in the isotopic profiles of bulk VOO from various geographical areas often come with considerable overlap, making direct differentiation difficult. Therefore, it is often necessary to combine bulk isotopic analysis with compound-specific isotopic analysis (Faberi et al., 2014; Bontempo et al., 2019) or other techniques and markers, such as elemental profile (Camin et al., 2010a; 2010b) or metabolite analysis (Lukić et al., 2020; Faberi et al., 2014; Portarena et al., 2017).

Untargeted metabolomics represents a cutting-edge approach for food authentication, providing more comprehensive data for fraud detection compared to traditional targeted methods. This is because untargeted methods are not limited to predefined compounds but consider comprehensive data, enhancing authentication efficiency (Ballin & Laursen, 2019; Amaral, 2020). Fingerprinting methods, using raw analytical data, combined with chemometric techniques such as partial least squares discriminant analysis (PLS-DA) have proven highly effective for authenticating VOO (Quintanilla-Casas et al., 2020a; 2020b; Torres-Cobos et al., 2021). However, in contrast to stable isotope analysis, the transferability of untargeted methods between laboratories is still challenging, particularly in the case of chromatographic fingerprinting methods, because of the absence of clear guidelines for assessing their analytical performance (Bayen et al., 2024; Riedl et al., 2015). Some strategies for in-house validation of chromatographic fingerprinting methods have yielded promising results (Quintanilla-Casas et al., 2020b), that encourage the additional analytical efforts that are still needed to achieve their full inter-laboratory transferability. This would be

particularly interesting for fingerprinting methods showing a high efficiency for VOO geographical authentication (Quintanilla-Casas et al., 2022a; 2022b; Torres-Cobos et al., 2021). Specifically, sesquiterpene hydrocarbons (SHs) have been identified as robust geographical markers for VOOs due to their strong association with olive tree cultivars and growing regions, and their stability against processing and storage conditions (Quintanilla-Casas et al., 2020a; Vichi et al., 2018). In fact, SH chromatographic fingerprints analysed using PLS-DA successfully distinguish VOOs based on their origin across different levels, including EU-wide and single-country labels, as well as adjacent PDOs (Quintanilla-Casas et al., 2022a; 2022b). This method, using affordable and automatable instrumentation and requiring minimal sample manipulation, offers extremely high classification accuracy and has a high potential as a fit-for-purpose screening tool for verifying the geographical origin of VOO.

Therefore, both stable isotope analysis and SH fingerprinting, selected as promising representatives of targeted and untargeted methods for VOO geographical authentication, respectively, offer distinct advantages and capabilities, and both appear as suitable strategies for Italian VOO authentication. However, the results in literature obtained by SH fingerprinting and stable isotope analysis for VOO geographical authentication are not easily comparable, underscoring the need for a systematic comparison under standardized conditions. This is needed to objectively evaluate and contrast their performance in authenticating VOO origin and to identify areas for further advancement in these approaches. Such a comparative study should involve testing both methods on the same sample set, applying consistent statistical treatments, and assessing their performance across various challenging scenarios in terms of sample variability and level of VOO geographical differentiation.

With this aim, PLS-DA classification models were developed on stable isotope data (bulk C, H, and O) and SH fingerprinting data obtained from the exact same sample set, comprising nearly 400 VOO samples, to differentiate between Italian and non-Italian VOOs, as well as among VOOs from closely situated Italian production regions. These oils were produced over multiple harvest years and showed significant variability in cultivars, producers, and processing techniques, providing a challenging scenario to test the methods under investigation. This pioneering study represents the first systematic comparison between stable isotope analysis and a metabolic fingerprinting

approach for VOO authentication, offering valuable insights into the strengths and weaknesses of each method. The results will be useful in guiding future research efforts aimed at enhancing their effectiveness and reliability in VOO geographical authentication.

2. Material and methods

2.1 Samples

The sample set consisted of 393 traceable VOOs from various countries and geographical regions (Table 1), produced from 2016-2017 to 2019-2020. Part of these samples were from Italy (ITA, n=242), while the remaining samples were oils from other five Mediterranean countries [non-ITA, n=151: Spain (ESP), n=51; Greece (GRC), n=39; Portugal (POR), n=23; Turkey (TUR), n=21; Tunisia (TUN), n=17]. Italian samples were produced in different regions: Apulia (PUG, n=73), Calabria (CAL, n=58), Sicily (SIC, n=40), and other regions (Lombardia/Emilia Romagna, n=16; Tuscany, n=7; Liguria, n=6; Sardinia, n=4; Basilicata, n=4; Umbria, n=3; Abruzzo, n=2; Campania, n=2; Marche, n=1) or were from not specified Italian regions (n=26) (Figure 1). Additional information about the samples is available in Table S1 of Supplementary information. The samples were stored under a nitrogen (N_2) atmosphere at -20°C until analysis.

Table 1. Geographical origin and harvest year of the 393 VOOs analysed.

Italy (ITA)	n	2016/2017	2017/2018	2018/2019	2019/2020
Apulia (PUG)	73	4	1	35	33
Calabria (CAL)	58	1	1	21	35
Sicily (SIC)	40	3	0	15	22
Other regions	45	6	6	27	6
No region specified	26	1	2	1	22
Total ITA	242	15	10	99	118
Other countries (non-ITA)	n				
Spain (ESP)	51	0	0	25	26
Greece (GRC)	39	0	0	9	30
Portugal (POR)	23	0	12	6	5
Turkey (TUR)	21	10	11	0	0
Tunisia (TUN)	17	1	5	0	11
Total non-ITA	151	11	28	40	72

2.2 Isotopic analysis of bulk VOO by Elemental Analysis-Isotope Ratio Mass spectrometry (EA-IRMS)

VOO samples were weighed (ca 0.3 mg) and placed in tin capsules to measure the $\delta^{13}C$ using an isotope ratio mass spectrometer (Elementar Analysensysteme GmbH, Langenselbold, Germany) after total combustion in an elemental analyser (Vario Isotope Cube; Elementar Analysensysteme GmbH).

The $\delta^{18}O$ and $\delta^{2}H$ ratios were obtained by weighing approximately 0.25 mg of the sample in a silver capsule and introducing it into a TC/EA (Finnigan DELTATC/EA, high temperature conversion elemental analyser, Thermo Scientific). The samples were measured in duplicate for $\delta^{13}C$ as well as for $\delta^{18}O$ and $\delta^{2}H$.

In accordance with IUPAC protocol, isotopic values are expressed as delta relative to the international standards: V-PDB (Vienna-Pee Dee Belemnite) for $\delta^{13}C$, V-SMOW (Vienna-Standard Mean Ocean Water) for $\delta^{2}H$ and $\delta^{18}O$, and air (atmospheric N_2) for $\delta^{15}N$, as described in Equation (1):

$$\delta^{i}(E_{sample/standard}) = \frac{R({}^{i}E/{}^{j}E)_{sample}}{R({}^{i}E/{}^{j}E)_{standard}} - 1 \qquad (1)$$

where 'standard' refers to the international measurement standard, 'sample' is the analysed specimen, and ${}^{i}E/{}^{j}E$ correspond to the isotope ratio between heavier and lighter isotopes. Delta values are multiplied by 1000 and are commonly expressed in per mil (‰) or, in accordance with the International System of Units (SI), as 'milliurey' (mUr).

Carbon isotopic values $\delta^{13}C$ were calculated against USGS 88 ($\delta^{13}C$ -16.06 ‰), while oxygen $\delta^{18}O$ and deuterium values $\delta^{2}H$ were calculated against 2 international standards (USGS 84 $\delta^{18}O$ 26.36‰, $\delta^{2}H$ -140.4‰ and USGS 86 $\delta^{18}O$ 18.76‰, $\delta^{2}H$ -207.4‰) through the creation of a linear equation. The accepted maximum standard deviations for repeatability were 0.3‰ for $\delta^{13}C$, 0.5‰ for $\delta^{18}O$, and 4‰ for $\delta^{2}H$.

2.3 Sesquiterpene fingerprinting by headspace-solid phase microextraction-gas chromatography-mass spectrometry (HS-SPME-GC-MS)

The SH fingerprint of VOO samples was analysed by GC–MS after extraction by HS-SPME according to Torres-Cobos et al. (2021), based on the original protocol from Vichi

et al. (2006). For this, a Combi-PAL autosampler (CTC Analytics, Zwingen, Switzerland) was used, in combination with an Agilent 6890N Network GC system coupled to a quadrupolar mass selective analyser Agilent 5975C Inert MSD (Agilent Technologies, Santa Clara, California, USA). Briefly, 2 g of oil was weighed into a 10 mL vial fitted with a PTFE/silicone septum and maintained at 70°C under constant agitation (250 rpm). After 10 minutes of sample conditioning, a divinylbenzene/carboxen /polydimethylsiloxane (DVB/CAR/PDMS) fiber (2 cm length, 50/30 µm film thickness) provided by Supelco (Bellefonte, PA) was exposed to the sample headspace for 60 minutes. The fiber was then desorbed in the GC injection port at 260°C for 10 minutes, with the injector operating in split-less mode for the first 5 minutes of desorption. Separation was performed on a Supelcowax-10 capillary column (60 m × 0.25 mm i.d., 0.25 µm film thickness) (Supelco, Bellefonte, PA), using helium as the carrier gas, at 1.5 mL/min.

Mass spectra acquisition was carried out in selected ion monitoring (SIM) mode, targeting *m/z* 93, 119, 157, 159, 161, 189, and 204, which are recognized as the main specific ions of SHs (Vichi et al., 2010). Extracted Ion Chromatograms (EICs) obtained for the target ions were considered from 21 min to 42 min (3197 scans for each EIC). A fingerprinting approach was then followed using the scan intensities of the EICs. A data matrix was constructed for each ion, with all samples (rows, n=393) and the scan intensities of each EIC as variables (columns) (7 different data matrices with 3197 scans × 7 ions).

To correct differences between injections, each EIC was normalized to the maximum intensity (row wise). Subsequently, the EICs of each ion matrix were aligned among them using the Correlation Optimized Warping (COW) algorithm in Matlab® (Nielsen et al., 1998) to correct the retention time shifts among samples. Finally, the 7 aligned EIC matrices were concatenated conforming a two-way unfolded matrix (393 samples × 22379 variables).

2.4 Statistical analysis

2.4.1 Univariate analysis of isotopic data

To examine differences in isotopic ratios across various origins (ITA/non-ITA, individual countries, Italian regions), we used statistical tests to evaluate population distribution

and compare population medians in IBM SPSS Statistics v29.0© (IBM Corp., Armonk, New York, USA). As isotopic ratios did not follow a normal distribution, determined by the Shapiro-Wilk test, or had fewer than 30 samples in the compared populations, the independent samples median test (non-parametric) was used to compare the medians of ITA vs non-ITA, the countries of origin and the Italian regions (CAL/SIC/PUG). In all cases, $p<0.05$ was considered significant.

2.4.2 Development of partial least square-discriminant analysis (PLS-DA) classification models

First, for both approaches, Principal Component Analysis (PCA) was carried out to explore the data (n = 393) and to identify potential outliers based on Hotelling's T^2 range and Q-residuals model parameters. No outliers were detected according to these parameters.

The data matrices obtained by stable isotope analysis and SH fingerprinting were separately used to construct and validate individual PLS-DA classification models with SIMCA v13.0© (Sartorius, Göttingen, Germany). For each method (stable isotope analysis and SH fingerprinting), two types of classification models were developed: i) a binary ITA/non-ITA model (n = 393) to differentiate ITA VOOs (n = 242) from those produced in five other major Mediterranean countries (non-ITA, n = 151); and ii) a multi-class regional model (n = 171) to distinguish among three major ITA producing regions (Apulia, n = 73; Calabria, n = 58; and Sicily, n = 40).

For each type of authentication model (ITA/non-ITA or regional) and method (stable isotope analysis or SH fingerprinting), the sample set was split following a stratified random sampling strategy into a training set [80% of samples from each category: ITA vs non-ITA model, n = 313 (ITA, n = 193; non-ITA, n = 120); regional model, n = 136 (Apulia, n= 58; Calabria, n = 46; and Sicily, n = 32)] and a validation set (20% of samples from each category: ITA/non-ITA model, n = 80; regional model, n = 35). This splitting process was repeated three times (3 iterations) to assess the impact of sample set composition and enhance the robustness of the external validation. Details of the sample set splitting, including the training and validation sets, are provided in **Table S1** of the **Supplementary Information**. To enable a rigorous comparison between

methods, the exact same splitting was applied to models based on stable isotope data and SH fingerprinting data.

2.4.3 Validation of partial least square-discriminant analysis (PLS-DA) classification models

First, with each training set of each of the three iterations, a PLS-DA model was calibrated and internally validated through leave-10 %-out cross-validation. In each iteration, the number of latent variables (LV) and optimal preprocessing (which was mean centering and scaling to unit variance) were chosen based on the lowest Root Mean Squared Error of Cross Validation (RMSEcv) criteria. Subsequently, potential overfitting of the models was assessed using permutation tests (n = 20 permutations) and ANOVA on the cross-validated predictive residuals (p-value). Then, each model was externally validated by predicting the class of samples in the corresponding validation set, which had not been used in model development.

In all PLS-DA binary models, classes were represented using PLS dummy variables (1 for non-ITA class and 0 for ITA class). In multi-class PLS-DA models, each class is modelled individually against the rest of samples. In this case, the dummy Y matrix contained vectors corresponding to each class, where each vector assigned a value of 1 to its specific class (Apulia, Calabria, Sicily) and 0 to all other classes (non-Apulia, non-Calabria, non-Sicily). Subsequently, each sample was assigned to the class corresponding to the vector with the highest PLS predicted value (PV), provided it exceeded the classification threshold. Samples that did not meet the threshold for any vector were left unassigned (no class).

The performance of each PLS-DA model was assessed by the Q^2 values and the percentage of correct classification in external validation, expressed as mean value of correct classification rate ± standard deviation obtained from the 3 iterations. For the binary models (ITA/non-ITA), the sensitivity and specificity were also assessed, according to Magnusson & Örnemark (2014).

2.4.4 Optimisation of classification thresholds by receiver operating characteristic (ROC) analysis

To maximize the performance of the developed models, classification thresholds were optimized generating the receiver operating characteristic (ROC) with PVs obtained

from leave-10%-out cross-validation. The ROC curve plots the sensitivity (true positives/ [true positives + false negatives]) against 1-specificity (1 - [true negatives/ (true negatives + false positives)]) resulting from varying the PV threshold to assign samples to a diagnostic category (ITA or non-ITA; and Apulia, Calabria or Sicily) (Fawcett, 2006). In this case, the positive classes were non-ITA, Apulia, Calabria and Sicily for the corresponding regional models. ROC analysis was applied on PV values from each individual PLS-DA model. Thus, a total of 24 ROC curves (3 random training sets for 4 diagnostic categories: 1 for the ITA/non-ITA model and 3 for the regional model) were generated for each method (stable isotope analysis and SH fingerprinting). The optimal thresholds for classifying the validation samples, detailed in **Table S2** of the **Supplementary Information**, were those that maximized the sum of sensitivity and specificity (Quintanilla-Casas et al., 2020b).

3. Results and discussion

3.1 Stable isotope analysis

Median $\delta^{18}O$ values determined in bulk VOOs produced across four harvest seasons, were significantly different between ITA ($\delta^{18}O$=24.0) and non-ITA ($\delta^{18}O$=25.8) classes ($p<0.001$), unlike $\delta^{13}C$ and $\delta^{2}H$ values (ITA: $\delta^{13}C$= -29.7; $\delta^{2}H$=-144.0; non-ITA: $\delta^{13}C$= -29.6; $\delta^{2}H$=-143.2) (**Table S3** of **Supplementary information**). As previous studies have indicated that the differentiation between geographical macro-areas is explained by the distinct characteristics of the sub-areas that compose them (Bontempo et al., 2019; Quintanilla-Casas et al., 2022), the behaviour of the different isotopic markers was evaluated across the six countries and the three Italian regions studied (**Table S3** of **Supplementary information**). The median isotopic values were consistent with previous reports (Bontempo et al., 2009; Bontempo et al., 2019; Camin et al., 2016, Chiocchini et al., 2016) and, in most cases, showed significant differences between countries and regions, explaining the observed significant differences between ITA and non-ITA classes. In particular, $\delta^{18}O$ presented the lowest value in ITA compared with all the non-ITA countries analysed, while $\delta^{13}C$ and $\delta^{2}H$ in ITA oils significantly differed from GRC, TUN, and TUR; and from GRC, POR, TUN, and TUR oils, respectively (**Table S3** of **Supplementary information**). However, despite the differences in median values between ITA and non-ITA oils, as well as between individual countries or regions, the

corresponding quartile and minimum-maximum ranges indicated that no single marker could clearly distinguish VOOs of any provenance.

This underscores the importance of investigating the potential of multi-isotopic analysis combined with multivariate techniques to achieve more accurate origin discrimination. Applying PLS-DA to the multi-isotopic data to differentiate VOO between ITA and non-ITA classes, as well as among Italian regions, achieved global classification rates of 74.5% and 65.2%, respectively, using leave-10%-out cross-validation (**Tables S4 and S5** of the **Supplementary Information**). To confirm these results, each model, across the three iterations, was externally validated by predicting the class of the corresponding validation samples, which were not used during model development. The external validation of the ITA/non-ITA PLS-DA models developed on the stable isotope data (**Table 2**) resulted in an overall classification rate >75%, with a good sensitivity (0.80) and acceptable specificity (0.73). Examining the identity of the misclassified samples in each validation set (**Table S1** of the **Supplementary Information**) provided valuable insights into the strengths and limitations of the classification model. For the isotopic ITA/non-ITA PLS-DA model, the non-ITA samples with the highest misclassification rates in external validation were from ESP, GRC, and especially from TUR, with 20%, 25%, and 50% of test samples misclassified across the three iterations, respectively. Some misclassification was expected between ESP and ITA samples, given that the preliminary univariate comparison showed only $\delta^{18}O$ to be significantly different between these classes. In contrast, higher classification efficiency was expected for GRC and TUR samples, as their median values for all stable isotopes tested were significantly different from ITA samples. This demonstrates that even when median isotopic values are significantly different between classes, this alone does not guarantee accurate discrimination, as it also depends on the overall dispersion of the samples. Therefore, when single thresholds for these target markers cannot efficiently distinguish sample classes, multivariate classification methods may be helpful for more accurately assessing their effective discrimination capacity.

Table 2. Results of the external validation of the ITA/non-ITA PLS-DA models developed on the stable isotope data[1]. Results are mean values ± standard deviation obtained from three iterations.

	n	Correct classification (%)	non-ITA (n)	ITA (n)	Sensitivity	Specificity
non-ITA	35	**80.2 ± 11.8**	25.7 ± 3.8	6.3 ± 3.8	0.80 ± 0.12	
ITA	48	**72.9 ± 2.1**	13.0 ± 1.0	35.0 ± 1.0		0.73 ± 0.02
Total	83	**75.8 ± 3.8**				

Abbreviations: ITA, Italy

[1]Training model (N = 313, 2 LVs) parameters: mean values obtained with the training sets from 3 iterations. Q^2 = 0.299, RMSEcv = 0.406. For all models, ANOVA p-value < 0.05.

Regarding misclassification of ITA samples, Apulian, Calabrian and Sicilian VOOs were the most frequently misclassified as non-ITA (**Table S1** of the **Supplementary Information**), which is reasonably attributable to their higher proportion in the sample set. Finally, although misclassified samples came from various harvest years and partially reflected their proportions in the sample set, the harvest year did seem to impact the misclassification rate. Specifically, even though the sample set was dominated by VOOs from 2018/2019, but particularly from 2019/2020, the 2018/2019 samples were misclassified more frequently. This higher misclassification rate for the 2018/2019 samples might be related to climatic differences registered across the Mediterranean countries during 2018 (Climate Change Knowledge Portal, 2018). In 2018, ITA experienced slightly higher temperatures and lower precipitation compared to other years, which might have made its climatic conditions more similar to those of other Mediterranean countries. Concurrently, some of these countries, such as ESP and TUR, recorded temperatures and especially rainfall that approached Italy's typical annual averages (**Table S6** of the **Supplementary information**). This might have led to the isotopic signatures of Italian VOOs from the 2018/2019 harvest resembling those of other Mediterranean VOOs more closely.

Contextualizing classification results of **Table 2** with respect to those reported in the literature is challenging because a direct comparison is not always feasible. While many available studies complement bulk isotopic values with compound-specific isotopic values or non-isotopic markers (Bontempo et al., 2019; Faberi et al., 2014), others focusing solely on bulk $\delta^{13}C$, $\delta^{2}H$, and $\delta^{18}O$ isotopic markers typically examine VOOs

from specific regions (Camin et al., 2010a) or single harvest seasons (Jiménez-Morillo et al., 2020). These studies cannot be directly compared to the present research, which includes broader geographical areas and multiple harvest years, thereby incorporating significant isotopic variability. Finally, comparisons are further complicated by varying data treatment methods across studies and the infrequent use of multivariate classification techniques for differentiating VOOs based on geographical origin (Bontempo et al., 2009; Camin et al., 2010b; 2016; Chiocchini et al., 2016; Portarena et al., 2014).

While direct comparison with previous results is difficult, the findings of this study can be regarded as highly satisfactory given the complexity of the sample set, which includes VOOs from various regions, five Mediterranean countries, and up to four harvest seasons, all based solely on bulk $\delta^{13}C$, $\delta^{2}H$, and $\delta^{18}O$ values. Additionally, the robustness of these findings is reinforced by the external validation process, which included three iterations of the sample set to prevent overly optimistic outcomes.

Regarding the differentiation between VOOs from the three adjacent Italian regions, Apulian and Sicilian samples showed satisfactory correct classification rates of 75.6% and 83.3%, respectively, in external validation. However, Calabrian samples showed a poor classification rate (36.1%), and they were often misclassified as Apulian VOOs, resulting in a lower overall classification accuracy (**Table 3**) compared to the ITA/non-ITA model. This higher misclassification was likely due to the regions' proximity and their geographical and climatic similarities. This aligns with previous studies that reported similar $\delta^{13}C$ and $\delta^{18}O$ values for Calabrian and Apulian oils, unlike other regions such as Sicily, with these values being correlated to the specific climatic conditions of each region (Chiocchini et al., 2016; Portarena et al., 2014). Sicilian VOOs also showed higher $\delta^{2}H$ values, but no previous comparisons of this marker among VOOs from SIC, CAL and APU are available in the literature. Similar to the ITA/non-ITA model, the regional model showed that the harvest year influenced classification accuracy, with samples from 2018/2019 exhibiting a higher tendency to be misclassified (**Table S1** of the **Supplementary Information**).

Table 3. Results of the external validation of the regional three-class PLS-DA models developed on the stable isotope data[1]. Results are mean values (± standard deviation) obtained from three iterations.

	n	Correct classification (%)	Apulia (n)	Calabria (n)	Sicily (n)	No class (n)
Apulia	15	**75.6 ± 3.8**	11.3 ± 0.6	2.3 ± 0.6	1.0 ± 1.0	0.3 ± 0.6
Calabria	12	**36.1 ± 21.0**	6.7 ± 1.5	4.3 ± 2.5	1.0 ± 1.0	0.0 ± 0.0
Sicily	8	**83.3 ± 19.1**	0.7 ± 1.2	0.3 ± 0.6	6.7 ± 1.5	0.3 ± 0.6
Total	35	**63.8 ± 11.5**				

[1]Training model (N = 136, 2-3 LVs) parameters: mean values obtained with the training sets from 3 iterations. Q^2 = 0.225, RMSEcv = 0.460. For all models, ANOVA p-value < 0.05.

3.2 Sesquiterpene fingerprinting

PLS-DA models developed using SH fingerprinting data achieved 99.9% and 100% classification accuracy for distinguishing between ITA and non-ITA samples and for identifying the Italian region of origin, respectively, based on leave-10%-out cross-validation (**Tables S4 and S5** of the **Supplementary Information**). As for the isotopic models, SH fingerprinting models were then externally validated by predicting the class of the samples in the corresponding validation sets, and expressing the results as mean values ± standard deviation obtained from the 3 iterations (**Tables 4 and 5**). The classification accuracy in external validation maintained over 90% for both classes (ITA and non-ITA) and reached an overall accuracy of 91.7%, providing sensitivity and specificity values close to 1. These results are in line with those obtained in previous models based on SH fingerprinting, which aimed to distinguish VOO based on their EU, non-EU, single country or PDO origin (Quintanilla-Casas et al., 2022a; 2022b), confirming the extraordinary efficiency of this method for VOO geographical authentication. Misclassified samples were not clearly related to any specific country of origin or harvest year (**Table S1** of the **Supplementary information**) but rather seemed to be individual samples with specific characteristics.

Table 4. Results of the external validation of the ITA/non-ITA PLS-DA models developed on the SH data[1]. Results are mean values (± standard deviation) obtained from three iterations.

	n	Correct classification (%)	non-ITA (n)	ITA (n)	Sensitivity	Specificity
non-ITA	32	**90.6 ± 6.3**	29.0 ± 2.0	3.0 ± 2.0	0.91 ± 0.06	
ITA	48	**92.4 ± 7.3**	3.7 ± 3.5	44.3 ± 3.5		0.92 ± 0.07
Total	80	**91.7 ± 6.2**				

Abbreviations: ITA, Italy.

[1] Training model (N= 313, 8-9 LVs) parameters: mean values obtained with the training sets from 3 iterations. $Q^2_{=}$ 0.721, RMSEcv = 0.228. For all models, ANOVA p-value < 0.05.

Table 5. Results of the external validation of the regional three-class PLS-DA models developed on the SH data[1]. Results are mean values (± standard deviation) obtained from three iterations.

	n	Correct classification (%)	Apulia (n)	Calabria (n)	Sicily (n)	No class (n)
Apulia	15	**82.2 ± 13.9**	12.3 ± 2.1	1.3 ± 1.5	0.7 ± 0.6	0.7 ± 0.6
Calabria	12	**83.3 ± 8.3**	1.0 ± 1.0	10.0 ± 1.0	0.7 ± 1.2	0.3 ± 0.6
Sicily	8	**79.2 ± 7.2**	0.3 ± 0.6	1.3 ± 0.6	6.3 ± 0.6	0.0 ± 0.0
Total	35	**81.9 ± 5.9**				

[1] Training model (N = 136, 13-14 LVs) parameters: mean values obtained with the training sets from 3 iterations. $Q^2_{=}$ 0.627, RMSEcv =0.292. For all models, ANOVA p-value < 0.05.

Considering the differentiation of VOOs from closely located Italian regions, classification models based on SH fingerprinting data achieved an accuracy close to 80% for all classes (**Table 5**). Interestingly, Calabrian VOOs, which were not satisfactorily classified by the isotopic model (36.1%), achieved the highest accuracy with the SH fingerprinting model (83.3%). Conversely, Sicilian VOOs, which had the highest correct classification rate with the isotopic PLS-DA model (83.3%), showed slightly lower accuracy with the SH fingerprinting model (79.2%). This can be attributed to the satisfactory relationship between the isotopic markers and Sicilian samples, and to the slightly lower representation of Sicilian samples in the model (**Table 1**), which likely impacted the performance of SH fingerprinting models more than it affected the isotopic models based on only three markers. This can be justified by the fact that models based on fingerprinting data operate with high-dimensional datasets, which require a larger number of samples to adequately represent the underlying data patterns associated with each class (Brereton, 2006). In contrast, targeted models with

fewer variables are less sensitive to sample size because their simpler relationships among variables result in lower noise and a reduced risk of overfitting.

3.3 Comparative analysis of strengths and aspects for improvement

This study provides a thorough comparison of the effectiveness of targeted isotopic profiling versus untargeted SH fingerprinting, both selected for their capability in authenticating the geographical origin of VOO. By applying both approaches to the exact same samples and using identical statistical treatments and validation sets, it was possible, for the first time, to compare the efficiency of these methods in terms of measurable indices, such as classification accuracy, sensitivity, and selectivity. This comparison was conducted within the complex scenario of verifying the origin of Italian VOO, a pressing and unresolved issue that demands effective solutions for detecting counterfeiting. This involved, on the one hand, distinguishing ITA VOOs from a highly diverse group of VOOs from various other producing countries, and on the other hand, discriminating between VOOs produced in closely located and relatively similar Italian regions. This context allowed for a thorough evaluation of both methods to assess their effectiveness and limitations. The ITA/non-ITA isotopic models demonstrated satisfactory discrimination power, achieving an overall classification accuracy of over 75% using just three isotopic markers determined on bulk VOO. This confirms that multi-isotopic methods are among the most effective targeted approaches for geographical authentication. However, authentication models based on untargeted SH fingerprinting achieved classification accuracies over 90% using the same experimental design. Moreover, they revealed a lower sensitivity to the harvest year in terms of sample misclassification. The ability to discriminate among VOOs from adjacent geographical regions further highlighted the differences between the two analytical approaches. Isotopic markers proved effective only when the origin regions were geographically or climatically distinct; otherwise, they struggled to differentiate VOOs, as seen with Calabrian oils (classification accuracy of 36%). In contrast, untargeted fingerprinting combined with PLS-DA demonstrated greater sensitivity to regional differences, achieving overall classification accuracy over 80%. Factors not only related with the presence of distinct local cultivars but also with slight variations in pedoclimatic conditions among neighbouring regions can significantly influence the SH

profile, resulting in specific features for each region, as previously reported (Quintanilla-Casas et al., 2022a).

On the other hand, it should be borne in mind that the performance of PLS-DA models based on fingerprinting data is more sensitive to sample size than isotopic models based on a limited number of target variables, due to the higher dimensionality, complexity, and variability of the data. For this reason, more samples are needed to, reduce noise, and prevent overfitting, thereby achieving reliable and accurate classification.

The results of this research showed that the untargeted SH fingerprinting method outperformed isotopic methods in several aspects, despite facing well-known transferability challenges compared to stable isotope analysis. Targeted isotopic methods offer greater applicability and versatility, as they can be more easily adapted and implemented across various contexts, making them more suitable for widespread use. Given the strong potential of SH fingerprinting for VOO geographical authentication demonstrated by this study, there is an urgent need to enhance the transferability of this method. This would help overcome complex authentication challenges, such as verifying the origin of Italian VOOs.

Funding

B. Torres-Cobos thanks the Spanish Ministry of Universities predoctoral fellowships FPU20/01454.

Declaration of competing 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.

Acknowledgments

INSA-UB Maria de Maeztu Unit of Excellence (Grant CEX2021-001234-M) funded by MICIU/AEI/FEDER, UE. INSA-UB authors are part of the LiBiFOOD Consolidated Research Group (2021-SGR-00854) recognised by AGAUR (Catalan Government). Unaprol - Italian Olive Consortium - for assistance in samples of Italian Virgin olive oil collecting and financial support.

CRediT authorship contribution statement

B. Torres-Cobos: Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft. **L. Bontempo**: Methodology, Writing – review & editing. **A. Roncone**: Formal analysis, Writing – review & editing. **B. Quintanilla-Casas**: Formal analysis, Investigation, Methodology, Writing – review & editing. **M. Servili**: Resources, Writing – review & editing; **F. Guardiola**: Supervision, Writing – review & editing. **S. Vichi**: Conceptualization, Methodology, Resources, Supervision, Writing – review & editing. **A. Tres**: Conceptualization, Methodology, Resources, Supervision, Writing – review & editing.

REFERENCES

1. Amaral., J.S. (2020). Target and non-target approaches for food authenticity and traceability. Foods, 10, Article 172. https://doi.org/10.3390/foods10010172

2. Bajoub, A., Bendini, A., Fernández-Gutiérrez, A., & Carrasco-Pancorbo, A. (2017). Olive oil authentication: A comparative analysis of regulatory frameworks with especial emphasis on quality and authenticity indices, and recent analytical techniques developed for their assessment. A review. Critical Reviews in Food Science and Nutrition, 58, 832–857. https://doi.org/10.1080/10408398.2016.1225666

3. Ballin, N.Z., & Laursen, K.H. (2019). To target or not to target? Definitions and nomenclature for targeted versus non-targeted analytical food authentication. Trends in Food Science & Technology, 86, 537–543. https://doi.org/10.1016/j.tifs.2018.09.025

4. Bannor, R.K., Arthur, K.K., Oppong, D., & Oppong-Kyeremeh, H. (2023). A comprehensive systematic review and bibliometric analysis of food fraud from a global perspective. Journal of Agriculture and Food Research, 14, Article 100686. https://doi.org/10.1016/j.jafr.2023.100686.

5. Bayen, S., Elliott, C., Arlorio, M., Ballin, N.Z., Birse, N., Brockmeyer, J., Chahal, S., Corradini, M.G., Hanner, R., Hann, S., Laursen, K.H., Mihailova, A., Steininger-Mairinger, T., Suman, M., Tian, L., van Ruth, S., & Xia, J. (2024). Towards a harmonized approach for food authenticity marker validation and accreditation, Trends in Food Science & Technology, 149, Article 104550. https://doi.org/10.1016/j.tifs.2024.104550

6. Bimbo, F., Roselli, L., Carlucci, C., & de Gennaro, B.C. (2020). Consumer misuse of country-of-origin label: Insights from the Italian extra-virgin olive oil market. Nutrients, 12, Article 2150. https://doi.org/10.3390/nu12072150

7. Bontempo, L., Camin, F., Larcher, R., Nicolini, G., Perini, M., & Rossmann, A. (2009). Coast and year effect on H, O and C stable isotope ratios of Tyrrhenian and Adriatic Italian olive oils. Rapid Communications in Mass Spectrometry, 23, 1043-1048. https://doi.org/10.1002/rcm.3968

8. Bontempo, L., Paolini, M., Franceschi, P., Ziller, L., García-González, D.L., & Camin, F. (2019). Characterisation and attempted differentiation of European and extra-European olive oils using stable isotope ratio analysis. Food Chemistry, 276, 782-789. https://doi.org/10.1016/j.foodchem.2018.10.077

9. Brereton, R. (2006). Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data. Trends in Analytical Chemistry, 25, 1103-1111. https://doi.org/10.1016/j.trac.2006.10.005

10. Camin, F., Boner, M., Bontempo, L., Fauhl-Hassek, C., Kelly, S. D., Riedl, J., & Rossmann, A. (2017). Stable isotope techniques for verifying the declared geographical origin of food in legal cases. Trends in Food Science & Technology, 61, 176-187. https://doi.org/10.1016/j.tifs.2016.12.007

11. Camin, F., Larcher, R., Nicolini, G., Bontempo, L., Bertoldi, D., Perini, M., Schlicht, C., e Schellenberg, A., Thomas,F., Heinrich, K., Voerkelius, S., Horacek, M., Ueckermann, H., Froeschl, H., Wimmer, B., Heiss, G., Baxter, M., Rossmann, A., & Hoogewerff, J. (2010a). Isotopic and elemental data for tracing the origin of European olive oils. Journal of Agricultural and Food Chemistry, 58, 570-577. https://doi.org/10.1021/jf902814s

12. Camin, F., Larcher, R., Perini, M., Bontempo, L., Bertoldi, D., Gagliano, G., Nicolini, G., & Versini, G. (2010b). Characterisation of authentic Italian extra-virgin olive oils by stable isotope ratios of C, O and H and mineral composition. Food Chemistry, 118, 901-909. https://doi.org/10.1016/j.foodchem.2008.04.059

13. Camin, F., Pavone, A., Bontempo, L., Wehrens, R., Paolini, M., Faberi, A., Marianella, R.M., Capitani, D., Vista, S., & Mannina, L. (2016). The use of IRMS, ^{1}H NMR and chemical analysis to characterise Italian and imported Tunisian olive oils. Food Chemistry, 196, 98-105. https://doi.org/10.1016/j.foodchem.2015.08.132

14. Cappelli, L., D'Ascenzo, F., Natale, L., Rossetti, F., Ruggieri, R., & Vistocco, D. (2017) Are Consumers willing to pay more for a "Made in" product? An empirical investigation on "Made in Italy". Sustainability, 9, Article 556. https://doi.org/10.3390/su9040556

15. Carbone, A., & Henke, R. (2023) Recent trends in agri-food Made in Italy exports. Agricultural and Food Economics, 11, 32. https://doi.org/10.1186/s40100-023-00274-7

16. Casadei, E., Valli, E., Panni, F., Donarski, J., Farrús Gubern, J., Lucci, P., Conte, L., Lacoste, F., Maquet, A., Brereton, P., Bendini, A., & Toschi, T. G. (2021). Emerging trends in olive oil fraud and possible countermeasures. Food Control, 124, Article 107902. https://doi.org/10.1016/j.foodcont.2021.107902

17. Chiocchini, F., Portarena, S., Ciolfi, M., Brugnoli, E., & Lauteri, M. (2016) Isoscapes of carbon and oxygen stable isotope compositions in tracing authenticity and geographical origin of Italian extra-virgin olive oils. Food Chemistry, 202, 291–301. https://doi.org/10.1016/j.foodchem.2016.01.146

18. Climate Change Knowledge Portal for Development Practitioners and Policy Makers. Retrieved from https://climateknowledgeportal.worldbank.org/country. Accessed July 25, 2024.

19. Conte, L., Bendini, A., Valli, E., Lucci, P., Moret, S., Maquet, A., Lacoste, F., Brereton, P., García-González, D., Moreda, W., & Gallina Toschi, T. (2020). Olive oil quality and authenticity: A review of current EU legislation, standards, relevant methods of analyses, their drawbacks and recommendations for the future. Trends in Food Science & Technology, 105, 483–493. https://doi.org/10.1016/j.tifs.2019.02.025

20. Everstine, K.D., Chin, H.B., Lopes, F.A., & Moore, J.C. (2024) Database of food fraud records: Summary of data from 1980 to 2022. Journal of Food Protection, 87, Article 100227. https://doi.org/10.1016/j.jfp.2024.100227

21. Faberi, A., Marianella, R.M., Fuselli, F., La Mantia, A., Ciardiello, F., Montesano, C., Mascini, M., Sergi, M., & Compagnone D. (2014). Fatty acid composition and $\delta^{13}C$ of bulk and individual fatty acids as marker for authenticating Italian PDO/PGI extra virgin olive oils by means of isotopic ratio mass spectrometry. Journal of Mass Spectrometry, 49, 840–849. https://doi.org/10.1002/jms.3399

22. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27, 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

23. International Olive Council (IOC), (2023) World olive oil and table olive figures. Retrieved from https://www.internationaloliveoil.org/what-we-do/economic-affairs-promotion-unit/. Accessed July 5, 2024.

24. Jiménez-Morillo, N.T., Palma, V., Garcia, R., Barrocas Dias, C., & Cabrita, M.J. (2020). Combination of stable isotope analysis and chemometrics to discriminate geoclimatically and temporally the virgin olive oils from three Mediterranean countries. Foods, 9, Article 1855. https://doi.org/10.3390/foods9121855

25. Joint Research Center of the European Commission (JRC). Monthly summary of articles on food fraud and adulteration. Retrieved from https://ec.europa.eu/knowledge4policy/food-fraud-quality/topic/food-fraud_en. Accessed July 5, 2024.

26. Kavallari, A., Maas, S., & Schmitz, P.M. (2011). Examining the determinants of olive oil demand in nonproducing countries: Evidence from Germany and the UK. Journal of Food Products Marketing, 17, 355–372. https://doi.org/10.1080/10454446.2011.548721

27. Laursen, K.H., Bontempo, L., Camin, F., & Rossmann, A. (2016). Advances in Isotopic Analysis for Food Authenticity Testing. In G. Downey G. Woodhead Publishing Series in Food Science, Technology and Nutrition, Advances in Food Authenticity Testing, Woodhead Publishing, 227-252. https://doi.org/10.1016/B978-0-08-100220-9.00009-6

28. Lukić, I., Da Ros, A., Guella, G., Camin, F., Masuero, D., Mulinacci, N., Vrhovsek, U., & Mattivi, F. (2020). Lipid profiling and stable isotopic data analysis for differentiation of extra virgin olive oils based on their origin. Molecules, 25, Article 4. https://doi.org/10.3390/molecules25010004

29. Magnusson, B., & Örnemark, U. Eurachem guide: The Fitness for Purpose of Analytical Methods - A Laboratory Guide to Method Validation and Related Topics, second ed., Eurachem, Belgium, 2014. ISBN 978-91-87461-59-0. Retrieved from https://www.eurachem.org/images/stories/Guides/pdf/MV_guide_2nd_ed_EN.pdf Accessed July 5, 2024.

30. Maléchaux, A., Le Dréau, Y., Artaud, J., & Dupuy, D. (2020) Exploring the scientific interest for olive oil origin: A bibliometric study from 1991 to 2018. Foods, 9, Article 556. https://doi.org/10.3390/foods9050556

31. Nielsen, N. P. V., Carstensen, J. M., & Smedsgaard, J. (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation

optimised warping. Journal of Chromatography A, 805, 17–35. https://doi.org/10.1016/S0021-9673(98)00021-1

32. Portarena, S., Baldacchini, C., & Brugnoli, E. (2017). Geographical discrimination of extra-virgin olive oils from the Italian coasts by combining stable isotope data and carotenoid content within a multivariate analysis. Food Chemistry, 215, 1–6. https://doi.org/10.1016/J.FOODCHEM.2016.07.135

33. Portarena, S., Gavrichkova, O., Lauteri, M., & Brugnoli, E.(2014). Authentication and traceability of Italian extra-virgin olive oils by means of stable isotopes techniques. Food Chemistry, 164, 12-16. https://doi.org/10.1016/j.foodchem.2014.04.115

34. Quintanilla-Casas, B., Bertin, S., Leik, K., Bustamante, J., Guardiola, F., Valli, E., Bendini, A., Gallina Toschi, T., Tres, A., & Vichi, S. (2020a). Profiling versus fingerprinting analysis of sesquiterpene hydrocarbons for the geographical authentication of extra virgin olive oils. Food Chemistry, 307, Article 125556 https://doi.org/10.1016/j.foodchem.2019.125556

35. Quintanilla-Casas, B., Marin, M., Guardiola, F., García-González, D. L., Barbieri, S., Bendini, A., Gallina Toschi, T., Vichi, S., & Tres, A. (2020b). Supporting the sensory panel to grade virgin olive oils: an in-house-validated screening tool by volatile fingerprinting and chemometrics. Foods, 9, Article 1509. https://doi.org/10.3390/foods9101509

36. Quintanilla-Casas, B., Torres-Cobos, B., Guardiola, F., Romero, A., Tres, A., & Vichi, S. (2022a) Geographical authentication of virgin olive oil by GC-MS sesquiterpene hydrocarbon fingerprint: Scaling down to the verification of PDO compliance. Food Control, 139, Article 109055. https://doi.org/10.1016/j.foodcont.2022.109055.

37. Quintanilla-Casas, B., Torres-Cobos, B., Guardiola, F., Servili, M., Alonso-Salces, R.M., Valli, E., Bendini, A., Gallina Toschi, T., Vichi, S., & Tres, A. (2022b). Geographical authentication of virgin olive oil by GC–MS sesquiterpene hydrocarbon fingerprint: Verifying EU and single country label-declaration. Food Chemistry, 378, Article 132104. https://doi.org/10.1016/j.foodchem.2022.132104.

38. Commission Delegated Regulation (EU) 2022/2104 of 29 July 2022 supplementing Regulation (EU) No 1308/2013 of the European Parliament and of the Council as regards marketing standards for olive oil, and repealing Commission Regulation (EEC) No 2568/91 and Commission Implementing Regulation (EU) No 29/2012, pp. 1–22.

39. Regulation (EU) No 2024/1143 on geographical indications for wine, spirit drinks and agricultural products, as well as traditional specialties guaranteed and optional quality terms for agricultural products, pp. 1–76.

40. Riedl, J., Esslinger, S., & Fauhl-Hassek, C. (2015). Review of validation and reporting of non-targeted fingerprinting approaches for food authentication. Analytica Chimica Acta, 885, 17-32. https://doi.org/10.1016/j.aca.2015.06.003

41. Spink, J., & Moyer, D.C. Defining the public Health threat of food fraud. (2011). Journal of Food Science, 76, 157-163. https://doi.org/10.1111/j.1750-3841.2011.02417.x

42. The EU Food Fraud Network and the System for Administrative Assistance & Food Fraud. Annual Reports 2014-2020. Retrieved from https://food.ec.europa.eu/safety/acn/reports-and-publications_en#agri-food-fraud-network-ffn. Accessed July 5, 2024.

43. Torres-Cobos, B., Quintanilla-Casas, B., Romero, A., Ninot, A., Alonso-Salces, R. M., Gallina Toschi, T., Bendini, A., Guardiola, F., Tres, A., & Vichi, S. (2021). Varietal authentication of virgin olive oil: Proving the efficiency of sesquiterpene fingerprinting for Mediterranean Arbequina oils. Food Control, 128, Article 108200. https://doi.org/10.1016/j.foodcont.2021.108200

44. Vichi, S., Guadayol, J. M., Caixach, J., López-Tamames, E., & Buxaderas, S. (2006). Monoterpene and sesquiterpene hydrocarbons of virgin olive oil by headspace solid phase microextraction coupled to gas chromatography/mass spectrometry. Journal of Chromatography A, 1125, 117–123. https://doi.org/10.1016/j.chroma.2006.05.029

45. Vichi, S., Lazzez, A., Grati-Kamoun, N., Lopez Tamames, E., & Buxaderas, S. (2010). Evolution of sesquiterpene hydrocarbons in virgin olive oil during fruit ripening. Journal of Agricultural and Food Chemistry, 58, 6972–6976. https://doi.org/10.1021/jf100497c

46. Vichi, S., Tres, A., Quintanilla-Casas, B., Bustamante, J., & Guardiola, F. (2018). Sesquiterpene hydrocarbons, a promising tool for virgin olive oil geographical authentication. In M. Kontominas (Ed.), Authentication and detection of adulteration of olive oil (chapter 11) New York, NY: Nova Science Publishers, Inc ISBN: 978-1-53614-596-0.

47. Yan, J., Erasmus, S.W., Aguilera Toro, M., Huanga, H., & van Ruth, S.M. (2020) Food fraud: Assessing fraud vulnerability in the extra virgin olive oil supply chain. Food Control, 111, Article 107081. https://doi.org/10.1016/j.foodcont.2019.107081

48. Zaroual, H., Chénè, C., El Hadrami, E. M., & Karoui, R. (2021). Application of new emerging techniques in combination with classical methods for the determination of the quality and

authenticity of olive oil: a review. Critical Reviews in Food Science and Nutrition, 62, 4526–4549. https://doi.org/10.1080/10408398.2021.1876624

4.2 Virgin olive oil varietal authentication

4.2.1 Sesquiterpene fingerprinting

To evaluate the suitability of SH fingerprint as a tool for VOO varietal authentication beyond geographical and technological factors, it should be tested on a sample set that reflects the real VOO production, where the same olive cultivar can be grown in various regions [42].

Bearing this in mind, a sample set of 404 VOOs from 4 consecutive campaigns (2015-2019) and various regions of origin, produced under real industrial conditions, was analysed by HS-SPME-GC-MS to obtain the SH fingerprint (**Figure 8**). Of these samples:

- 178 were oils of the 'Arbequina' cultivar produced in different countries,

- 226 where oils from 37 other cultivars and their coupages (blends of cultivars).

A binary PLS-DA model was built using the SH unfolded matrix to discriminate between 'Arbequina' VOOs from 'non-Arbequina' VOOs, which included all the other cultivars. Models were externally validated by predicting the class of the samples in the validation set (20% of each class of the full sample set).

Finally, an exploration of the regression coefficients of the model was carried out to investigate which variables contributed the most to the cultivar discrimination.

Furthermore, a preliminary multi-class PLS-DA model (n = 256) was built to distinguish among 7 different cultivars (represented by at least 10 samples), to prospectively evaluate the capability of the SH fingerprint to differentiate other cultivars (Publication 3).

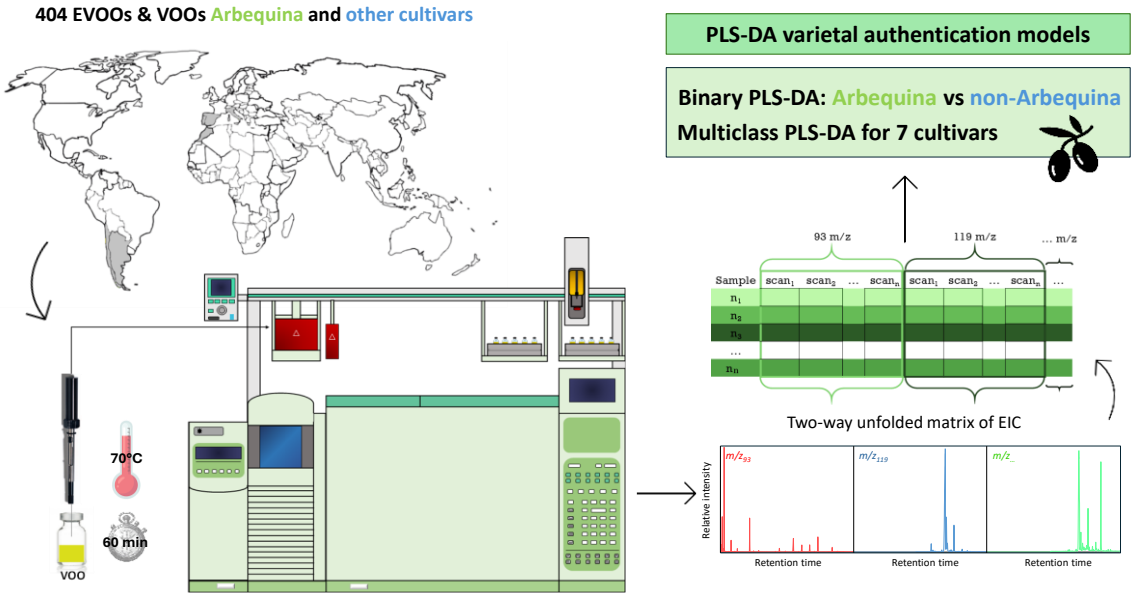


Figure 8. Graphical abstract of Publication 3. EVOO: extra virgin olive oil, VOO: virgin olive oil, EIC: extracted ion chromatogram, PLS-DA: partial least square-discriminant analysis.

4.2.2 Publication 3

Varietal authentication of virgin olive oil: Proving the efficiency of sesquiterpene fingerprinting for Mediterranean Arbequina oils

Berta Torres-Cobos, Beatriz Quintanilla-Casas, Agustí Romero , Antonia Ninot, Rosa M. Alonso-Salces, Tullia Gallina Toschi, Alessandra Bendini, Francesc Guardiola, Alba Tres, Stefania Vichi

*Food Control, **2021**, 128, 108200. https://doi.org/10.1016/j.foodcont.2021.108200*

Supplementary material available in **Annex 2**

Food Control 128 (2021) 108200

Contents lists available at ScienceDirect

Food Control

journal homepage: www.elsevier.com/locate/foodcont

Varietal authentication of virgin olive oil: Proving the efficiency of sesquiterpene fingerprinting for Mediterranean Arbequina oils

Berta Torres-Cobos [a,b], Beatriz Quintanilla-Casas [a,b], Agustí Romero [c], Antonia Ninot [c], Rosa M. Alonso-Salces [d], Tullia Gallina Toschi [e], Alessandra Bendini [e], Francesc Guardiola [a,b], Alba Tres [a,b,*], Stefania Vichi [a,b]

[a] Departament de Nutrició, Ciències de L'Alimentació I Gastronomia, Campus de L'Alimentació Torribera, Facultat de Farmàcia I Ciències de L'Alimentació, Universitat de Barcelona. Av Prat de La Riba, 171, 08921, Santa Coloma de Gramenet, Spain
[b] Institut de Recerca en Nutrició I Seguretat Alimentària (INSA-UB), Universitat de Barcelona. Av Prat de La Riba, 171, 08921, Santa Coloma de Gramenet, Spain
[c] IRTA-Mas Bové Ctra, Reus-El Morell Km 3,8 43120-Constantí, Spain
[d] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Departamento de Biología, Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Mar Del Plata (UNMdP), Funes 3350, 7600, Mar Del Plata, Argentina
[e] Dipartimento di Scienze e Tecnologie Agro-alimentari - Centro Interdipartimentale di Ricerca Industriale (CIRI Agroalimentare), Alma Mater Studiorum - Università di Bologna, Bologna, Italy

ARTICLE INFO

Keywords:
Virgin olive oil
Cultivar
Sesquiterpene
Fingerprint
HS-SPME-GC-MS
PLS-DA

ABSTRACT

Virgin olive oil (VOO) is a highly appreciated product fundamental in the Mediterranean diet. Since its sensory attributes are greatly influenced by the olive cultivar, the varietal authentication of VOOs is needed to protect consumers from misleading information. The present study aims to evaluate the suitability of sesquiterpene hydrocarbon (SH) fingerprint as VOO cultivar marker beyond geographical, agronomical and processing conditions. The study was mainly focused on Mediterranean Arbequina oils. SH profile of more than 400 VOOs from 6 counties and 38 different cultivars and coupages was analysed by Headspace Solid Phase Microextraction-Gas Chromatography-Mass Spectrometry (HS-SPME-GC-MS). Partial Least Square-Discriminant Analysis (PLS-DA) classification models were built with the aligned chromatograms. A binary PLS-DA model was built to distinguish 'Arbequina' oils from those of other cultivars (non-'Arbequina' class) and it was externally validated. The results of the external validation showed a 95.1% of overall correct classification confirming the suitability of SH fingerprint as a screening method for the authentication of Arbequina VOO. Also, the discrimination capacity of SH fingerprinting to authenticate VOOs from other cultivars was preliminary explored and promising results were obtained.

1. Introduction

Qualitative characteristics of virgin olive oil (VOO) are not only determined by the processing and storage conditions but also by the cultivar and geographical origin (Dias et al., 2014; Montealegre et al., 2010). Hence, oils within the same commercial category can present very different compositional and sensory characteristics depending on the olive cultivar. Monovarietal VOOs that include information about the olive cultivar on the label, as well as VOOs included into a given Protected Designation of Origin which are required to use specific traditional olive cultivars (Council Regulation (EC) 510/2006), increase consumers' perceived quality and lead them to pay a higher price (Cabrera et al., 2015; Cicerale et al., 2016). Therefore, verifying the label-declared cultivar in VOO has become relevant to protect consumers from misleading information (Bajoub et al., 2018) and, currently, it can only be achieved by auditing traceability documents.

The varietal characterization of VOOs has been widely studied by addressing several major and minor compounds and by applying multiple analytical techniques and chemometric approaches (Montealegre et al., 2010; Aparicio et al., 2013; Bajoub et al., 2018). However, except DNA based methods (Agrimonti et al., 2011), which are costly to be used for routine analysis (Bajoub et al., 2018), reliable markers for VOO cultivar authentication are still unavailable. This is largely due to the

* Corresponding author. Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Campus de l'Alimentació Torribera, Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona. Av Prat de la Riba, 171. Edifici Gaudí, 08921, Santa Coloma de Gramenet, Spain.
E-mail address: atres@ub.edu (A. Tres).

https://doi.org/10.1016/j.foodcont.2021.108200
Received 29 November 2020; Received in revised form 22 April 2021; Accepted 23 April 2021
Available online 28 April 2021

Abbreviations

COW	Correlation Optimized Warping
EIC	Extracted Ion Chromatogram
LV	Latent Variables
PC	Principal Component
PCA	Principal Component Analysis
PLS-DA	Partial Least Discriminant Analysis
RMSEcv	Root Mean Squared Error of Cross Validation
SEcv	Standard Error of Cross-Validation
SH	sesquiterpene hydrocarbon
SIM	Selected Ion Monitoring
SPME-GC-MS	Solid Phase Microextraction coupled to Gas Chromatography-Mass Spectrometry
VOO	Virgin olive oil

high concomitant influence of other factors on VOO composition, such as the degree of maturation, phytosanitary status, harvest year or processing and storage conditions (García-González & Aparicio, 2010; Montealegre et al., 2010; Vichi et al., 2018). In this regard, most of the available studies on VOO cultivar authentication are addressed to monovarietal VOOs produced within a limited geographical area (Osorio-Bueno et al., 2005; Papadia et al., 2011; Piccinonna et al., 2016; Sacco et al., 2000; Sayago et al., 2019). They have been fundamental to demonstrate the genetic influence on VOO composition with other factors being similar, but they do not take into account the effect of technological and, in particular, environmental conditions beyond the olive cultivar itself. This is not fully representative for the actual VOO production, in which the same olive cultivar may be cultivated in different regions or countries (Tous, 2017). Marini et al. (2004) developed authentication models relying on a representative number of samples (n > 500) from cultivars produced in different regions of Southern Italy, achieving a satisfactory classification by applying supervised chemometric methods to several official quality and purity parameters. However, the need for multiple analytical techniques or time-expensive procedures hinders its application for the screening of large number of samples, and moreover, its suitability should be confirmed also for VOOs from other major production areas. Hence, there is a need for screening tools capable of authenticating VOOs according to their cultivar beyond geographical and technological factors.

Previous research has evidenced that the presence of sesquiterpene hydrocarbons (SHs) in VOO is highly dependent the olive cultivar and the growing area (Bortolomeazzi et al., 2001; Damascelli & Palmisano, 2013; Quintanilla-Casas et al., 2020; Vichi et al., 2006, 2010, 2018) while it is barely affected by technological factors such as olive post-harvest processing and oil storage conditions (Vichi et al., 2018). Specifically, the effect of genetic factors on VOO SHs was demonstrated when oils from different cultivars, produced in the same geographical region, presented significant differences in the SH composition (Guinda et al., 1996; Osorio-Bueno et al., 2005; Vichi et al., 2010).

Also recently, a fingerprinting approach has been applied to SH chromatograms obtained by SPME-GC-MS (Quintanilla-Casas et al., 2020). In a fingerprinting approach, chemometric tools are applied to highly dimensional analytical data, such as a chromatogram, to find specific patterns of a certain quality characteristic that are known as fingerprints, not requiring peak identification or quantitation (Ballin & Laursen, 2019; Berrueta et al., 2007; Bosque-Sendra et al., 2012; Quintanilla-Casas et al., 2020). Compared to the target approach, chromatographic fingerprinting considers more information as the whole analytical signal is used, and overcomes drawbacks related to the SH identification or quantitation as they are a wide category of compounds with high structural diversity but very similar mass spectra (Degenhardt et al., 2009; Quintanilla-Casas et al., 2020). This was demonstrated in our previous study, in which Partial Least Discriminant Analysis (PLS-DA) was able to find features in the SH fingerprint that were common between samples from the same region even if they belonged to different cultivars, and thus, models to authenticate VOO geographical origin could be developed by selecting the country of origin as the grouping variable to supervise the PLS-DA (Quintanilla-Casas et al., 2020). But, since the SH composition in olive fruit is known to be driven by both genetic and environmental factors, our hypothesis was that if the cultivar type was selected as the variable to supervise the PLS-DA analysis, the PLS-DA model would find different features on the SH fingerprint of VOO that would be characteristic of the cultivar.

On this basis, the aim of the present work was to assess the suitability of SH fingerprint as VOO cultivar marker beyond geographical, agronomical and processing conditions. For this purpose, we developed and validated a varietal authentication model based on the SH fingerprint obtained by SPME-GC–MS, particularly focusing on the discrimination of 'Arbequina' VOOs from the rest of VOOs, using a sample set of more than 400 VOOs from different cultivars produced under real processing conditions in different harvest seasons and in various EU and non-EU countries and regions. Moreover, to explore whether analogous models could also be developed to distinguish VOOs from other cultivars, the discrimination capacity of SH fingerprinting was preliminarily assessed for the seven main cultivars included in the sample set.

2. Material and methods

2.1. Sampling

The sample set was composed by 404 traceable VOOs and EVOOs from different countries and geographical regions (Table 1). They were obtained in the framework of the Projects OLEUM (EC H2020 Programme 2014–2020) and Autenfood (ACCIÓ- Programa Operatiu FEDER Catalunya 2014–2020), and under the surveys implemented by the Institut de Recerca i Tecnologia Agroalimentària (IRTA). Of these samples, 178 were from 'Arbequina' cultivar and 226 were monovarietal oils from 37 other cultivars (n = 144) and coupages (which are blends of different cultivars) that did not contain 'Arbequina' oil (n = 82). The VOO and EVOO samples pertained to virgin and extra virgin olive oil categories according to the European Commission regulation (ECC) No 2591/91 of July 11, 1991 and its amendments and were produced at real industrial conditions during 4 different campaigns (harvests from 2015/16 to 2018/19). More information about VOO and EVOO samples is available in Table S1 (Supplementary material). Samples were stored under N_2 atmosphere at −20 °C until analysis. The full sample set was analysed in four main batches throughout 2017–2019.

2.2. Headspace-solid phase microextraction (HS-SPME)

The SH fingerprint of VOO samples was analysed using a Combi-pal autosampler (CTC Analytics, Zwingen, Switzerland) at the conditions reported by Vichi et al. (2006). An aliquot of 2 g of oil was weighed into

Table 1
Number and geographical origin of VOO and EVOO samples from 'Arbequina' and non-'Arbequina' cultivars.

Origin		'Arbequina' (n)	Other cultivars[a] (n)
Argentina		4	37
Chile		1	1
Italy		1	35
Morocco		9	29
Portugal		3	13
Spain	Catalonia	155	93
	Andalusia	2	15
	Other regions	3	3
total samples		178	226

[a] 37 different cultivars plus 20 coupages.

a 10 mL vial fitted with a PTFE/silicone septum and kept at 70 °C under constant agitation. After 10 min of sample conditioning, a divinylbenzene/carboxen/polydimethylsiloxane (DVB/CAR/PDMS) fiber (2 cm length, 50/30 μm film thickness) provided by Supelco (Bellefonte, PA) was exposed to the sample headspace for 60 min. Then, it was desorbed in the gas chromatograph injection port at 260 °C for 10 min. During the desorption step, the injector was maintained in split-less mode during 5 min.

2.3. Gas chromatography-mass spectrometry (GC-MS)

The SHs fingerprint was acquired by an Agilent 6890N Network GC system coupled to a quadrupolar mass selective analyser Agilent 5975C Inert MSD (Agilent Technologies, Santa Clara, California, USA) using helium as carrier gas, at a flow of 1.5 mL/min. Analytes were separated on a Supelcowax-10 capillary column (60 m × 0.25 mm i.d., 0.25 μm film thickness) (Supelco, Bellefonte, PA). Column temperature was held at 40 °C for 3 min, increased to 100 °C at 4 °C/min, then to 200 °C at 5 °C/min and to 260 °C at 15 °C/min, holding the last temperature for 5 min. The temperatures of the ion source and the transfer line were 230 and 280 °C, respectively. Mass spectra were recorded at 5.1 scan/s and the electron energy was 70 eV. Acquisition was performed in the selected ion monitoring (SIM) mode, by analysing the Extracted Ion Chromatogram (EIC) of each of the following SH specific ions: *m/z* 93, 119, 157, 159, 161, 189 and 204, which had been reported to be specific for SHs (Vichi et al., 2010). Therefore, only the chromatographic data belonging to SH compounds was studied.

2.4. Fingerprinting approach

The intensities of scans between minutes 21 and 42 (3197 scans) were considered for each ion (3197 scans × 7 ions = 22379 variables per sample). A data matrix was built for each ion, with scans' intensities of each EIC (columns) from each sample (rows). For each selected ion, the EICs of the 404 samples were aligned by Correlation Optimized Warping (COW) algorithm in Matlab® (Nielsen et al., 1998). Then, the 7 matrices of the aligned chromatograms were concatenated conforming a two-way unfolded matrix (404 samples × 22379 variables).

2.5. Chemometrics

2.5.1. Data pre-processing and exploration

The pre-processing of the aligned data matrix was performed with SIMCA software v13.0© (Umetrics AB, Sweden). Multiple pre-processing treatments were tested (mean centring, scaling to unit variance, log10, derivatives) until finding the optimal one for each data set. A Principal Component Analysis (PCA) was performed for the exploration of data (n = 404) and to identify potential outliers (according to Hotelling's T^2 range and distance to the model parameters).

2.5.2. Partial least squares discriminant analysis (PLS-DA): Arbequina vs non-Arbequina oils

A binary PLS-DA model was built after applying a first derivative and log 10 pre-processing to classify the 404 samples into the 'Arbequina' and the non-'Arbequina' classes, the latter including coupages and monovarietal oils from other cultivars,.

The model was internally validated through leave 10%-out cross-validation. The optimal number of Latent Variables (LV) of the PLS-DA model were selected according to the lowest Root Mean Squared Error of Cross Validation (RMSEcv) criteria. To assess the model over-fitting, permutation test and ANOVA on the cross-validated predictive residuals (p-value) were carried out. The Q^2 values and the percentage of correct classifications were assessed to evaluate the suitability of each PLS-DA model.

2.5.3. External validation

The binary PLS-DA model was then externally validated by predicting the class of samples that had not been used to develop the model. For this, the full data set (n = 404) was randomly split into a training set (80% of the sample set, n = 323) and a validation set (20% of the sample set, n = 81), maintaining a balance in the proportions of 'Arbequina' and non-'Arbequina' samples and of samples from different analytical batches and geographical origins (Table S2, Supplementary material). This was carried out seven times, obtaining seven different training and validation sets. The efficiency of the classification was assessed as mean percentage of correct classification.

2.5.4. Evaluation of PLS-DA regression coefficients

The regression coefficients of the binary PLS-DA model developed with the full sample set were evaluated to explore the contribution of the variables obtained by each *m/z* ion. Regression coefficients were considered as significant when a jack-knife standard error of cross-validation (SEcv) was lower than the given coefficient value.

2.5.5. Multi-class PLS-DA model for 7 cultivars

Once the Arbequina vs non-Arbequina model was found suitable, it was explored whether SHs fingerprinting would be suitable to develop models to verify the identity of other cultivars. Thus, a multi-class PLS-DA model was developed as a preliminary model (after autoscaling, log10 and first derivative pre-processing) to evaluate the discrimination of the 7 cultivars included in the sample set that were represented by at least 10 samples (n = 256). The model was fitted and internally validated as described above.

3. Results and discussion

3.1. Data pre-processing and exploratory analysis

Seven ions (*m/z* 93, 119, 157, 159, 161, 189 and 204) were selected according to previous studies that indicated them to be specific of SH (Vichi et al., 2010). The data points of the EIC within the interval of elution of SH (21–42 min) were used as variables (22379 variables), following the fingerprinting approach applied by Quintanilla-Casas et al. (2020).

To solve the retention time shifting between samples, the EICs were aligned by Correlation Optimized Warping (COW) algorithm in Matlab®. This alignment algorithm was selected since it was specifically designed for chromatographic data. The COW method aligns the chromatographic profiles by piecewise linear stretching and compression, also known as warping, of the time axis of one of the profiles (Nielsen et al., 1998). Once aligned, the matrix obtained (22379 variables and 404 samples) was imported to SIMCA software v13.0c (Umetrics AB, Sweden) to develop and optimize the classification models.

After data pre-processing, a PCA was performed to explore the data and to detect potential outliers (3 Principal Components (PCs) accounted for 80.7% of the total variance explained). According to the Hotelling's T^2 range and distance to the model parameters, no outliers were detected. A first examination of the PCA score plot revealed that even under a non-supervised analysis, VOOs naturally clustered according to their cultivar. In fact, even if there was some overlap, 'Arbequina' samples tended to shape into a differentiated group at the upper part of the plot, while non-'Arbequina' samples located at the lower part (Fig. 1a), evidencing that the variability linked to the cultivar of origin was mainly explained by PC3 (2.0% of total explained variance). Certain clustering was also observed for other monovarietal VOOs (Fig. 1b), endorsing the behaviour observed for 'Arbequina' and non-'Arbequina' samples. To exclude that this natural clustering was only induced by the geographical origin of each cultivar, the score plot was coloured by sample's provenance (Fig. 1c). In this way, it revealed that samples from cultivars grown in different regions tended to group by cultivar rather than by geographic origin, leading to unclear clusters by country. This

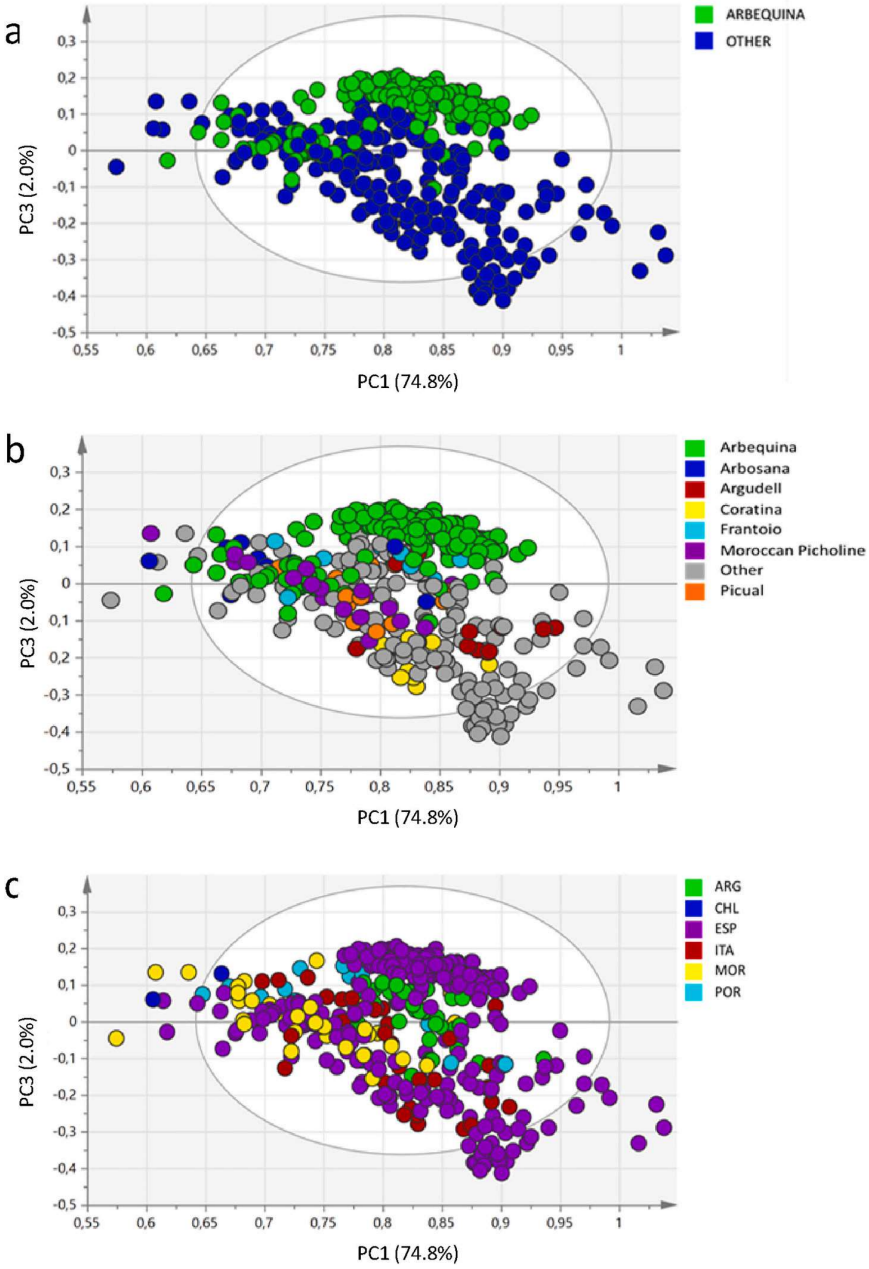


Fig. 1. First and third principal components of the PCA model (n = 404 samples, first derivative and log 10 pre-processing; variance explained by each factor is shown between parentheses), based on VOO sesquiterpene data, coloured by a) 'Arbequina'/non-'Arbequina' cultivars; b) Olive cultivar; c) Country of origin. (ARG: Argentina, CHL: Chile, ESP: Spain, ITA: Italy, MOR: Morocco and POR: Portugal).

suggested that the cultivar had a higher effect on the SH fingerprint than the geographical origin.

3.2. PLS-DA authentication models and internal validation

A binary PLS-DA classification model was built with the SH fingerprint of 404 samples to distinguish 'Arbequina' oils from those of other cultivars (non-'Arbequina' class). The model with 5 LV was internally validated through leave 10%-out cross-validation, resulting in a RMSEcv of 0.23 and a 99.8% of correct classification with all samples assigned to the correct class, except one (Table 2). Permutation tests, which display the prediction capacity of 20 random models, and ANOVA results ($p <$

Table 2
Results of the leave 10%-out cross-validation of the 'Arbequina' *vs* non-'Arbequina' PLS-DA classification model.

	N	'Arbequina'	Other	Correct class (%)
'Arbequina'	178	178	0	100
non-'Arbequina'	226	1	225	99.6
Total	404			**99.8**

N = 404, 5 LVs, $Q^2 = 0.82$, RMSEcv = 0.23, ANOVA p-value <0.05.

0.05) showed that the model had a high discrimination capacity and was not over-fitted. The results were very promising, especially considering that the 'Arbequina' class included samples from various geographical origins, and that the non-'Arbequina' class included a high number of other monovarietal and coupage oils.

In view of the encouraging results provided by the binary model based on the extensive 'Arbequina' and non-'Arbequina' sampling, and since the PCA exploratory analysis already revealed certain natural separation between other cultivars, we investigated whether analogous models could potentially be developed to authenticate VOOs from other cultivars. As a proof of concept for this scope, we performed a preliminary multi-class PLS-DA model to discriminate between the seven cultivars of the sample set that were represented by at least 10 samples. Only monovarietal samples were included in this PLS-DA model (n = 256). The results of the leave 10%-out cross-validation were promising (Table 3), displaying high percentages of correct classification (generally above 90%) for most of the cultivars and achieving a global 94.9% of correct classification. In addition, the ANOVA ($p < 0.05$) and the permutation test (Q^2 values of permuted models < 0) indicated the absence of a random classification and of model overfitting. The lowest percentage of correct classification was observed for the 'Frantoio' set (69.2%), probably due to the very different geographical origin of its samples (Italy, n = 3 and Argentina, n = 9). The low number of samples of this preliminary submodel might have not been enough to compensate such a source of confusion, but we can hypothesize that with a suitable sampling the authentication of this cultivar would not be precluded. Overall, based on these preliminary results, we could infer that developing future models to authenticate VOOs from any cultivar based on the SH fingerprint would be possible, providing that appropriate sampling was available.

3.3. External validation

Disposing of meaningful results is crucial for the relevant implementation of any authentication tool. In this case, verifying the real predictive ability of the developed model is necessary to exclude possible over-optimistic results and to prove the suitability of the method to assess VOO varietal origin. To confirm the reliability of the predictions obtained by internal validation, we carried out an external validation in which models were applied to samples that had not been included in their development to classify them as 'Arbequina' and non-'Arbequina' VOOs. For this, we randomly split the sample set into a training set (n = 323) and a validation set (n = 81). A PLS-DA model was developed with the training set and cross-validated by leave-10%-out, and it was then applied to predict the class of the 81 samples conforming the validation set. To increase the robustness of the validation, and to minimize the effect of the sample sets' composition, this process was run seven times and the results were expressed as mean values of the seven sets of external validation. The internal validation results of PLS-DA models built with the training sets were in agreement with those of the model developed with the global sample set (mean overall correct classification of 100%). The results of the external validation, summarized in Table 4, were extremely satisfactory. On average, 93.5% of 'Arbequina' samples and 96.4% of non-'Arbequina' samples were correctly classified, resulting in a 95.1% of overall correct classification.

On closer inspection, the external validation revealed that the model was not able to correctly classify 'Arbequina' oils produced in Argentina. Unfortunately, due to the reduced number of 'Arbequina' VOOs from Argentina, only 3 samples could be randomly included in each training set and only one in each validation set, while other non-'Arbequina' oils from Argentina were much more represented in both sets. This probably contributed to the fact that each Argentinian 'Arbequina' sample, randomly selected in each validation set, was incorrectly classified (Table 5). This seems to support our previous hypothesis formulated for the internal validation results obtained for the 'Frantoio' cultivar by the preliminary multi-class PLS-DA model. This outcome could be due to the extreme compositional differences reported between VOOs from the southern hemisphere and those from the Mediterranean area. Romero (2017) reported that 'Arbequina' VOO from Argentina and Australia has a very different fatty acid composition from other regions and discussed about the high temperature effect during the maturation season. In fact, the southern hemisphere 'Arbequina' oils may present compositions whose differences may even be outside of some of the limits set by the current trade standards (Aparicio et al., 2013). Apparently, for these samples, the geographical origin outperformed the cultivar differences (Rondanini et al., 2011; Torres et al., 2009), possibly leading to a better matching of the SH pattern of Argentinian 'Arbequina' oils with that of other Argentinian oils rather than with that of the rest of 'Arbequina' Mediterranean samples, even if the cultivar was the classification variable. Since only a scarce number of Argentinian 'Arbequina' samples was available (n = 4), the results are not conclusive in this aspect. Varietal misclassification of samples from non-Mediterranean region could be presumably resolved by widening the sampling with a representative number of VOOs from the region of interest. By this, it would be possible to determine if the model would be able to find proper common SH traits between Mediterranean and Argentinian 'Arbequina' oils, or if a specific model 'Arbequina' vs non-'Arbequina' would be required for Argentinian oils.

Even so, samples from different EU and non-EU Mediterranean

Table 4
Results of external validation of the 'Arbequina' *vs* non-'Arbequina' PLS-DA models. Results are mean values obtained from seven randomly selected validation sets.

	n	Arbequina	Other	Correct class (%)
'Arbequina'	36	34 ± 1	2 ± 2	93.5 ± 3.8
non-'Arbequina'	45	2 ± 1	43 ± 2	96.4 ± 1.3
Total	81			95.1 ± 2.4

N = 404, 5 LVs, $Q^2 > 0.80$, RMSEcv < 0.27, ANOVA p-value <0.05.

Table 3
Results of leave 10%-out cross-validation of the multi-class PLS-DA model for the classification of main cultivars in the sample set.

	N	'Arbequina'	'Picual'	'Arbosana'	'Moroccan Picholine'	'Coratina'	'Frantoio'	'Argudell'	No class[a]	Correct class (%)	RMSEcv
'Arbequina'	178	173	0	0	1	0	0	0	4	97.2	0.30
'Picual'	12	0	12	0	0	0	0	0	0	100	0.19
'Arbosana'	11	0	0	10	0	0	0	0	1	90.9	0.16
'Moroccan Picholine'	20	1	0	0	19	0	0	0	0	95.0	0.19
'Coratina'	10	0	0	0	0	10	0	0	0	100	0.15
'Frantoio'	13	0	0	0	0	0	9	0	4	69.2	0.20
'Argudell'	12	2	0	0	0	0	0	10	0	83.3	0.17
Total	256	176	12	10	20	10	9	10	9	**94.9**	

N = 256, 7 LVs, $Q^2 = 0.47$, ANOVA p-value <0.05.
[a] YPred<0.5.

Table 5
Number, country and cultivar of origin of samples misclassified in the external validation of the binary PLS-DA model, for each validation set. The number of misclassified samples is reported with respect to the total number of samples from the same country and class ('Arbequina' and non-'Arbequina') that were included in the validation set.

	Validation set	Number of misclassified samples	Country of origin of the misclassified samples	Cultivar of the misclassified samples
'Arbequina' samples classified as non-'Arbequina'	1	1/1	Argentina	'Arbequina'
	2	1/1	Argentina	'Arbequina'
		1/32	Spain	'Arbequina'
	3	1/1	Argentina	'Arbequina'
	4	1/1	Argentina	'Arbequina'
		1/32	Spain	'Arbequina'
	5	1/1	Argentina	'Arbequina'
		1/32	Spain	'Arbequina'
	6	1/1	Argentina	'Arbequina'
		1/32	Spain	'Arbequina'
	7	1/1	Argentina	'Arbequina'
Non-'Arbequina' samples classified as 'Arbequina'	1	1/22	Spain	'Empeltre'
	2	1/7	Italy	'Casaliva'
		1/22	Spain	'Empeltre'
	3	1/3	Portugal	Coupage
		1/22	Spain	'Empeltre'
	4	2/22	Spain	'Empeltre'
	5	1/7	Italy	'Casaliva'
		1/7	Italy	Coupage
		3/22	Spain	'Empeltre'
	6	1/6	Morocco	'Moroccan Picholine'
		3/22	Spain	'Empeltre'
	7	1/7	Italy	Coupage

regions were correctly assigned to their corresponding varietal class, even when their representativeness in the training set was low, such is the case of the Portuguese and Moroccan 'Arbequina' oils (Table 1). These results evidenced the suitability of the model for Mediterranean VOO cultivar authentication regardless of the oil's geographical origin. Although the present sampling was specially focused on Spanish 'Arbequina' oils, the results of the external validation proved the efficiency of the authentication model to distinguish 'Arbequina' oils produced also in the rest of the Mediterranean basin. To improve the predictive ability of the model for 'Arbequina' samples coming from other specific geographical origins, new samples from those regions should be included in the training set.

3.4. Exploration of PLS-DA regression coefficients

To study the variables that contributed the most to the discrimination between 'Arbequina' and other cultivar VOOs, we examined the significant regression coefficients of the 'Arbequina' vs non-'Arbequina' PLS-DA model. More than 200 variables resulted relevant to the model according to the significance of their regression coefficients (as defined according to their value and standard error), highlighting the key advantage of high-dimensional data approaches, such as fingerprinting, over other conventional approaches such as the target or even multi-target ones (Quintanilla-Casas et al., 2020).

The plotting of regression coefficients against the variables of the unfolded matrix (Fig. 2a) revealed that each EIC provided relevant variables to the model throughout different regions of the chromatogram. In particular, EICs of *m/z* 93, 119 and 204 provided the highest number of variables important for the discrimination of both the 'Arbequina' and the non-'Arbequina' classes. The same figure revealed that significant regression coefficients corresponded to sections of the unfolded matrix presenting either major and very minor variables. As an illustration of this, amplifying a section of the EIC of *m/z* 93 (34–38 min) (Fig. 2b), revealed that some of the highest regression coefficients corresponded to minor SHs or not well-resolved chromatographic peaks that could only be studied through a fingerprinting approach. This is in agreement with our previous findings dealing with the application of SH fingerprinting for VOO geographical authentication (Quintanilla-Casas et al., 2020).

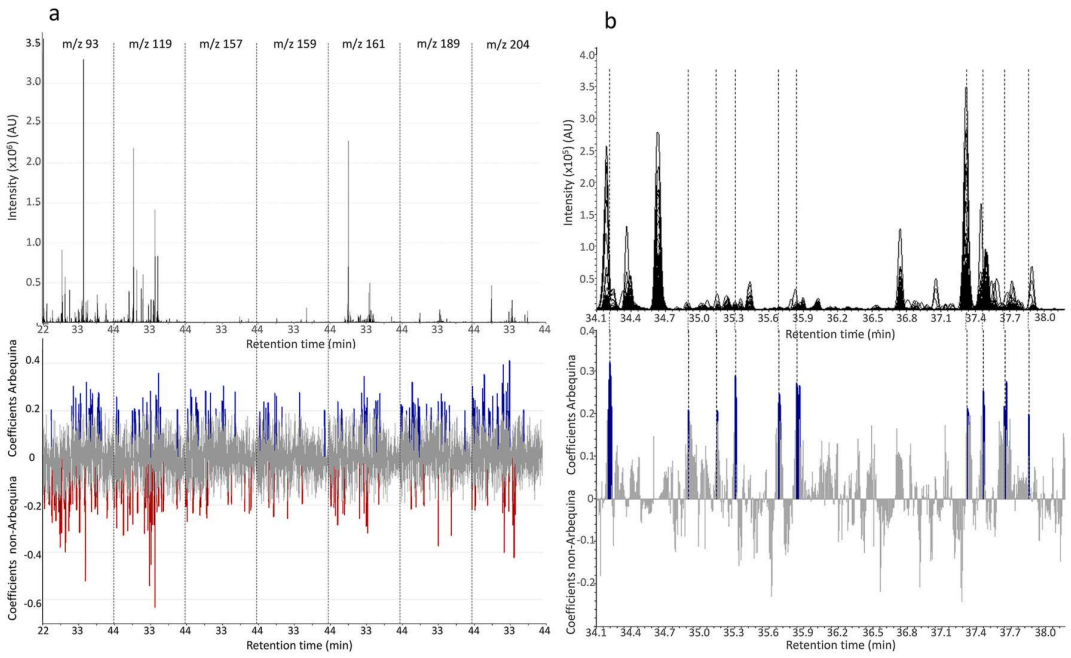


Fig. 2. PLS-DA regression coefficients of the Arbequina *vs* non-'Arbequina' model against the variables of the unfolded matrix. 'Arbequina' positive and negative relevant coefficients (selected from the significant ones according to an arbitrary threshold of 0.2) are highlighted in blue or in red, respectively. a) The sections of the unfolded matrix corresponding to each Extracted Ion Chromatogram are marked. b) Section of the ion *m/z* 93 chromatogram (34.059–38.126 min) plotted against its corresponding regression coefficients. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

4. Conclusions

In conclusion, SH fingerprint proved to be a suitable screening method for the varietal authentication of VOO, enabling an extremely satisfactory efficiency in the classification of 'Arbequina' and non-'Arbequina' oils produced under real industrial conditions in different regions of the Mediterranean basin, as assessed by external validation (95.1% of correct classification). Combining SH fingerprinting with PLS-DA allowed recognizing characteristic patterns relevant for each cultivar class minimizing those variables related with other factors such as the geographical origin. This confirmed our hypothesis that genetic and environmental factors exert distinct effects on the particularly complex SH fingerprint, which may provide suitable markers of both VOO geographical (Quintanilla-Casas et al., 2020) and varietal origin (as evidenced in this study), which would be revealed by PLS-DA depending on the variable selected for supervising the pattern recognition analysis. The VOOs produced in the Southern Hemisphere were not satisfactorily classified by the PLS-DA model built using mainly Mediterranean samples, evidencing the need to include more samples from this region to improve the predictive ability for 'Arbequina' oils from this geographical origin.

Furthermore, a preliminary multi-class PLS-DA model to discriminate between the other cultivars represented in the sample set resulted in a 94.9% of overall correct classification by leave 10%-out cross-validation, suggesting that successful classification models analogous to that developed and validated for 'Arbequina' samples could be potentially developed to authenticate other VOO cultivars.

Finally, the exploration of PLS-DA regression coefficients revealed that a high number of variables contributed to the discrimination model, several of which corresponded to minor SHs or not well-resolved chromatographic peaks that could only be studied through a fingerprinting approach. This confirmed the advantage of high-dimensional non-targeted data approaches like fingerprinting over other conventional approaches.

CRediT authorship contribution statement

Berta Torres-Cobos: Formal analysis, Data curation, Investigation, Methodology, Validation, Writing – original draft. **Beatriz Quintanilla-Casas:** Formal analysis, Data curation, Investigation, Methodology, Validation, Visualization, Writing – review & editing. **Agustí Romero:** Resources, Writing – review & editing. **Antonia Ninot:** Resources. **Rosa M. Alonso-Salces:** Resources, Writing – review & editing. **Tullia Gallina Toschi:** Conceptualization, Funding acquisition, Project administration, Resources, Writing – review & editing. **Alessandra Bendini:** Conceptualization, Project administration, Resources, Writing – review & editing. **Francesc Guardiola:** Investigation, Supervision, Writing – review & editing. **Alba Tres:** Investigation, Funding acquisition, Methodology, Supervision, Validation, Writing – review & editing. **Stefania Vichi:** Conceptualization, Funding acquisition, Methodology, Resources, Validation, Writing – original draft, Supervision.

Declaration of competing interest

None.

Acknowledgments

Authors are grateful to Catalan Cooperatives for providing traceable VOO samples in the frame of the Autenfood project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodcont.2021.108200.

Funding

This work was developed in the context of the project AUTENFOOD, supported by ACCIÓ-Generalitat de Catalunya (Spain) and the European Union through the Programa Operatiu FEDER Catalunya 2014–2020 [grant number COMRDI-15-1-0035] and the project OLEUM, supported by the European Commission within the Horizon 2020 Program [grant agreement no. 635690]. This work was also supported by Generalitat de Catalunya (Spain) [grant number 2020FI_B00595]; and the Spanish Ministry of Science, Innovation and Universities and the FEDER Program [grant numbers FPU16/01744, RYC-2017-23601]. These funding sources had no involvement in the study design, collection, analysis or interpretation of data; or in the writing of the manuscript, or in the decision to submit the manuscript for publication. The information and views set out in this article are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

References

Agrimonti, C., Vietina, M., Pafundo, S., & Marmiroli, N. (2011). The use of food genomics to ensure the traceability of olive oil. *Trends in Food Science & Technology, 22*, 237–244. https://doi.org/10.1016/j.tifs.2011.02.002

Aparicio, R., Morales, M. T., Aparicio-Ruíz, R., Tena, N. D., & García-González, D. L. (2013). Authenticity of olive oil: Mapping and comparing official methods and promising alternatives. *Food Research International, 54*, 2025–2038. https://doi.org/10.1016/j.foodres.2013.07.039

Bajoub, A., Bendini, A., Fernández-Gutiérrez, A., & Carrasco-Pancorbo, A. (2018). Olive oil authentication: A comparative analysis of regulatory frameworks with especial emphasis on quality and authenticity indices, and recent analytical techniques developed for their assessment. A review. *Critical Reviews in Food Science and Nutrition, 58*, 832–857. https://doi.org/10.1080/10408398.2016.1225666

Ballin, N. Z., & Laursen, K. H. (2019). To target or not to target? Definitions and nomenclature for targeted versus non-targeted analytical food authentication. *Trends in Food Science & Technology, 86*, 537–543. https://doi.org/10.1016/j.tifs.2018.09.025

Berrueta, L. A., Alonso-Salces, R. M., & Héberger, K. (2007). Supervised pattern recognition in food analysis. *Journal of Chromatography A, 1158*, 196–214. https://doi.org/10.1016/j.chroma.2007.05.024

Bortolomeazzi, R., Berno, P., Pizzale, L., & Conte, L. (2001). Sesquiterpene, alkene, and alkane hydrocarbons in virgin olive oils of different varieties and geographical origins. *Journal of Agricultural and Food Chemistry, 49*, 3278–3283. https://doi.org/10.1021/jf001271w

Bosque-Sendra, J. M., Cuadros-Rodríguez, L., Ruiz-Samblás, C., & de la Mata, P. (2012). Combining chromatography and chemometrics for the characterization and authentication of fats and oils from triacylglycerol compositional data–a review. *Analytica Chimica Acta, 724*, 1–11. https://doi.org/10.1016/j.aca.2012.02.041

Cabrera, E. R., Arriaza, M., & Rodríguez-Entrena, M. (2015). Is the extra virgin olive oil market facing a process of differentiation? A hedonic approach to disentangle the effect of quality attributes. *Grasas Y Aceites, 66*, 4. https://doi.org/10.3989/gya.0253151

Cicerale, S., Liem, G., & Keast, R. S. J. (2016). Consumer perception, attitudes, liking and preferences for olive oil. In D. Boskou, & M. Clodoveo (Eds.), *Products from olive tree*. London, UK: IntechOpen. https://doi.org/10.5772/64554.

Council Regulation (EC). No 510/2006 of 20 March 2006 on the protection of geographical indications and designations of origin for agricultural products and foodstuffs. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32006R0510.

Damascelli, A., & Palmisano, F. (2013). Sesquiterpene fingerprinting by headspace SPME–GC–MS: Preliminary study for a simple and powerful analytical tool for traceability of olive oils. *Food Analytical Methods, 6*, 900–905. https://doi.org/10.1007/s12161-012-9500-9

Degenhardt, J., Köllner, T. G., & Gershenzon, J. (2009). Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry, 70*, 1621–1637. https://doi.org/10.1016/j.phytochem.2009.07.030

Dias, L. G., Fernandes, A., Veloso, A. C. A., Machado, A. A. S. C., Pereira, J. A., & Peres, A. M. (2014). Single-cultivar extra virgin olive oil classification using a potentiometric electronic tongue. *Food Chemistry, 160*, 321–329. https://doi.org/10.1016/j.foodchem.2014.03.072

García-González, D. L., & Aparicio, R. (2010). Research in olive oil: Challenges for the near future. *Journal of Agricultural and Food Chemistry, 24*, 12569–12577. https://doi.org/10.1021/jf102735n

Guinda, Á., Lanzón, A., & Albi, T. (1996). Differences in hydrocarbons of virgin olive oils obtained from several olive varieties. *Journal of Agricultural and Food Chemistry, 44*, 1723–1726. https://doi.org/10.1021/jf950571o

Marini, F., Balestrieri, F., Bucci, R., Magri, A. L., & Marini, D. (2004). Supervised pattern recognition to authenticate Italian olive oil varieties. *Chemometrics and Intelligent Laboratory Systems, 73*, 85–93. https://doi.org/10.1016/j.chemolab.2003.12.007

Montealegre, C., Alegre, M. L. M., & García-Ruíz, C. (2010). Traceability markers to the botanical origin in olive oils. *Journal of Agricultural and Food Chemistry, 58*, 28–38. https://doi.org/10.1021/jf902619z

Nielsen, N. P. V., Carstensen, J. M., & Smedsgaard, J. (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A, 805*, 17–35. https://doi.org/10.1016/S0021-9673(98)00021-1

Osorio-Bueno, E., Sánchez-Casas, J., Montaño-García, A., & Gallardo-González, L. (2005). Discriminating power of the hydrocarbon content from virgin olive oil of Extremadura cultivars. *Journal of the American Oil Chemists' Society, 82*, 1–6. https://doi.org/10.1007/s11746-005-1034-0

Papadia, P., Del Coco, L., Muzzalupo, I., Rizzi, M., Perri, E., Cesari, G., Simeone, V., Mondelli, D., Schena, F. P., & Fanizzi, F. P. (2011). Multivariate analysis of 1H-NMR spectra of genetically characterized extra virgin olive oils and growth soil correlations. *Journal of the American Oil Chemists' Society, 88*, 1463–1475. https://doi.org/10.1007/s11746-011-1812-1

Piccinonna, S., Ragone, R., Stocchero, M., Del Coco, L., De Pascali, S. A., Schena, F. P., & Fanizzi, F. P. (2016). Robustness of NMR-based metabolomics to generate comparable data sets for olive oil cultivar classification. An inter-laboratory study on Apulian olive oils. *Food Chemistry, 199*, 675–683. https://doi.org/10.1016/j.foodchem.2015.12.064

Quintanilla-Casas, B., Bertin, S., Leik, K., Bustamante, J., Guardiola, F., Valli, E., Bendini, A., Gallina-Toschi, T., Tres, A., & Vichi, S. (2020). Profiling versus fingerprinting analysis of sesquiterpene hydrocarbons for the geographical authentication of extra virgin olive oils. *Food Chemistry, 307*, 125556. https://doi.org/10.1016/j.foodchem.2019.125556

Romero, A. (2017). Olive oil composition affected by environmental factors. In A. Kiritsakis, & F. Shahidi (Eds.), *Olives and olive oil as functional food. Bioactivity, chemistry and processing* (pp. 69–76). New York, NY: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119135340.ch4.

Rondanini, D. P., Castro, D. N., Searles, P. S., & Rousseaux, M. C. (2011). Fatty acid profiles of varietal virgin olive oils (Olea europaea L.) from mature orchards in warm arid valleys of Northwestern Argentina (La Rioja). *Grasas Y Aceites, 62*, 4. https://doi.org/10.3989/gya.125110

Sacco, A., Brescia, M. A., Liuzzi, V., Reniero, F., Guillou, C., Ghelli, S., & Van der Meer, P. (2000). Characterization of Italian olive oils based on analytical and nuclear magnetic resonance determinations. *Journal of the American Oil Chemists' Society, 77*, 619–625. https://doi.org/10.1007/s11746-000-0100-y

Sayago, A., González-Domínguez, R., Urbano, J., & Fernández-Recamales, A. (2019). Combination of vintage and new-fashioned analytical approaches for varietal and geographical traceability of olive oils. *Lebensmittel-Wissenschaft und -Technologie-Food Science and Technology, 111*, 99–104. https://doi.org/10.1016/j.lwt.2019.05.009

Torres, M. M., Pierantozzi, P., Cáceres, M. E., Labombarda, P., Fontanazza, G., & Maestri, D. M. (2009). Genetic and chemical assessment of Arbequina olive cultivar grown in Córdoba province, Argentina. *Journal of the Science of Food and Agriculture, 89*, 523–530. https://doi.org/10.1002/jsfa.3483

Tous, J. (2017). The influence of growing region and cultivar on olives and olive oil characteristics and on their functional constituents. In A. Kiritsakis, & F. Shahidi (Eds.), *Olives and olive oil as functional food. Bioactivity, chemistry and processing* (pp. 45–80). New York, NY: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119135340.ch4.

Vichi, S., Guadayol, J. M., Caixach, J., López-Tamames, E., & Buxaderas, S. (2006). Monoterpene and sesquiterpene hydrocarbons of virgin olive oil by headspace solid-phase microextraction coupled to gas chromatography/mass spectrometry. *Journal of Chromatography A, 1125*, 117–123. https://doi.org/10.1016/j.chroma.2006.05.029

Vichi, S., Lazzez, A., Grati-Kamoun, N., Lopez Tamames, E., & Buxaderas, S. (2010). Evolution of sesquiterpene hydrocarbons in virgin olive oil during fruit ripening. *Journal of Agricultural and Food Chemistry, 58*, 6972–6976. https://doi.org/10.1021/jf100497c

Vichi, S., Tres, A., Quintanilla-Casas, B., Bustamante, J., & Guardiola, F. (2018). Sesquiterpene hydrocarbons, a promising tool for virgin olive oil geographical authentication. In M. Kontominas (Ed.), *Authentication and detection of Adulteration of olive oil*. New York, NY: Nova Science Publishers, Inc., ISBN 978-1-53614-596-0

CHAPTER 5.

HAZELNUT GEOGRAPHICAL AND VARIETAL AUTHENTICATION

The present **chapter 5** addresses the development of several analytical tools for the **authentication of hazelnut geographical and varietal origin.** Methods based on isotopic (**objective 3**) and metabolic markers (**objective 4.1**) have been explored, developed, validated, and compared (**objective 4.2**) to establish a reliable and efficient method for hazelnut authenticity. In this context, multi-isotopic approaches have been explored to verify the label-declared geographical origin (**objective 3**). Additionally, and based on the available literature (section 1.3.4), SH, UF, TAG, and spectroscopic fingerprints have been studied as metabolic approaches to simultaneously authenticate both the provenance and botanical origin of hazelnuts (**objective 4.1**).

First, PoC studies using reduced sample sets were conducted to assess the feasibility of each approach. For certain markers, it was also necessary to develop the analytical methods for their determination in hazelnuts. Subsequently, when the PoC studies yielded promising results, the approaches were evaluated including greater natural variability (**Figure 6**). This evaluation included an expanded sample set comprising more samples of different categories from various producing areas, cultivars, and harvest years.

5.1 Development of isotopic methods for hazelnut geographical authentication: Proof-of-concept study

5.1.1 Multi-isotopic approach

To evaluate the efficacy of various isotopic markers in verifying hazelnut provenance and to develop a robust multi-isotopic method, a PoC study was conducted using a selected, reduced sample set (n = 40). This sample set comprised samples from the same cultivar (TG) collected between 2019 and 2021 across four regions located in different countries: CHL (n = 10), ESP (n = 10), GEO (n = 10) and ITA (n = 10).

The elemental isotopic profile of light bioelements ($\delta^{13}C$, $\delta^{2}H$, $\delta^{18}O$, $\delta^{15}N$, $\delta^{34}S$), heavy geoelements ($^{86}Sr/^{87}Sr$), and the $\delta^{13}C$, $\delta^{2}H$ of the FAMEs were analysed (**Figure 9**). This multi-isotopic approach was combined with chemometric data analysis tools to build PLS-DA classification models to discriminate hazelnuts according to their origin. Additionally, the effect of fertilisation on isotopic ratios was assessed to avoid selecting markers affected by agronomic practices. The regression coefficients of the PLS-DA

models were studied to identify the variables driving the discrimination of hazelnuts by geographical origin, and consequently, the most promising isotopic markers for hazelnut provenance authentication. The results of this study are detailed in Publication 4.

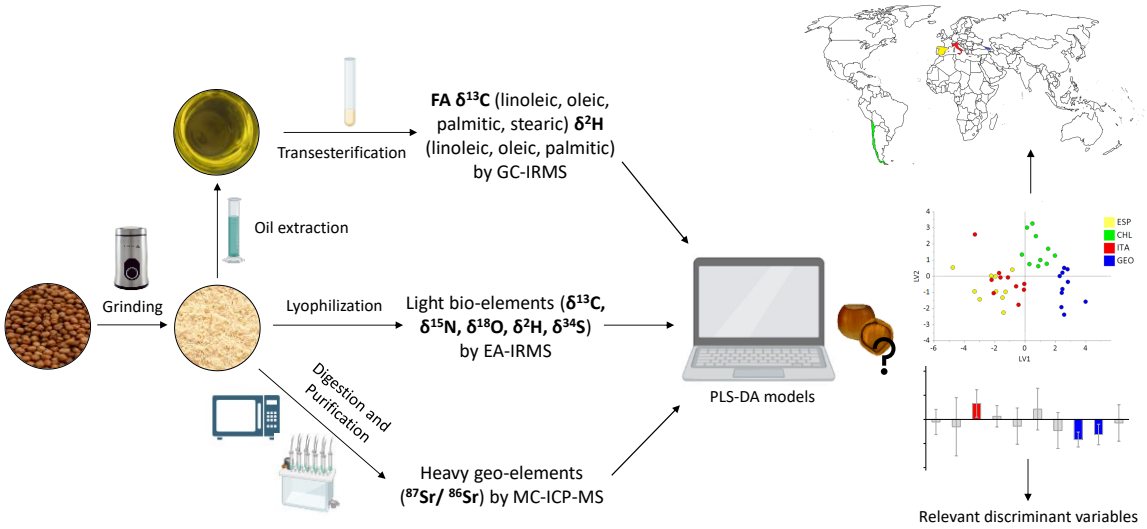


Figure 9. Graphical abstract of Publication 4. FA: fatty acids, GC-IRMS: gas chromatography-isotope ratio mass spectrometry, EA-IRMS: elemental analysis-isotope ratio mass spectrometry, MC-ICP-MS: multi collector-inductively coupled plasma-mass spectrometry, PLS-DA: partial least square-discriminant analysis, CHL: Chile, ESP: Spain, GEO: Georgia, ITA: Italy.

5.1.2 Publication 4

Investigating isotopic markers for hazelnut geographical authentication: Promising variables and potential applications

Berta Torres-Cobos, Mònica Rossell, Albert Soler, Mercè Rovira, Agustí Romero, Francesc Guardiola, Stefania Vichi, Alba Tres

Food Chemistry, ***2024****, 449, 139083. https://doi.org/10.1016/j.foodchem.2024.139083*

Supplementary material available in **Annex 3**

Food Chemistry 449 (2024) 139083

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem

Investigating isotopic markers for hazelnut geographical authentication: Promising variables and potential applications

B. Torres-Cobos [a,b], M. Rosell [c], A. Soler [c], M. Rovira [d], A. Romero [d], F. Guardiola [a,b], S. Vichi [a,b,*], A. Tres [a,b]

[a] Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Universitat de Barcelona. Av Prat de La Riba, 171, 08921 Santa Coloma de Gramenet, Spain
[b] Institut de Recerca en Nutrició i Seguretat Alimentària (INSA-UB), Universitat de Barcelona. Av Prat de La Riba, 171, 08921 Santa Coloma de Gramenet, Spain
[c] Grup MAiMA, Mineralogia Aplicada, Geoquímica i Hidrogeologia, Departament de Mineralogia, Petrologia i Geologia Aplicada, Institut de Recerca de l'Aigua (IdRA), Universitat de Barcelona, Martí i Franqués s/n, 08028 Barcelona, Spain
[d] Institute of Agrifood Research and Technology (IRTA), Ctra. de Reus – El Morell Km 3.8, Constantí 43120, Spain

ARTICLE INFO

Keywords:
Hazelnut
Geographical authentication
Stable isotopes
Strontium
Chemometrics

ABSTRACT

Hazelnuts' features and price are influenced by their geographical origin, making them susceptible to fraud, especially counterfeit claims regarding their provenance. Stable isotope analysis is a recognised approach to establish the geographical origin of foods, yet its potential in hazelnut authentication remains unexplored. In this prospective study, we assessed multiple isotopic markers in hazelnuts from different origins and evaluated the most promising variables for geographical authentication by chemometric tools. Our findings indicate that bulk $\delta^{18}O$, along with $\delta^{2}H$ and $\delta^{13}C$ in the main fatty acid methyl esters, exhibit significant potential in discriminating geographical origins, and $^{87}Sr/^{86}Sr$ analysis could serve as a proficient confirmatory tool. Though no single marker alone can differentiate between all the studied origins, employing a multi-isotopic approach based on PLS-DA models achieved up to 92.5 % accuracy in leave-10 %-out cross-validation. These findings will probably lay the groundwork for developing robust models for hazelnut geographical authentication based on larger datasets.

1. Introduction

Hazelnuts are a widely used and prized ingredient in both sweet and savoury foods. Their sensory and qualitative characteristics are heavily influenced by their growing region (Król & Gantner, 2020) and their market value fluctuates accordingly. As an example, in 2021, the price range for in-shell hazelnuts from Georgia and Italy was between 1449 USD/T and 4174 USD/T, respectively (FAOSTAT, 2021). Hazelnuts' high economic value makes them particularly susceptible to frauds, including the counterfeiting of geographical origin. Even though the official Regulation (Regulation (EU) No, 1169) defines geographical origin as a label claim that must be verified by official inspection services, when necessary (Mahalovich et al., 2016), the lack of effective methods for detecting this fraud provides counterfeiters with an opportunity to exploit the situation. Fraudulent activities can have a significant impact on the industrial sector that uses these commodities. This holds particular significance for foods that fall under EU recognition, including Protected Designation of Origin, as the origin of the food product serves as the basis of these quality schemes. Hence, the implementation of effective methods to detect and prevent fraudulent activities related to hazelnut origin claims is necessary to protect the industry and consumers.

Currently, stable isotope composition analysis is among the most acknowledged approaches to establish the geographical origin of foods, as the isotopic composition of both light bio-elements, namely carbon (C), hydrogen (H), oxygen (O), nitrogen (N), and sulphur (S) and heavy geo-elements, such as strontium (Sr) are strongly influenced by factors that are indicative of the geographical origin, such as geology and hydrogeology (Kelly et al., 2005; Laursen et al., 2016, Podio et al., 2013; De Rijke et al., 2016). Regarding light elements such as H and O, phase changes of water (solid–liquid-vapour) lead to isotopic fractionation because of the different saturation vapour pressures of different isotopologues. This isotopic fractionation resulting from evaporation, condensation, and precipitation of meteoric water is influenced by

* Corresponding author at: Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Universitat de Barcelona. Av Prat de La Riba, 171, 08921 Santa Coloma de Gramenet, Spain.
E-mail address: stefaniavichi@ub.edu (S. Vichi).

https://doi.org/10.1016/j.foodchem.2024.139083
Received 28 July 2023; Received in revised form 16 February 2024; Accepted 18 March 2024
Available online 23 March 2024

factors like temperature, latitude, altitude, and proximity to the coast. These factors give rise to notable variations in δ^2H and δ^{18}O within groundwater across different geographical regions. By analysing these isotopic signatures in plant samples, valuable insights can be obtained regarding their geographical provenance (Kelly et al., 2005; De Rijke et al., 2016). On the other hand, the δ^{13}C value in plants is primarily determined by their specific types of photosynthetic carboxylation reactions. However, it has been noted that environmental factors can also exert an influence on carbon isotope fractionation, making of δ^{13}C a potential indicator of the origin of plants (Yoneyama et al., 2000; Branch et al., 2003). The isotopic composition of S and N in plant tissues can be influenced by the application of isotopically different sulphur and nitrogen-containing fertilisers (Vitòria et al., 2004), making them more dependent on agricultural practices rather than geographical factors. Nevertheless, the full extent of this impact is not entirely understood and requires evaluation specific to the plant species of interest. For instance, regarding δ^{15}N, it has been observed that it can be influenced by the plant's nitrogen uptake process (Anderson and Smith, 2006). Additionally, nitrogen isotopes are subject to the influence of geoclimatic factors (soil type, temperatures, precipitation), which vary across geographical regions and can significantly contribute to shaping the isotopic signature of plant materials (Yoneyama et al., 2000; Brescia et al., 2002; Anderson and Smith, 2006). Furthermore, the isotopic analysis of heavy geo-elements, like strontium, offers an enhanced connection between soil characteristics and primary agricultural products due to the negligible isotopic fractionation because of the small mass difference between the two main Sr isotopes (^{87}Sr and ^{86}Sr). Only ^{87}Sr is radiogenic; it is produced by the long-term decay from the radioactive ^{87}Rb present in the distinct rock types and geological formations (Kelly et al., 2005, Laursen et al., 2016). Also, in this case, traces of geological Sr found in fertilisers made by phosphate from mining could potentially impact the Sr isotopic composition in crops (Vitòria et al., 2004; Techer et al., 2017).

Given these considerations, the isotopic analysis has been extensively utilized for the geographical verification of a diverse range of food products (Bertoldi et al., 2019; Camin et al., 2017; Kelly et al., 2005, Perini et al., 2018; Podio et al., 2013; De Rijke et al., 2016). The consistently satisfactory results obtained from this analysis have prompted the proposal of employing this method even in legal cases (Camin et al., 2017). As demonstrated by previous studies, measuring multiple parameters within a food product enhances the accuracy in determining its geographical origin (Bertoldi et al., 2019; Luykx & Van Ruth, 2008; Podio et al., 2013), especially when multi-isotope-ratio analysis is paired with molecular-specific isotopic data (Bontempo et al., 2019). However, only a limited number of studies focus on single or multi-isotopic composition specifically for nut products. Encouraging results have been obtained from the analysis of bio-element isotopic composition in walnuts (Di Pierro et al., 2018; Krauß et al., 2020), pistachios (Anderson and Smith, 2006), and North American pine nuts (Mahalovich et al., 2016) and from the evaluation of ^{87}Sr/^{86}Sr isotopes in pistachios (Zannella et al., 2017) and peanuts (Zhu et al., 2014). No data are available on the application of isotopic markers for hazelnut geographical authentication. Finally, a key element in authentication by multi-isotope-ratio analysis is the data treatment and chemometric analysis (Drivelos & Georgiou, 2012; Podio et al., 2013; De Rijke et al., 2016). When a large number of variables are studied, multivariate techniques such as Principal Component Analysis (PCA) and Partial Least Squares Discriminant Analysis (PLS-DA), among others, are crucial to extract all the relevant information.

The objective of this prospective study was to evaluate the efficacy of specific isotopic markers to identify hazelnut geographical origin. To achieve this purpose, we analysed the elemental isotopic profile (δ^{13}C, δ^2H, δ^{18}O, δ^{15}N, δ^{34}S, and ^{86}Sr/^{87}Sr) and the δ^{13}C and δ^2H of the fatty acid methyl esters (FAMEs) of the hazelnut oil of 40 samples of raw hazelnuts from four different origins. Chemometric tools were used within a multi-isotopic approach to compare and identify the most promising variables for hazelnut geographical authentication.

2. Material and methods

2.1. Samples

Forty samples of hazelnuts from the 'Tonda di Giffoni' Italian cultivar were produced over one to three harvest seasons in four different countries: Spain (n = 10; 2019, 2020, 2021), Italy (n = 10; 2019, 2020, 2021), Georgia (n = 10; 2021) and Chile (n = 10; 2019, 2020). All samples were collected when hazelnuts were ripe (harvested in August-October for Georgia Italy and Spain and in March-April for Chile). Each sample was obtained from one single tree. Geographical coordinates and fertilisation data are reported in Table 1. Samples were provided in the framework of the TRACENUTS project (PID2020-117701RB). Collected hazelnuts were shelled at the laboratory. Kernels were stored under vacuum and refrigerated at 4 °C until analysis.

2.2. Bulk isotopic analysis by Elemental Analysis-Isotope ratio Mass spectrometry (EA-IRMS)

2.2.1. Sample preparation

Preliminary tests were conducted to determine the most suitable sample treatment. For this purpose, three aliquots of 30 g each were taken from the same homogeneous hazelnut sample (500 g) and subjected to different grinding methods: cryogenic milling (Cryogenic Mill 6850, SPEX Certiprep, Metuchen, New Jersey, USA), domestic grinder (Aromatic, Taurus, Oliana, Spain) and a combination of grinding followed by cryogenic milling. Then, 5 g were taken in triplicate from each ground sample and lyophilized in a Telstar Cryodos-45 freeze dryer (Telstar, Terrassa, Spain). The lyophilization time until constant weight was assessed, and the amount of water removed from the sample was calculated.

The repeatability of the method chosen for sample treatment was assessed on C/N ratios, δ^{15}N‰, and δ^{13}C‰, determined on 10 aliquots obtained from the same ground and lyophilized sample, and extrapolated to the rest of the light elements. Mean values and standard deviation of each parameter were calculated.

The sample treatment selected for the isotope bulk analysis of δ^{13}C, δ^2H, δ^{18}O, δ^{15}N, δ^{34}S involved grinding 30 g of the raw hazelnuts into fine powder using a domestic grinder (Aromatic, Taurus, Oliana, Spain), followed by a three-day lyophilizing process.

A dual-water equilibration test was performed to estimate any possible exchange of H between the samples and the ambient according to previous protocols (Sauer et al., 2009; Qi and Coplen, 2011). For this, a set of 10 aliquots extracted from the same ground and lyophilized sample, plus the standards, were weighed and loaded into individual silver capsules (Lüdi Swiss, Flawil, Switzerland). Then, five of them were equilibrated in a glass desiccator with water depleted in ^{2}H (Milli-Q water, δ^2H = −43 ‰) and the other five were equilibrated with deuterated water (δ^2H = +100 ‰). In each glass desiccator, a set of standards was also included. Samples were equilibrated for seven days at ambient temperature (25 °C). Prior to analysis, samples equilibrated with light and heavy water were dried in separate desiccators filled with Sicapent (P_2O_5) for seven days.

2.2.2. EA-IRMS

2.2.2.1. C (δ^{13}C-values in ‰) and N (δ^{15}N-values in ‰) isotope analysis. About 0.8 mg of the powdered sample was weighed into individual tin capsules for the determination of bulk δ^{13}C- and δ^{15}N-values (‰). Samples were measured in a Flash IRMS™ Elemental Analyser coupled to a Delta V Advantage (IRMS) via Conflo IV interface (Thermo Fisher Scientific, Waltham, Massachusetts, USA). In the elemental analyser, each sample was combusted with oxygen added to the helium stream at a temperature of 900 °C in a reactor comprised of copper oxide and silvered oxides of cobalt (ThermoFisher Scientific, Waltham,

Table 1
Geographical coordinates and the applied fertilization data of the hazelnut production parcels during the studied years, as provided by the producers.

	Geographical coordinates	Fertiliser (as N source) (Units N/ha)				Organic N
		Straight NH_4NO_3	NPK (N: NO_3; NH_4)	NPK (N: Urea)	NPK (N: not specified)	
Chile	35°15′36″ S, 71°32′60″ W	–	–	–	91–115	–
Italy	42°25′23″ N, 12°4′45″ E	–	–	–	–	–
Georgia	42°27′34″ N, 41°51′31″ E	40	–	–	30	–
Spain	41°10′15″ N, 1°10′09″ E	–	45	27	–	mulching[a]

[a] : spontaneous vegetation, mowed 2–3 times per year and left on the ground between trees.

Massachusetts, USA) and then NO_x gases were reduced at 650 °C with electrolytic copper to produce N_2. The gases obtained were carried through water traps (granular magnesium perchlorate, Elemental Microanalysis Ltd, Okehampton, UK) and N_2 and CO_2 were separated in a stream of helium N 60 (grade 99.9999 % purity, Air Liquide, Madrid, Spain) at 180 mL/min with a GC column (length 2 m, diameter 6 x 5 mm) set to 40 °C and transferred into the isotope ratio mass spectrometer. The certified international standards used were IAEA CH7 ($\delta^{13}C$ = –32.15 ‰) and IAEA N1 ($\delta^{15}N$ = 0.4 ‰) both from Vienna (Austria). UCGEMA K (keratin, $\delta^{13}C = -14.97$ ‰, $\delta^{15}N = 13.52$ ‰), UCGEMA CH (chitin, $\delta^{13}C = -22.08$ ‰, $\delta^{15}N = -4.81$ ‰), fructose ($\delta^{13}C = -10.8$ ‰) and UCGEMA P (animal hair, $\delta^{15}N = +7.6$ ‰) were used as internal secondary standards. The reference gases were N_2 ($\delta^{15}N = -1.0$ ‰) and CO_2 ($\delta^{13}C = -41.2$ ‰). For sample combustion, O_2 was injected for 2 s at 250 mL/min and for the $\delta^{13}C$ analysis the sample was diluted with the carried gas to 78 %. Samples were analysed in duplicate.

2.2.2.2. H (δ^2H-values in ‰) isotope analysis. About 0.1 mg of the powdered sample was transferred into silver capsules for solids. To prevent potential further H exchange, the capsules were sealed immediately after being removed from the desiccator and were analysed right after the encapsulation and pressing of the sample, where the sample remains relatively isolated from the atmosphere. Samples were analysed in a TC/EA-IRMS Delta Plus XP (Thermo Fisher Scientific) equipped with a conventional autosampler (Sample Tray, N° 2, MAS200R autosampler) and a pyrolysis reactor, heated to 1450 °C. Specifically, the filling of this 450 mm ceramic reactor from the bottom to the top is as follows: 20 mm quartz wool (Lüdi Swiss), 30 mm of graphite (IVA Analysentechnik, Meerbusch, Germany). Then, glassy carbon reactor (IVA) is introduced inside the ceramic reactor and filled with 120 mm of graphite (IVA) and then the graphite crucible (IVA); all this procedure is according to the Thermo Fisher Scientific manual. Helium was used as carrier gas (pressure 90 kPa) and the reference gas was hydrogen (H_2) with a δ^2H-value of −115.6 ‰. The certified standards used were IAEA-CH-7 (polyethylene, $\delta^2H = -100.3$ ‰), coumarin ($\delta^2H = 82.3$ ‰, reference material from Schimmelmann Research: Indiana University Bloomington), icosanoic acid ($\delta^2H = -166.7$ ‰) and biphenyl ($\delta^2H = -41.2$ ‰). Samples were analysed in duplicate in less than 12 h after their sealing.

2.2.2.3. O (δ^{18}O-values in ‰) isotope analysis. About 0.3 mg of the powdered sample was introduced into silver capsules (Lüdi Swiss, Flawil, Switzerland). Samples were analysed in the same TC/EA-IRMS Delta Plus XP mentioned above but the pyrolysis reactor temperature was 1445 °C. The filling of the reactor was the same as reported for ^{2}H analyses. Helium was used as carrier gas (pressure 62 kPa) and the reference gas was carbon monoxide (CO) with a δ^{18}O-value of −8.68 ‰ (pressure 180 kPa). The certified standard IAEA-601 (benzoic acid, $\delta^{18}O = +23.3$ ‰) and the internal secondary standards, UB-YCEM ($\delta^{18}O = +17.6$ ‰) and UB-ASC ($\delta^{18}O = +13.2$ ‰), both barium sulphates, were used fitting the range of the samples. Samples were analysed in duplicate.

2.2.2.4. S (δ^{34}S-values in ‰) isotope analysis. About 8 mg of the powdered sample was weighed into tin capsules (Elemental Microanalyses, Okehampton, UK). V_2O_5 was added as a catalyser. Samples were analysed in the same TC/EA-IRMS Delta Plus XP with a pyrolysis reactor at 1035 °C. Specifically, the filling of the 450 mm quartz reactor from the bottom to the top consists of: 30 mm of quartz wool (Lüdi Swiss, Flawil, Switzerland), 90 mm of copper wires (Elemental Microanalysis), 45 mm of quartz chips (Lüdi Swiss) and 45 mm of tungsten (VI) oxide (WO_3) (Elemental Microanalysis). Helium was used as carrier gas (pressure 65.5 kPa) and the reference gas was sulphur dioxide (SO_2) with a δ^{34}S-value of 1.266 ‰ (pressure 50 kPa). The certified standards, IAEA S-2 (silver sulphide, $\delta^{34}S = +22.7$ ‰), IAEA SO-5 (barium sulphate, $\delta^{34}S = +0.5$ ‰) and IAEA SO-6 (barium sulphate, $\delta^{34}S = -34.1$ ‰), and the secondary standard UB-YCEM (barium sulphate, $\delta^{34}S = +12.8$ ‰) were used. Samples were analysed in duplicate.

2.3. Isotopic analysis of FAMEs by gas Chromatography-Isotope ratio Mass spectrometry (GC-IRMS)

2.3.1. Sample preparation

2.3.1.1. Oil extraction. Lipid fraction was extracted with 50 mL of diethyl ether from 25 g of ground hazelnuts and the organic solvent was evaporated to dryness.

2.3.1.2. Preparation of FAMEs. An aliquot of 100 mg of hazelnut oil was dissolved in 2 mL of hexane and 200 μL of 2 M methanolic potassium hydroxide solution was added (Hrastar et al., 2009). The mixture was centrifuged, and the supernatant was analysed.

2.3.2. GC-IRMS

The analysis of H and C isotopes of individual FAMEs was carried out in duplicate using a Trace GC Ultra gas chromatograph with a Triplus Autosampler coupled to an Isotope Ratio Mass Spectrometer Delta V Advantage through a GC Isolink interface (Thermo Fisher Scientific). A total of 1 μL of sample was injected with a split ratio of 1:5. Analytes were separated on a VF-23ms capillary column (60 × 0.32 mm I.D., 0.15 μm of Agilent Technologies, Santa Clara, California, USA). The initial GC oven temperature was 60 °C and was held for 1 min, then, it was increased to 160 °C at 6 °C/min and held for 10 min. Finally, it was increased to 240 °C at 6 °C/min.

Helium was the carrier gas, at a flow rate of 1.8 mL/min. The temperature of the injector was 240 °C. The commercial NiO/CuO-NiO-Pt combustion reactor (P/N 1255321, Thermo Fisher Scientific) operated at 1000 °C for CO_2. In the case of H_2, the commercial high temperature reactor (P/N 1255330, Thermo Fisher Scientific) was set at 1400 °C. This pyrolysis reactor is empty, but an inner layer of carbon is formed that covers the walls of the ceramic when conditioning it, this layer serves to avoid contact with the oxygen of the walls (which is Al_2O_3)

and, on the other hand, to catalyse the reaction. The certificated standards, androstane, USGS76, coumarin, and a secondary standard, FAME C19, were used for the δ^2H analysis. Icosane, FAME C16, USGS76, FAME C19, USGS72, phenanthrene 16/0020 were used for the δ^{13}C analysis.

2.4. Isotopic analysis of $^{87}Sr/^{86}Sr$ by multi Collector-Inductively coupled Plasma-Mass spectrometry (MC-ICP-MS)

2.4.1. Sample preparation and digestion

A 0.5 g-powdered sample was treated with 3 mL of concentrated HNO_3 and 1 mL of water, both ultrapure, into a closed quartz reactor using a microwave assisted digestion system Ultrawave ECR (Milestone Slr, Sorisole, Italy) at 240 °C (ramp for 30 min and maintenance for 15 min). After complete digestion, 5 mL of ultrapure water were added, and the solution was weighed. The final volume was calculated by the solution weight and weight/volume ratio.

2.4.2. Preliminary quantification of Sr and Rb by Inductively coupled Plasma-Mass spectrometry (ICP-MS)

Samples were analysed by a Nexion 350d ICP-MS (PerkinElmer Life & Analytical Sciences, Waltham, Massachusetts, USA) in standard mode without collision and/nor reaction gas in 1/20 ratio.

Standard solutions of concentrations in the range of the samples, prepared from 1 g/L Rb and Sr certified standards (Inorganic Ventures, Christiansburg, Virginia, USA), were used for calibration.

2.4.3. Sample purification

After digestion, samples were evaporated to dryness inusing Savillex ® PFA beakers using ultra clean evaporation stations (ISO 5) at the LIRA (Laboratori d'Isótops Radiogenics i Ambientals) ultraclean Lab (UB). Then, samples were brought back into solution in ultrapure (double distilled) 8 N HNO_3 acid (1 mL). Sr was purified using Sr-SPEC Resin (100–150 μm mesh; Triskem International) packed in 1 mL polypropylene cartridges with porous PFA mesh used as frits and connected into a Triskem24-manifold vacuum box attached to a vacuum pump system. Sr elution started by cleaning the Sr-SPEC resin with 20 mL of 0.05 N HNO_3 and conditioning with 5 mL of 8 N HNO_3 before loading the sample in 1 mL of 8 N HNO_3 (a previous concentration check is made to assure loading around 500 ng of Sr). Next, matrix was eluted with 5 mL of 8 N HNO_3 to mainly get rid of sample matrix as well as Rb, which presents isobaric interferences during the spectrometric measurements. The purified Sr was finally eluted with 5 mL of ultrapure 0.05 N HNO_3, evaporated and dissolved in 2 % HNO_3 acid for the MC-ICP-MS measurements.

2.4.4. MC-ICP-MS

The determination of Sr isotope composition was performed using a Plasma 3 Multi Collector Inductively Coupled Plasma Mass Spectrometer (Nu Instruments-AMETEK) at Centres Científics i Tecnològics of the Universitat de Barcelona (CCiTUB). Samples and standards were matched in concentration. Procedural blanks contained less than 360 pg of Sr, which was negligible compared to the Sr amounts in the samples (over 250 ng) and were systematically corrected. The contribution of ^{87}Rb to the ^{87}Sr signal was mathematically corrected from the measurement of the ^{85}Rb signal, assuming a $^{87}Rb/^{85}Rb$ ratio of 0.38562. The ^{86}Kr interference on ^{86}Sr, caused by Kr impurities in the argon gas, was also corrected by measuring the ^{83}Kr signal, and assuming a $^{83}Kr/^{86}Kr$ value of 0.66453. Mass bias was corrected with the exponential model, using the traditionally accepted $^{86}Sr/^{88}Sr$ ratio value of 0.1194 (Nier, 1938). A further normalization by the sample-standard bracketing method was performed by analysing the NIST SRM 987 isotopic certified standard before and after each sample. Results were provided relative to this standard, assuming a reference value for the $^{87}Sr/^{86}Sr$ ratio of 0.710249 (Azmy et al., 1999). The reproducibility of the analysis was typically better than 0.000030 (2SD).

2.5. Data treatment and statistical analysis

First, the possible effect of the fertilisation on the δ^{15}N, δ^{34}S and $^{87}Sr/^{86}Sr$ was studied. δ^{15}N values were plotted versus δ^{34}S values and 1/[Sr] (μg/g) was plotted versus $^{87}Sr/^{86}Sr$ values as found in Epova et al. (2019). Then, these values were compared with those reported in the literature according to the type of fertilisers and to the geogenic values of each geographical area. After evaluating the results obtained in our study, the variables δ^{15}N, δ^{34}S were excluded from the development of geographical modelling, while $^{87}Sr/^{86}Sr$ was further considered.

A data matrix was built, consisting of the 40 samples (rows) and 11 variables presumed to be unaffected by agronomic practices (bulk δ^{13}C, δ^2H and δ^{18}O; δ^{13}C of C16:0, C18:0, C18:1, C18:2 FAMEs and δ^2H of C16:0, C18:1, C18:2 FAMEs; $^{87}Sr/^{86}Sr$) (columns).

Principal Component Analysis (PCA) was performed with SIMCA v13.0© (Umetrics AB, Sweden) to explore the data and to identify any potential outliers according to the Hotelling's T^2 range and model residuals parameters. No outliers were detected.

Afterwards, the data matrix was used to develop and validate Partial Least Square-Discriminant Analysis (PLS-DA) classification models to discriminate samples according to their geographical origin (SIMCA v13.0©). In PLS-DA multi-class models, a dummy Y matrix with as many classification vectors as classes was used, each vector had values of 1 for one class (a specific country of origin) and 0 for all the other classes (the other countries). Then, each sample was classified into the class corresponding to the vector leading to the highest PLS predicted value (PV); but samples whose PV did not reach the classification threshold (PV < 0.5) for any vector were not assigned to any country (no class).

The regression coefficients of the global PLS-DA model were studied to determine which variables were more relevant to discriminate among origins. To evaluate the significance of the regression coefficients, the jack-knife standard error of cross-validation (SEcv) was used. Coefficients were considered significant if their values were higher than their corresponding SEcv. The variables exhibiting significant coefficients were used to develop optimized PLS-DA classification models. Two models were built: one based on isotopic profiling of light elements alone, which are suitable for routine analysis, and another model that incorporated also $^{87}Sr/^{86}Sr$ isotopic composition, to further evaluate its discriminatory potential as a heavy geo-element.

Models' performance was assessed through internal validation with leave-10 %-out cross-validation, and the optimal number of latent variables (LV) was selected according to the lowest Root Mean Squared Error of Cross Validation (RMSEcv) criteria. The optimal pre-processing for all the models was mean centring and scaling to the unit of variance. Permutation test and ANOVA on the cross-validated predictive residuals (p-value) were carried out to assess the models' overfitting. The suitability of the PLS-DA models was evaluated by the Q^2 values and the percentage of correct classification of each class.

3. Results and discussion

3.1. Sample preparation for bulk isotopic analysis

Homogeneity can be a critical factor that limits the precision of bulk isotopic analysis when applied to solid samples with distinct morphological components (Carter & Fry, 2013), such as hazelnuts. Moreover, the high lipid content in hazelnut kernels can pose challenges in obtaining a fine powder necessary for accurately weighing small and uniform amounts of this sample matrix. For this reason, different grinding methods were tested, a small domestic grinder, a cryogenic mill and a combination of both. The lyophilization efficiency achieved after the application of each grinding method was evaluated (**Table S1, Supplementary material**). All the replicates obtained using a domestic grinder were completely lyophilized within 72 h, while only two of the cryogenic mill replicates achieved the complete lyophilization, after 127 h. The combination of grinder and cryogenic mill led to a paste that

could not be lyophilized within 127 h. Considering these results, the cryogenic mill was excluded.

Repeatability results for $\delta^{15}N$ and $\delta^{13}C$ bulk analysis (n = 10) were –28.17 ± 0.05 ‰ and 6.9 ± 0.2 ‰, respectively, confirming that the selected method produces a sufficiently homogeneous sample, ensuring acceptable precision in the isotopic measurements. It was assumed that the observed repeatability also held valid for the remaining elements analysed by EA-IRMS.

The extent of exchangeable H was evaluated to assess the impact of recent exposure of hazelnuts to water or vapor on the H isotope ratios. The dual-water equilibration assay revealed identical $\delta^{2}H$ ‰ values (−149 ± 2 and −152 ± 2) between the samples equilibrated with the two isotopically different waters (Milli-Q grade water −43 ‰ and deuterated water +100 ‰, respectively). Therefore, labile H appeared not to be isotopically relevant in the hazelnut samples.

3.2. N, S and Sr isotopic profile of hazelnuts and possible impact of fertilisation practices

Mean isotopic compositions of N, S and Sr in hazelnuts from the four origins are reported with the corresponding standard deviation in Table 2. The plot of $\delta^{15}N$ ‰ versus $\delta^{34}S$ ‰ values (Fig. 1) reveals a lack of discernible patterns in the $\delta^{34}S$ values of hazelnuts across the four distinct geographical areas. According to Vitòria et al., 2004, these $\delta^{34}S$ values are in the range of fertilisers, so geographical factors that may differentiate these zones could have been masked by the diverse agronomic practices employed in each plantation, thereby losing value as a geographical marker in this study. On the contrary, $\delta^{15}N$ differentiates almost completely the samples from each of the four provenances, showing the lowest values in Chilean hazelnuts, followed by Italian, Georgian and Spanish ones. The N isotopic composition of both Chilean and Italian samples aligns with the expected values based on current fertilisation agricultural practices (Table 1) and previous literature (Vitòria et al., 2004; Bateman and Kelly, 2007; Laursen et al., 2013). Specifically, the application of synthetic fertilisers justifies the depletion of ^{15}N in the Chilean samples compared to the non-fertilized Italian samples. Conversely, also synthetically fertilised Georgian and Spanish samples exhibited enriched ^{15}N values compared to the latter. These results may be attributed to the specific type of fertilisation employed or the long-term fertilisation history of such soils, given that residual nitrogen from fertilizers, as forecasted by Sebilo et al. (2013), possesses

Table 2
Mean and standard deviation results of all the isotopic markers for the 4 origins (Chile, Georgia, Italy and Spain).

	Chile	Georgia	Italy	Spain
$\delta^{15}N$ (‰)	−0.6 ± 0.5	2.5 ± 0.4	0.9 ± 0.6	4.0 ± 0.4
$\delta^{34}S$ (‰)	7.2 ± 0.9	5.6 ± 1.2	6.4 ± 0.5	6.1 ± 1.3
$\delta^{13}C$ (‰)	−26.7 ± 0.8	−27.7 ± 0.4	−26.9 ± 0.6	−26.7 ± 0.5
$\delta^{2}H$ (‰)	−169.2 ± 8.3	−172.4 ± 3.1	−161.8 ± 7.8	−159.0 ± 7.9
$\delta^{18}O$ (‰)	23.7 ± 1.0	20.1 ± 0.5	24.4 ± 0.7	23.6 ± 0.4
$\delta^{13}C$-Palmitic (‰)	−30.1 ± 0.6	−30.9 ± 0.4	−30.1 ± 0.7	−30.1 ± 0.5
$\delta^{13}C$-Stearic (‰)	–32.2 ± 1.1	–33.5 ± 1.2	–32.2 ± 1.2	–31.8 ± 1.0
$\delta^{13}C$-Oleic (‰)	−28.3 ± 0.7	−29.0 ± 0.4	−28.3 ± 0.8	−28.3 ± 0.6
$\delta^{13}C$-Linoleic (‰)	−30.6 ± 0.7	−31.4 ± 0.4	−31.2 ± 0.9	−30.9 ± 0.9
$\delta^{2}H$-Palmitic (‰)	−183.9 ± 7.4	−187.5 ± 5.6	−175.4 ± 5.4	−164.9 ± 8.4
$\delta^{2}H$-Oleic (‰)	−195.2 ± 4.8	−200.8 ± 2.7	−187.4 ± 3.1	−179.8 ± 6.1
$\delta^{2}H$-Linoleic (‰)	−218.4 ± 3.6	–223.3 ± 4.9	−207.3 ± 4.9	−202.2 ± 6.6
$^{87}Sr/^{86}Sr$	0.70436 ± $4x10^{-5}$	0.706 ± $1x10^{-3}$	0.709 ± $1x10^{-3}$	0.7087 ± $1x10^{-4}$

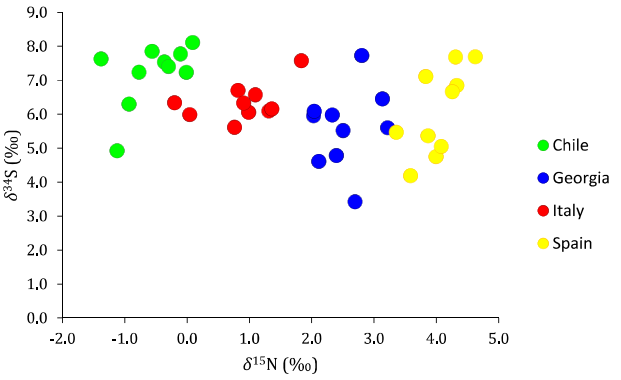


Fig. 1. Plot of $\delta^{15}N$ ‰ values versus $\delta^{34}S$ ‰ values in hazelnuts from Chile (green), Georgia (blue), Italy (red), and Spain (yellow).

the capacity to sustain crop nutrition for a duration extending up to five decades. In Georgia, significant amounts of straight fertiliser NH_4NO_3 were used in addition to NPK, which has been reported to yield relatively higher enriched ^{15}N water-soluble products ($\delta^{15}N_{NO3}$ = +5.6 ‰) in some cases (Vitòria et al., 2004). On the other hand, the higher $\delta^{15}N$ values observed in Spanish samples could not be easily related to the current fertilisation practises, as producers said to apply a mixture of NPK and leave spontaneous vegetation mowed 2–3 times per year as mulch between the rows of trees. This vegetation cover (legumes or non-legumes) would enhance the homogeneity of the soil nitrogen supply and reduce the fertilizer influence on the isotopic composition of hazelnuts (Krauß et al., 2020; Giannioti et al., 2024), so we propose that it should be the result of previous history of soil fertilization.

The findings from this study, along with previous research, indicate that the complexity arising from the variability of synthetic fertilisers' isotopic signatures, plant physiological factors, potential bacterial-mediated reactions resulting in N isotopic fractionation, the potential influence of climate and soil characteristics and remaining fertilizer N still residing in the soil present challenges in elucidating the impact of agricultural practices and, specifically, in predicting N isotopic composition based on geographical origin.

Regarding Sr, combining its concentration and isotopic composition in hazelnuts permitted a clear discrimination of the samples according to their origin (Fig. 2). Despite the reported isotopic differences in trace amounts of Sr present in fertilisers (Vitòria et al., 2004), the $^{87}Sr/^{86}Sr$ composition in hazelnuts align with the geogenic factors of each production area after checking geological maps (Jones et al., 1994). Spanish and Italian samples presented the highest values, consistent with Holocene formations, whereas Georgian and Chilean samples showed lower ones, indicative of more recent Neogene formations and recent

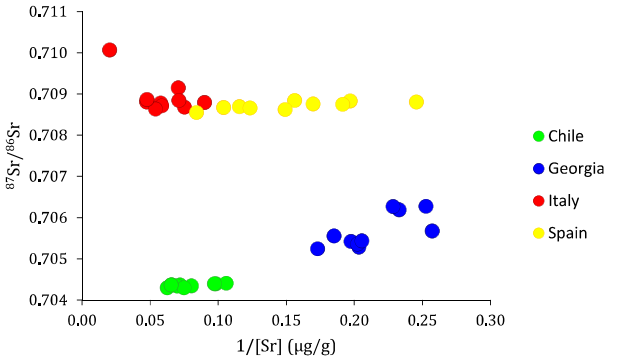


Fig. 2. Plot of 1/[Sr] (µg/g) versus $^{87}Sr/^{86}Sr$ values in hazelnuts from Chile (green), Georgia (blue), Italy (red) and Spain (yellow).

volcanism, respectively. Furthermore, the Sr abundance in hazelnuts varied according to their origin, with Italian and Chilean hazelnuts exhibiting higher levels, followed by Spanish and Georgian samples. Although Sr isotopic composition can potentially be influenced by the use of different fertilisers, the impact of these variations does not seem to be critical in the analysed samples. Therefore, it is deemed appropriate to further consider this marker for the development of geographical classification models, while also acknowledging that the practical application of this determination in routine analysis may present challenges.

3.3. Evaluating relevant isotopic markers for hazelnut geographical authentication by Partial Least Square-Discriminant analysis (PLS-DA)

3.3.1. Evaluation of light element isotopic composition: Regression coefficients of global PLS-DA model

In Table 2, mean and standard deviation results of all the isotopic markers for the four origins are reported. After excluding $\delta^{15}N$ and $\delta^{34}S$ for the reasons given above, noticeable differences between the means can be appreciated, especially for the $\delta^{18}O$, $\delta^{2}H$ of the FAMEs and $^{87}Sr/^{86}Sr$, but no marker, by itself, can differentiate between all the origins. For this reason, it is more advisable to study the complete isotopic profile applying multivariate techniques such as PLS-DA. This approach provides a more comprehensive understanding of the samples by utilizing a broader range of information, thus enabling better discrimination. Additionally, it permits the identification of variables with the highest discriminant power. Indeed, one of the primary objectives of this study was to identify highly promising isotopic markers that can be utilized in future classification models, with a focus on using the minimum number of variables that demonstrate the highest discrimination power. This approach was aimed at streamlining the analytical procedure while maximizing accuracy. To accomplish this, the regression coefficients of the prospective PLS-DA classification models, developed with hazelnut samples from four distinct origins, were evaluated.

A multiclass PLS-DA model was built to discriminate samples according to their geographical origin, including light element bulk $\delta^{2}H$, $\delta^{13}C$, $\delta^{18}O$, and $\delta^{2}H$, $\delta^{13}C$ of the FAMEs as the variables. Elements whose isotopic value could potentially be influenced by specific agronomic practices or demonstrated negligible variation in relation to hazelnut origin, such as N and S (Section 3.2), were omitted from the model to prevent any potential bias. Moreover, to focus first on light elements alone, Sr was also excluded from modelling at this stage. The PLS-DA model cross-validation results provided an 85 % of global correct classification, with percentages higher than 70 % for each individual class. Only two samples were misclassified, while four samples were not assigned to any category (Table 3**a**). Permutation tests, which display the prediction capacity of 20 random models, and ANOVA of the cross-validation ($p < 0.05$) indicated the absence of a random classification and of model overfitting.

PLS regression coefficients (Fig. 3**a**) indicated that among bulk isotopic markers, only $\delta^{18}O$ seemed to contribute relevantly to the discrimination between geographical origins, showing significant positive correlation with Chile and Italy classes and significant negative correlation with Georgia and Spain ones. The $\delta^{2}H$ and $\delta^{13}C$ of some of the main FAMEs ($\delta^{2}H$ of palmitic and oleic acid; $\delta^{13}C$ of linoleic acid) were also relevant for the discrimination of Italian and Spanish samples, respectively. Neither bulk $\delta^{13}C$ or $\delta^{2}H$ seemed to be relevant for the classification of any of the tested hazelnut origins. The mean values of the bulk $\delta^{2}H$ and $\delta^{2}H$ of the FAMEs of each origin (Table 2) were coherent with the isotopic ratios of the precipitations at the specified locations: the Georgian fruits and water ($\delta^{2}H_{water} = -53$ ‰) exhibited the most depleted values in ^{2}H, followed by Chile ($\delta^{2}H_{water} = -41$ ‰), Italy ($\delta^{2}H_{water} = -38$ ‰) and Spain ($\delta^{2}H_{water} = -34$ ‰) (Bowen & Revenaugh, 2003; Bowen, 2024; IAEA/WMO, 2015). Conversely, the $\delta^{18}O$ mean values of each origin did not presented a clear correlation with precipitation (Georgia $\delta^{18}O_{water} = -8.4$ ‰; Chile $\delta^{18}O_{water} = -6.4$ ‰; Italy $\delta^{18}O_{water} = -6.2$ ‰; Spain $\delta^{18}O_{water} = -5.6$ ‰) (Bowen et al., 2024; Bowen and Revenaugh, 2003; IAEA/WMO, 2015). This observation may be attributed to the fact that organic hydrogen exclusively derives from water in the hydrosphere, whereas oxygen can originate from various sources, including atmospheric oxygen and photosynthesized carbon dioxide (De Rijke et al., 2016).

Table 3
Results of the leave-10 %-out cross-validation of the: a) Global PLS-DA classification model that included the variables $\delta^{2}H$, $\delta^{13}C$, $\delta^{18}O$, and $\delta^{2}H$, $\delta^{13}C$ of the main FAMEs; b) PLS-DA classification model that included the variables $\delta^{18}O$, and $\delta^{2}H$, $\delta^{13}C$ of the main FAMEs and c) PLS-DA classification model that included the variables Sr^{87}/Sr^{86}, $\delta^{18}O$, and $\delta^{2}H$, $\delta^{13}C$ of the main FAMEs.

a)

	n	Chile	Georgia	Italy	Spain	Correct class (%)	No Class*
Chile	10	9	0	0	0	90	1
Georgia	10	0	10	0	0	100	0
Italy	10	0	0	7	1	70	2
Spain	10	0	0	1	8	80	1
Total	40					**85**	

N = 40, 5 LVs, $Q^2 = 0.38$, RMSEcv = 0.43, ANOVA p-value < 0.05.* Not assigned samples (PV < 0.5)

b)

	n	Chile	Georgia	Italy	Spain	Correct class (%)	No Class*
Chile	10	7	0	0	0	70	3
Georgia	10	0	10	0	0	100	0
Italy	10	0	0	9	0	90	1
Spain	10	0	0	1	8	80	1
Total	40					**85**	

N = 40, 4 LVs, $Q^2 = 0.45$, RMSEcv = 0.38, ANOVA p-value <0.05.* Not assigned samples (PV < 0.5)

c)

	n	Chile	Georgia	Italy	Spain	Correct class (%)	No Class*
Chile	10	10	0	0	0	100	0
Georgia	10	0	10	0	0	100	0
Italy	10	1	0	8	1	80	1
Spain	10	0	0	0	9	90	1
Total	40					**92.5**	

N = 40, 4 LVs, $Q^2 = 0.62$, RMSEcv = 0.34, ANOVA p-value < 0.05.* Not assigned samples (PV < 0.5).

3.3.2. Optimised PLS-DA model based on light element isotopic composition

A new PLS-DA model was built with the most relevant variables not influenced by other factors: $\delta^{18}O$, and $\delta^{2}H$, $\delta^{13}C$ of the FAMEs. The absence of random classification and model overfitting was evidenced by the permutation tests and ANOVA ($p < 0.05$).

The performance of the new simplified model was equivalent to the previous one (Table 3**b**), yielding slightly better and slightly worse results for Italian and Chilean samples, respectively, but also achieving an 85 % of global correct classification. This confirms that isotopic markers omitted in this simplified model did not play a crucial role in the discrimination. Besides, only one sample was misclassified, and five were not classified.

The assessment of the regression coefficients assessment demonstrated that most of the included variables were relevant in discriminating at least one origin (Fig. 3**b**). For the classification of Georgian, Chilean and Italian samples, the $\delta^{18}O$ was a relevant variable, negatively correlated with the first one and positively correlated with the other two origins. The $\delta^{2}H$ of FAMEs was significant for the samples from Spain and Chile. Additionally, for this last origin and for Italian samples, $\delta^{13}C$ of the FAMEs was also a relevant variable.

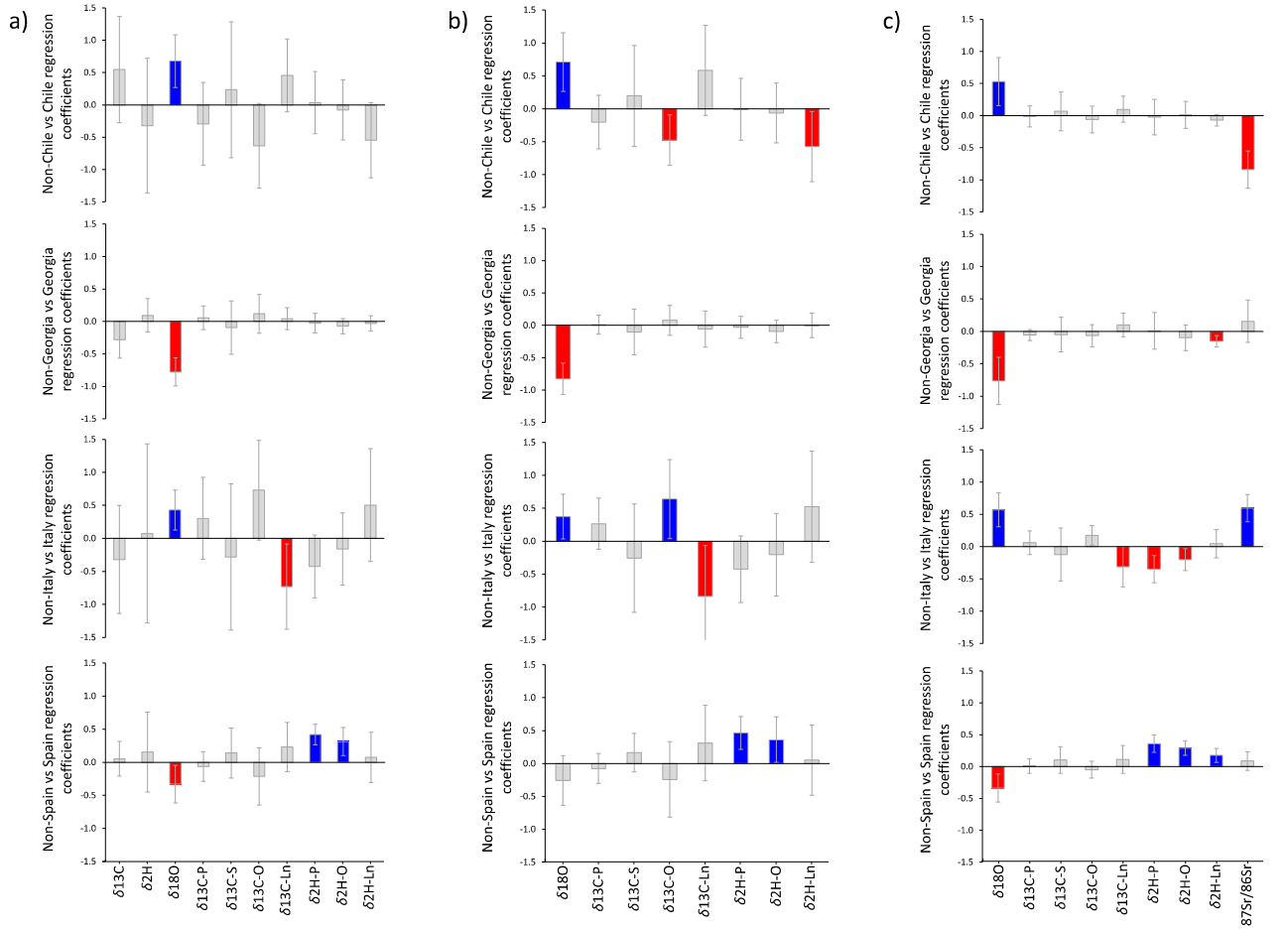


Fig. 3. Regression coefficients of the: a) Global PLS-DA model developed including the variables δ^2H, δ^{13}C, δ^{18}O, and δ^2H, δ^{13}C of the main FAMEs (δ^2H of palmitic δ^2H-P, oleic- δ^2H-O, and linoleic acids δ^2H-Ln; δ^{13}C of palmitic δ^{13}C-P, stearic δ^{13}C-S, oleic δ^{13}C-O and linoleic acids δ^{13}C-Ln); b) PLS-DA model developed including the variables δ^{18}O, and δ^2H, δ^{13}C of the main FAMEs and c) PLS-DA classification model that included the variables Sr^{87}/Sr^{86}, δ^{18}O, and δ^2H, δ^{13}C of the main FAMEs. Significant coefficients are highlighted in blue or red colour ('Chile' (blue) vs 'non-Chile' (red), 'Georgia' (blue) vs 'non-Georgia' (red), 'Italy' (blue) vs 'non-Italy' (red) and 'Spain' (blue) vs 'non-Spain' (red).

3.3.3. PLS-DA model including relevant light element and Sr isotopic composition

Isotopic composition of Sr is expected to provide a strong correlation with the geographical origin of agricultural products, as it is influenced by the geological characteristics of the soil. Indeed, the results obtained in the present study (presented in section 3.2), demonstrated that $^{87}Sr/^{86}Sr$ composition enabled to clearly distinguish Chilean and Georgian hazelnuts among them and from Spanish and Italian hazelnuts (Fig. 2). It is important to note that isotopic analysis of Sr involves labour-intensive purification steps and higher operating costs compared to analysing more abundant light elements (Katerinopoulou et al., 2020; Laursen et al., 2016), making this marker difficult to apply for routine analysis. However, it can be evaluated as a confirmatory tool for samples that resulted uncertain or unclassified by the model based on light element isotopic data, thus limited to a smaller number of samples.

To evaluate the suitability of $^{87}Sr/^{86}Sr$ composition as such a confirmatory tool, a classification model was developed including the relevant light isotopic markers and $^{87}Sr/^{86}Sr$ composition. The classification results were improved including the $^{87}Sr/^{86}Sr$ (Table 3**c**), achieving a 92.5 % of global correct classification and classifying three of the five samples that resulted as not classified in the previous model. In this way, for each individual class, the percentages of classification were higher than 80 %. ANOVA and permutation tests showed that the model had a high discrimination capacity and was not overfitted.

The evaluation of the model regression coefficients (Fig. 3**c**) evidenced that most of the included variables were still significant for one or more origins, being the $^{87}Sr/^{86}Sr$ especially relevant for the Chilean (negative correlation) and Italian (positive correlation) hazelnuts.

These results demonstrate the potential use of $^{87}Sr/^{86}Sr$ as confirmatory tool to classify uncertain samples.

4. Conclusions

Multiple specific isotopic markers were assessed in this prospective study as potential tools to authenticate hazelnut geographical origin. Based on the obtained results, it can be concluded that the most promising variables are bulk δ^{18}O, and δ^2H, δ^{13}C of the main FAMEs, since they are minimally influenced by external factors such as fertilisation treatment and have shown to be relevant in discriminating among geographical origins. On the other hand, the analysis of $^{87}Sr/^{86}Sr$, could be valuable to discriminate uncertain samples or as a confirmatory tool. These findings serve as a starting point for the development of efficient and robust models based on large-scale datasets, which should be further validated to assess their efficiency as hazelnut geographical

authentication tools.

5. Funding

This work was developed in the context of the project TRACENUTS, PID2020-117701RB100 financed by MCIN/AEI/https://doi.org/10.13039/501100011033. MAiMA group is part of MAGH project (2021-SGR-00308), financed by the AGAUR (Catalan Government). B. Torres-Cobos thanks the Spanish Ministry of Universities predoctoral fellowships FPU20/014540. A. Tres received a Ramon y Cajal grant (RYC-2017–23601) funded by MCIN/AEI/ https://doi.org/10.13039/501100011033 and by "ESF Investing in your future".

CRediT authorship contribution statement

B. Torres-Cobos: Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft. **M. Rosell:** Conceptualization, Data curation, Investigation, Methodology, Supervision, Writing – review & editing. **A. Soler:** Supervision, Writing – review & editing. **M. Rovira:** Conceptualization, Resources, Writing – review & editing. **A. Romero:** Conceptualization, Resources, Writing – review & editing. **F. Guardiola:** Supervision, Writing – review & editing. **S. Vichi:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing. **A. Tres:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of competing 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

Data will be made available on request.

Acknowledgments

INSA-UB Maria de Maeztu Unit of Excellence (Grant CEX2021-001234-M) funded by MICIN/AEI/FEDER, UE. INSA-UB authors are part of the LiBiFOOD Consolidated Research Group (2021-SGR-00854) recognised by AGAUR (Catalan Government). The authors would like to express their gratitude to Ferrero Hazelnut Company and Tuscia University (Department of Agriculture and Forest Science) for providing the hazelnut samples from Chile and Georgia and Italy, respectively. They also thank the technicians at MAiMA (Roger Puig, Pedro Valenzuela and Tatiana Palomeque), CCiT-UB (Raul Carrey, Anna Navarro, Pilar Rubio and Pilar Teixidor) and LIRA-PANTHALASSA (Antoni Padró, Leopoldo Pena, Eduardo Paredes and Ester García) for their collaboration and technical support in this project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2024.139083.

References

Anderson, K. A., & Smith, B. W. (2006). Effect of season and Variety on the differentiation of geographic growing origin of pistachios by stable isotope profiling. *Journal of Agricultural and Food Chemistry, 54*, 1747–1752. https://doi.org/10.1021/jf052928m

Azmy, K., Veizer, J., Wenzel, B., Bassett, M. G., & Copper, P. (1999). Silurian strontium isotope stratigraphy. *Geological Society of America Bulletin, 111*, 475–483. https://doi.org/10.1130/0016-7606(1999)111<0475:SSIS>2.3.CO;2

Bateman, A. S., & Kelly, S. D. (2007). Fertilizer nitrogen isotope signatures. *Isotopes in environmental and health studies, 43*(3), 237–247. https://doi.org/10.1080/10256010701550732

Bowen, G. J. (2024). The Online Isotopes in Precipitation Calculator, version 3.2. http://www.waterisotopes.org.

Bowen, G. J., & Revenaugh, J. (2003). Interpolating the isotopic composition of modern meteoric precipitation. *Water Resources Research, 39*(10), 1299. https://doi.org/10.1029/2003WR002086

Bertoldi, D., Cossignani, L., Blasi, F., Perini, M., Barbero, A., Pianezze, S., & Montesano, D. (2019). Characterization and geographical traceability of italian goji berries. *Food Chemistry, 275*, 585–593. https://doi.org/10.1016/j.foodchem.2018.09.098

Bontempo, L., Paolini, M., Franceschi, P., Ziller, L., García-González, D. L., & Camin, F. (2019). Characterisation and attempted differentiation of european and extra-european olive oils using stable isotope ratio analysis. *Food Chemistry, 276*, 782–789. https://doi.org/10.1016/j.foodchem.2018.10.077

Branch, S., Burke, S., Evans, P., Fairman, B., & Wolldd-Briche, C. S. (2003). A preliminary study in determining the geographical origin of wheat using isotope ratio inductively coupled plasma mass spectrometry with 13C, 15N mass spectrometry. *Journal of Analytical Atomic Spectrometry, 18*, 17–22. https://doi.org/10.1039/B207055N

Brescia, M. A., Di Martino, G., Guillou, C., Reniero, F., Sacco, A., & Serrra, F. (2002). Differentiation of the geographical origin of durum wheat semolina samples on the basis of isotopic composition. *Rapid Communications in Mass Spectrometry, 16*, 2286–2290. https://doi.org/10.1002/rcm.860

Camin, F., Boner, M., Bontempo, L., Fauhl-Hassek, C., Kelly, S. D., Riedl, J., & Rossmann, A. (2017). Stable isotope techniques for verifying the declared geographical origin of food in legal cases. *Trends in Food Science & Technology, 61*, 176–187. https://doi.org/10.1016/j.tifs.2016.12.007

Carter, J. F., & Fry, B. (2013). Ensuring the reliability of stable isotope ratio data—beyond the principle of identical treatment. *Analytical and Bioanalytical Chemistry, 405*, 2799–2814. https://doi.org/10.1007/s00216-012-6551-0

De Rijke, E., Schoorl, J. C., Cerli, C., Vonhof, H. B., Verdegaal, S. J. A., Vivó-Truyols, G., Lopatka, M., Dekter, R., Bakker, D., Sjerps, M. J., Ebskamp, M., & De Koster, C. G. (2016). The use of δ2H and δ18O isotopic analyses combined with chemometrics as a traceability tool for the geographical origin of bell peppers. *Food Chemistry, 204*, 122–128. https://doi.org/10.1016/j.foodchem.2016.01.134

Di Pierro, E. A., Ziller, L., Tonon, A., Bianco, L., Franceschi, P., Troggio, M., & Camin, F. (2018, June). *Characterization and valorisation of the Italian walnut (Juglans regia L.): a first application of stable isotope ratio analysis to determine walnut geographical origin.* Presentation at the 2nd Isotope Ratio MS Day, Messina, Italy, Società Chimica Italiana. http://hdl.handle.net/10449/49708.

Drivelos, S. A., & Georgiou, C. A. (2012). Multi-element and multi-isotope-ratio analysis to determine the geographical origin of foods in the European Union. *Trends in Analytical Chemistry, 40*, 38–51. https://doi.org/10.1016/j.trac.2012.08.003

Epova, E. N., Bérail, S., Séby, F., Vacchina, V., Bareille, G., Médina, B., Sarthou, L., & Donard, O. F. X. (2019). Strontium elemental and isotopic signatures of Bordeaux wines for authenticity and geographical origin assessment. *Food Chemistry, 294*, 35–45. https://doi.org/10.1016/j.foodchem.2019.04.068

FAOSTAT, Food and Agriculture Organization of the United Nations, agricultural data 2021 http://www.fao.org/faostat/en/ (accessed 24/03/2023).

Giannioti, Z., Ogrinc, N., Suman, M., Camin, F., & Bontempo, L. (2024). Isotope ratio mass spectrometry (IRMS) methods for distinguishing organic from conventional food products: A review. *Trends in Analytical Chemistry, 170, 117476*. https://doi.org/10.1016/j.trac.2023.117476. ISSN 0165–9936.

Hrastar, R., Petrišič, M. G., Ogrinc, N., & Košir, I. J. (2009). Fatty acid and stable Carbon isotope Characterization of Camelina sativa oil: Implications for authentication. *Journal of Agriculture and Food Chemistry, 57*, 579–585. https://doi-org.sire.ub.edu/10.1021/jf8028144.

IAEA/WMO (2015). Global Network of Isotopes in Precipitation. The GNIP Database. https://nucleus.iaea.org/wiser (accessed 22/01/2024).

Jones, C. E., Jenkyns, H. C., Coe, A. L., & Stephen, H. P. (1994). Strontium isotopic variations in jurassic and cretaceous seawater. *Geochimica et Cosmochimica Acta, 58*, 3061–3074. https://doi.org/10.1016/0016-7037(94)90179-1

Katerinopoulou, K., Kontogeorgos, A., Salmas, C. E., Patakas, A., & Ladavos, A. (2020). Geographical origin authentication of Agri-food products: A review. *Foods, 9*, 489. https://doi.org/10.3390/foods9040489

Kelly, S., Heaton, K., & Hoogewerff, J. (2005). Tracing the geographical origin of food: The application of multi-element and multi-isotope analysis. *Trends in Food Science & Technology, 16*, 555–567. https://doi.org/10.1016/j.tifs.2005.08.008

Król, K., & Gantner, M. (2020). Morphological traits and chemical composition of hazelnut from different geographical origins: A review. *Agriculture, 10*, 375. https://doi.org/10.3390/agriculture10090375

Krauß, S., Vieweg, A., & Vetter, W. (2020). Stable isotope signatures (δ2H-, δ13C-, δ15N-values) of walnuts (Juglans regia L.) from different regions in Germany. *Journal of the Science of Food and Agriculture, 100*, 1625–1634. https://doi.org/10.1002/jsfa.10174

Laursen, K. H., Bontempo, L., Camin, F., & Roßmann, A. (2016). Advances in isotopic analysis for food authenticity testing. In G. In, & D. G. Woodhead (Eds.), *Publishing Series in Food Science, Technology and Nutrition, Advances in Food Authenticity Testing* (pp. 227–252). Woodhead Publishing. https://doi.org/10.1016/B978-0-08-100220-9.00009-6.

Laursen, K. H., Mihailova, A., Kelly, S. D., Epov, V. N., Bérail, S., Schjoerring, J. K., Donard, O. F. X., Larsen, E. H., Pedentchouk, N., Marca-Bell, A. D., Halekoh, U., Olesen, J. E., & Husted, S. (2013). Is it really organic? – multi-isotopic analysis as a tool to discriminate between organic and conventional plants. *Food Chemistry, 141*, 2812–2820. https://doi.org/10.1016/j.foodchem.2013.05.068

Luykx, D. M. A. M., & van Ruth, S. M. (2008). An overview of analytical methods for determining the geographical origin of food products. *Food Chemistry, 107*, 897–911. https://doi.org/10.1016/j.foodchem.2007.09.038

Mahalovich, M. F., Kimsey, M. J., Fortin-Noreus, J. K., & Robbins, C. T. (2016). Isotopic heterogeneity in whitebark pine (Pinus albicaulis engelm.) nuts across geographic, edaphic and climatic gradients in the northern rockies (USA). *Forest Ecology and Management, 359*, 174–189. https://doi.org/10.1016/j.foreco.2015.09.047

Nier, A. O. (1938). The isotopic constitution of strontium, Barium, bismuth, thallium and mercury. *Physical Review Journals, 54*, 275–278. https://doi.org/10.1103/PhysRev.54.275

Regulation (EU) No 1169/2011 on the provision of food information to consumers (2011).

Perini, M., Giongo, L., Grisenti, M., Bontempo, L., & Camin, F. (2018). Stable isotope ratio analysis of different european raspberries, blackberries, blueberries, currants and strawberries. *Food Chemistry, 239*, 48–55. https://doi.org/10.1016/j.foodchem.2017.06.023

Podio, N. S., Baroni, M. V., Badini, R. G., Inga, M., Ostera, H. A., Cagnoni, M., Gautier, E. A., García, P. P., Hoogewerff, J., & Wunderlin, D. A. (2013). Elemental and isotopic fingerprint of argentinean wheat. matching soil, water, and crop composition to differentiate provenance. *Journal of Agricultural and Food Chemistry, 61*(16), 3763–3773. https://doi.org/10.1021/jf305258r

Qi, H., & Coplen, T. B. (2011). Investigation of preparation techniques for δ2H analysis of keratin materials and a proposed analytical protocol. *Rapid Communications in Mass Spectrometry, 25*, 2209–2222. https://doi.org/10.1002/rcm.5095

Sauer, P. E., Schimmelmann, A., Sessions, A. L., & Topalov, K. (2009). Simplified batch equilibration for D/H determination of non-exchangeable hydrogen in solid organic material. *Rapid Communications in Mass Spectrometry, 23*, 949–956. https://doi.org/10.1002/rcm.3954

Sebilo, M., Mayer, B., Nicolardot, B., Pinay, G., & Mariotti, A. (2013). Long-term fate of nitrate fertilizer in agricultural soils. *Proceedings of the National Academy of Sciences, 110*, 18185-18189. https://doi. org/10.1073/pnas.1305372110, 18185–9.

Techer, I., Medini, S., Janin, M., & Arregui, M. (2017). Impact of agricultural practice on the sr isotopic composition of food products: Application to discriminate the geographic origin of olives and olive oil. *Applied Geochemistry, 82*, 1–14. https://doi.org/10.1016/j.apgeochem.2017.05.010

Vitòria, L., Otero, N., Soler, A., & Canals, A. (2004). Fertilizer Characterization: Isotopic data (N, S, O, C, and sr). *Environmental Science & Technology, 38*, 3254–3262. https://doi.org/10.1021/es0348187

Yoneyama, T., Fujiwara, H., & Engelaar, W. M. H. G. (2000). Weather and nodule mediated variations in δ13C and δ15N values in field-grown soybean (Glycine max L.) with special interest in the analyses of xylem fluids. *Journal of Experimental Botany, 51*, 559–566. https://doi-org.sire.ub.edu/10.1093/jexbot/51.344.559.

Zannella, C., Carucci, F., Aversano, R., Prohaska, T., Vingiani, S., Carputo, D., & Adamo, P. (2017). Genetic and geochemical signatures to prevent frauds and counterfeit of high-quality asparagus and pistachio. *Food Chemistry, 237*, 545–552. https://doi.org/10.1016/j.foodchem.2017.05.158

Zhu, Y., Hioki, A., & Chiba, K. (2014). Measurement of strontium isotope ratio in nitric acid extract of peanut testa by ICP-Q-MS after removal of rb by extraction with pure water. *Talanta, 119*, 596–600. https://doi.org/10.1016/j.talanta.2013.11.058

5.2 Enhancement and validation of isotopic methods for hazelnut geographical authentication

5.2.1 Multi-isotopic approach

The results of the PoC study (Publication 4) demonstrated that bulk $\delta^{18}O$, and $\delta^{2}H$, $\delta^{13}C$ of the FAMEs, were the most promising isotopic markers for hazelnut geographical authentication, while Sr could be proposed as a possible confirmatory marker for uncertain samples. These markers were not influenced by agronomic practices and proved to be the most relevant isotopic variables in differentiating hazelnuts according to their geographical area of origin.

However, their effectiveness as a tool for hazelnut origin authentication needed to be properly evaluated on larger sample sets, including more samples from each category, multiple harvest years, and various production areas. Therefore, a subsequent study was conducted with an enriched sample set to build and externally validate PLS-DA classification models for discriminating between different hazelnut provenances (**Figure 10**).

This sample set included all TG samples conforming an expanded hazelnut sample set, that included 207 samples of TG cultivar hazelnuts across four regions in different countries (CHL, n=40), Spain (ESP, n=91), Georgia (GEO, n=40), and Italy (ITA, n=36). Samples were collected over four harvest seasons (2019-2022). The models' performance was assessed in terms of fitting and classification accuracy in external validation (Publication 5).

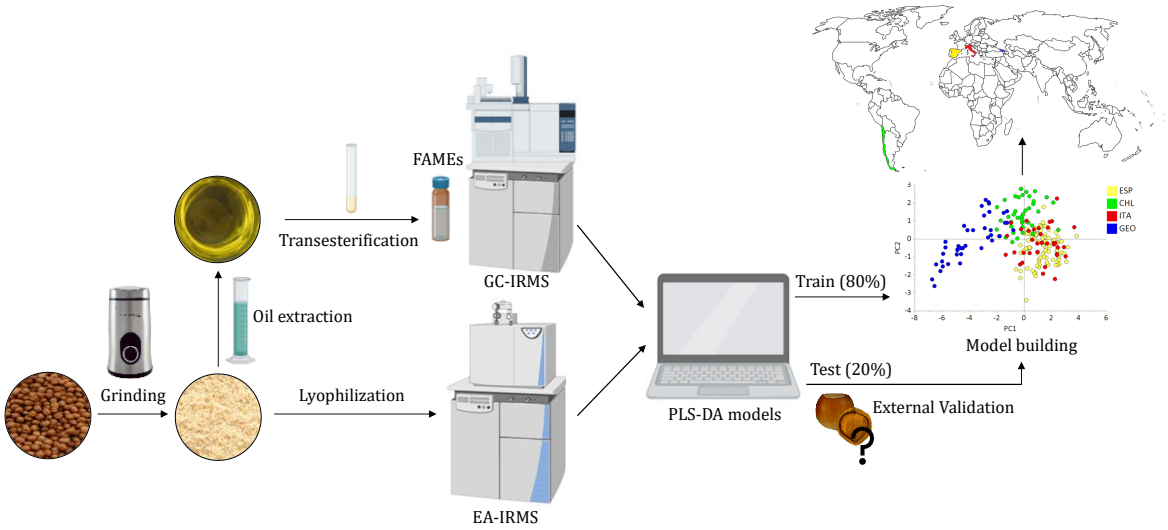


Figure 10. Graphical abstract of Publication 5. FAMEs: fatty acid methyl esters, GC-IRMS: gas chromatography-isotope ratio mass spectrometry, EA-IRMS: elemental analysis-isotope ratio mass spectrometry, PLS-DA: partial least square-discriminant analysis, CHL: Chile, ESP: Spain, GEO: Georgia, ITA: Italy.

5.2.2 Publication 5

A multi-isotopic chemometric approach for tracing hazelnuts origin

Berta Torres-Cobos, Mònica Rossell, Albert Soler, Mercè Rovira, Agustí Romero, Francesc Guardiola, Stefania Vichi, Alba Tres

*Foods, **2024,** 13(21), 3399. https://doi.org/10.3390/foods13213399*

Supplementary material available in **Annex 4**

foods

MDPI

Article

A Multi-Isotopic Chemometric Approach for Tracing Hazelnut Origins

Berta Torres-Cobos [1,2], Mònica Rosell [3], Albert Soler [3], Mercè Rovira [4], Agustí Romero [4], Francesc Guardiola [1,2], Stefania Vichi [1,2,*] and Alba Tres [1,2]

[1] Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Universitat de Barcelona, Av Prat de La Riba, 171, 08921 Santa Coloma de Gramenet, Spain; bertatorres@ub.edu (B.T.-C.); fguardiola@ub.edu (F.G.); atres@ub.edu (A.T.)
[2] Institut de Recerca en Nutrició i Seguretat Alimentària (INSA-UB), Universitat de Barcelona, Av Prat de La Riba, 171, 08921 Santa Coloma de Gramenet, Spain
[3] Grup MAiMA, Mineralogia Aplicada, Geoquímica i Hidrogeologia, Departament de Mineralogia, Petrologia i Geologia Aplicada, Institut de Recerca de l'Aigua (IdRA), Universitat de Barcelona, Martí i Franqués s/n, 08028 Barcelona, Spain; monica.rosell@ub.edu (M.R.); albertsolergil@ub.edu (A.S.)
[4] Institute of Agrifood Research and Technology (IRTA), Ctra. de Reus–El Morell Km 3.8, 43120 Constantí, Spain; merce.rovira58@gmail.com (M.R.); agusti.romero@irta.cat (A.R.)
* Correspondence: stefaniavichi@ub.edu

Abstract: High-value products, such as hazelnuts, are particularly vulnerable to fraud due to their price dependence on geographical origin. Guaranteeing hazelnuts' authenticity is essential for consumer trust and safety. Stable isotope analysis has become a reference method for origin authentication as it is reliable, robust, and easily transferable across laboratories. However, multiple isotopic markers coupled with chemometric techniques are often needed to authenticate food provenance accurately. In this study, we focused on assessing the potential of bulk $\delta^{18}O$, along with $\delta^{2}H$ and $\delta^{13}C$ of the main fatty acids, as hazelnut-origin authenticity markers. PLS-DA classification models were developed to differentiate samples (n = 207) according to their region of origin. This multi-isotopic approach provided promising external validation results, achieving a 94% global correct classification rate in discriminating hazelnuts from regions with distinct geographical and environmental conditions. This study lays the groundwork for further model development and evaluation across additional production areas and harvest years.

Keywords: hazelnut; stable isotopes; geographical authentication; PLS-DA; food fraud

Citation: Torres-Cobos, B.; Rosell, M.; Soler, A.; Rovira, M.; Romero, A.; Guardiola, F.; Vichi, S.; Tres, A. A Multi-Isotopic Chemometric Approach for Tracing Hazelnut Origins. *Foods* **2024**, *13*, 3399. https://doi.org/10.3390/foods13213399

Academic Editors: Biagia Musio and Vito Gallo

Received: 18 September 2024
Revised: 17 October 2024
Accepted: 22 October 2024
Published: 25 October 2024

1. Introduction

Ensuring the authenticity of food products is paramount for fostering consumer trust and safety throughout the food chain. Verifying authenticity and traceability demands robust analytical methods, which is especially crucial for regulatory bodies. Over the past few decades, stable isotope analysis has progressively consolidated its position as a reference authentication method, as isotopic data can be reliably measured in routine work and effectively compared across laboratories. This reliability has also led to the utilization of these methods in legal cases as part of enforcement exercises [1]. According to various studies, stable isotope analysis is effective at detecting adulteration and verifying farming practices in various food products [2–5], as well as in determining the geographical origin of food commodities [1,6–11], which is a crucial aspect of assessing food authenticity. The effectiveness of isotopic measurements in verifying food's geographical origin stems from the dependence of isotopic ratios on the geological characteristics and environmental conditions of the area where it was produced. Regarding plant-derived food specifically, the isotopic ratio of light elements ($\delta^{13}C$, $\delta^{2}H$, and $\delta^{18}O$) is significantly shaped by geological and environmental conditions, such as soil composition and precipitation

patterns [4,6–10,12]. The $\delta^{13}C$ ratio in plants is closely related to their botanical origin, allowing discrimination between C3 and C4 plants, with the former being more depleted in the heavier isotope [13,14]. However, other studies have shown that it might be influenced by several environmental factors such as temperature, precipitation, water stress, humidity, and ambient CO_2 concentration [15–17]. The δ^2H and $\delta^{18}O$ values in plants are determined by the meteoric water isotopic ratios of the specific cultivation territory. These ratios result from isotopic fractionation processes occurring during water phase changes (solid–liquid–vapor) such as evaporation, condensation, and precipitation, which are driven by factors such as latitude, altitude, and temperature [6,8]. Nevertheless, while organic hydrogen (H) in plants derives solely from water in the hydrosphere, oxygen (O) can originate from other sources, such as the atmosphere or photosynthesized CO_2 [8,10]. Therefore, numerous environmental and climatic factors can impact the $\delta^{18}O$ ratio, rendering it a particularly promising candidate for geographical authentication.

Recent studies have shifted their focus beyond the bulk analysis of light bio-elements to include compound-specific isotopic data [4,10]. This approach allows for more insight into isotopic data by assessing the isotopic ratios of elements in specific food components such as proteins, lipids, or carbohydrates [4]. This strategy has been applied for origin authentication in a wide number of foods, including honey [18], olive oil [10,19], pumpkin [20], and shitake mushrooms [21], among others.

Geographical origin determination often requires considering multiple stable isotope parameters [1]. In fact, in most cases, a single isotopic marker is insufficient to authenticate food origin [7,9,10,22], but the combination of multiple isotopic markers can satisfactorily achieve this purpose, offering more comprehensive information and enhancing method accuracy.

In addition to selecting the appropriate isotopic markers, a crucial aspect of the authentication strategy lies in data analysis, especially when dealing with multi-isotope ratios. Supervised chemometric techniques, such as partial least squares discriminant analysis (PLS-DA), offer an advantageous alternative to unsupervised methods, enabling the construction of classification models and the accurate prediction of unknown samples. Specifically, PLS-DA is focused on finding the maximum correlation between data and the selected class variable, minimizing those variables not related to the selected category. It has been proven to be an efficient tool for authentication purposes [23–25]. Various studies have successfully applied multi-isotope analysis combined with chemometrics to verify the origin of various food products, aiming to protect consumers from misleading information [6–9,11,23–25].

Given that market prices frequently rely on the geographical area of production, certain high-value products become particularly vulnerable to fraudulent practices, such as falsifying a label's declared origin. This is motivated by the potential for illicit profit and is often aggravated by a lack of efficient authentication methods. This is the case with hazelnuts, whose price and sensorial characteristics are highly dependent on their region of origin [26,27]. These nuts are highly valued in the food industry, both in their raw and processed forms. Italian and Spanish hazelnuts, in particular, present market prices up to twice as high as those from other origins [27], with hazelnuts under quality schemes, such as Protected Denominations of Origin and Protected Geographical Indications, further enhancing their value [28].

Few studies have addressed nut authenticity through isotope-based methods. Promising findings have been obtained from analyzing the isotopic ratios of bio-elements in pistachios [29], walnuts [12,30,31], peanuts [32], and pine nuts [33]. To the best of the authors' knowledge, only Sammarco et al. [34] have applied isotopic markers to authenticate the origin of hazelnuts, combining this approach with multi-elemental analysis. To further explore the potential of isotopic data in verifying hazelnuts' origin, preliminary research was carried out to optimize the analytical procedures and to identify the most suitable isotopic markers for implementation in a multi-isotopic approach [35]. This prospective study, conducted on a restricted sample set, revealed that bulk $\delta^{18}O$, together with δ^2H

Foods **2024**, *13*, 3399

and $\delta^{13}C$ measured in the main fatty acid methyl esters (FAMEs) were the most relevant in discriminating hazelnut samples from different origins when excluding isotopic markers influenced by external factors such as fertilization treatment. Therefore, exploring the potential of multi-isotopic markers by constructing and testing origin classification models based on large sample sets, including different provenances and harvest years, could provide new, reliable tools to ensure hazelnut origin.

Considering this previous knowledge, the main aim of the present study is to explore the potential of several isotopic markers that have formerly shown promise as hazelnut geographical authentication tools and to assess their suitability. For this purpose, PLS-DA classification models based on bulk $\delta^{18}O$, along with $\delta^{2}H$, and $\delta^{13}C$ of the main FAMEs were developed to discriminate hazelnuts according to their region of origin. These models were built on a sample set of 207 hazelnuts from four different geographical origins and evaluated through external validation.

2. Materials and Methods

2.1. Samples

The sample set consisted of 207 'Tonda di Giffoni' hazelnuts from four harvest seasons and four different regions located in Chile (CHL, n = 40; 2019, 2020), Spain (ESP, n = 91; 2019, 2020, 2021, 2022), Georgia (GEO, n = 40; 2021, 2022), and Italy (ITA, n = 36; 2019, 2020, 2021), respectively. Geographical coordinates for each production region are reported in Table 1.

Table 1. Harvest year and geographical coordinates of parcels where hazelnuts were produced.

	2019 (n)	2020 (n)	2021 (n)	2022 (n)	Geographical Coordinates
Chile	20	20	-	-	35°15′36″ S, 71°32′60″ W
Spain	23	23	23	22	41°10′15″ N, 1°10′09″ E
Georgia	-	-	20	20	42°27′34″ N, 41°51′31″ E
Italy	12	12	12	-	42°25′23″ N, 12°4′45″ E

Each hazelnut sample was taken from an individual tree and collected at the time of ripening, with harvests taking place in March–April for CHL and in August–October for ESP, GEO, and ITA. Sample information is summarized in Table S1 of the Supplementary Materials.

Samples were obtained within the TRACENUTS project (PID2020-117701RB-I00) and collected from distinct hazelnut trees by producers. Hazelnuts were shelled at the laboratory, and kernels were stored under vacuum at 4 °C until analysis.

2.2. Bulk Isotopic Analysis by Elemental Analysis–Isotope Ratio Mass Spectrometry (EA-IRMS)

2.2.1. Sample Preparation for Bulk Isotopic Analysis

Sample preparation was carried out, as reported by Torres-Cobos et al. [35]. About 30 g of raw hazelnuts were ground into fine powder using a domestic grinder (Aromatic, Taurus, Oliana, Spain). Then, ground samples were lyophilized for three days in a Telstar Cryodos-45 freeze dryer (Telstar, Terrassa, Spain).

2.2.2. EA-IRMS

Samples were analyzed in the conditions reported by Torres-Cobos et al. [35]. An aliquot of 0.3 mg of the ground sample was placed into silver capsules (Lüdi Swiss, Flawil, Switzerland) and analyzed in a TC/EA-IRMS Delta Plus XP (Thermo Fisher Scientific) equipped with an autosampler (Sample Tray, N° 2, MAS200R autosampler) and a 450 mm ceramic pyrolysis reactor, heated to 1445 °C. The filling of the reactor was carried out according to the Thermo Fisher Scientific manual [35]. Helium was the carrier gas (pressure 62 kPa), and carbon monoxide (CO) was the reference gas with a $\delta^{18}O$-value of −8.68 ‰ (pressure 180 kPa). The certified standard IAEA-601 (benzoic acid, $\delta^{18}O$ = +23.3 ‰) and the

internal secondary standards, UB-YCEM ($\delta^{18}O$ = +17.6 ‰) and UB-ASC ($\delta^{18}O$ = +13.2 ‰), both barium sulfates, were used fitting the range of samples. Samples were analyzed in duplicate.

2.3. Isotopic Analysis of FAMEs by Gas Chromatography–Isotope Ratio Mass Spectrometry (GC-IRMS)

2.3.1. Sample Preparation for Isotopic Analysis of FAMEs

Samples were prepared, as reported by Torres-Cobos et al. [35]. For each sample, the lipid fraction of 25 g of ground hazelnuts was extracted with 50 mL of diethyl ether and evaporated to dryness. Subsequently, FAMEs were prepared, dissolving 100 mg of hazelnut oil in 2 mL of hexane and adding 200 µL of a 2M methanolic potassium hydroxide solution [2]. After centrifugation, the supernatant was analyzed.

2.3.2. GC-IRMS

The analysis was carried out as reported by Torres-Cobos et al. [35]. The H and C isotopes of individual FAMEs were analyzed in duplicate using a Trace GC Ultra gas chromatograph with a Triplus Autosampler, coupled to an Isotope Ratio Mass Spectrometer Delta V Advantage through a GC Isolink interface (Thermo Fisher Scientific, Waltham, MA, USA). The injection volume was 1 µL, with a split ratio of 1:5. The separation of FAMEs was performed on a VF-23ms capillary column (60 × 0.32 mm I.D., 0.15 µm of Agilent Technologies, Santa Clara, CA, USA). The GC oven temperature was initially set at 60 °C, held for 1 min, increased to 160 °C at a rate of 6 °C/min, and held for 10 min. Lastly, it was raised to 240 °C at a rate of 6 °C/min.

Helium was the carrier gas at a flow rate of 1.8 mL/min. The injector temperature was 240 °C. The commercial NiO/CuO-NiO-Pt combustion reactor (P/N 1255321, Thermo Fisher Scientific, Waltham, MA, USA) operated at 1000 °C for CO_2. For the analysis of H_2, the commercial high-temperature reactor (P/N 1255330, Thermo Fisher Scientific, Waltham, MA, USA) was used and adjusted to 1400 °C. The certificated standards used for the analysis of δ^2H were Androstane ($\delta^2H = -293.2 \pm 1.0$ ‰), USGS76 ($\delta^2H = -210.8 \pm 0.9$ ‰), Coumarin ($\delta^2H = 82.3 \pm 1.2$ ‰), and a secondary standard, FAME C19 ($\delta^2H = -215.25 \pm 0.6$ ‰). For the analysis of $\delta^{13}C$, the standards used were Icosane ($\delta^{13}C = -40.91 \pm 0.02$ ‰), FAME C16 ($\delta^{13}C = -30.78 \pm 0.02$ ‰), USGS76 ($\delta^{13}C = -31.36 \pm 0.04$ ‰), FAME C19 ($\delta^{13}C = -30.32 \pm 0.02$ ‰), USGS72 ($\delta^{13}C = -1.54 \pm 0.03$ ‰), and Phenanthrene 16/0020 ($\delta^{13}C = -23.90 \pm 0.14$ ‰).

2.4. Data Treatment and Statistical Analysis

A data matrix was built, with samples in the rows (n = 207) and variables in the columns (bulk $\delta^{18}O$; $\delta^{13}C$ of C16:0, C18:0, C18:1 and C18:2 FAMEs; and δ^2H of C16:0, C18:1 and C18:2 FAMEs). To study the effect of interannual variability and the differences in isotopic ratios across origins, independent sample t-tests and one-way analysis of variance (ANOVA), together with Scheffé's post hoc test, were performed in IBM SPSS Statistics v29.0© (IBM Corp., Armonk, NY, USA) for isotopic ratios that followed a normal distribution according to the Shapiro–Wilk normality test, whereas the Mann–Whitney U and Kruskal–Wallis tests were applied to those variables that did not present a normal distribution to determine if there were significant differences between the isotopic ratios throughout the different harvest years and origins.

Model development was performed on SIMCA v13.0© (Sartorius, Göttingen, Germany). Principal Component Analysis (PCA) was used for preliminary data exploration and to detect possible outliers based on Hotelling's T^2 range and model residual parameters.

The sample set was randomly split into the training set (80% of the samples of each class, n = 166) and validation set (20% of the samples of each class, n = 41), assuring that samples from all origins and harvest years were included in both sets. This splitting was run seven times (seven iterations) to evaluate the effect of the sample's set composition

Foods **2024**, *13*, 3399

and to increase the robustness of external validation. The sample set splitting information, validation, and training sets are summarized in Table S1 of the Supplementary Materials.

Partial least square–discriminant analysis (PLS-DA) classification models were developed and validated to classify samples according to their geographical origin (SIMCA v13.0©). First, a PLS-DA model was built with each training set (n = 166) to discriminate between the four origins: Chile (CHL, n = 32), Spain (ESP, n = 73), Georgia (GEO, n = 32), and Italy (ITA, n = 29), respectively (see Section 3.1). After evaluating the results from this four-class model, a three-class model was developed to discriminate between CHL, GEO, and a third class that combined both ESP and ITA samples.

In PLS-DA multi-class models, a dummy Y matrix is employed, featuring classification vectors equivalent to the number of classes. Each vector assigns a value of 1 to a specific class (representing a specific region of origin) and 0 to all other classes (representing other regions). Afterward, each sample is categorized into the class associated with the vector yielding the highest PLS predicted value (PV), provided it is higher than the classification threshold. Samples failing to surpass the classification threshold for any vector remain unassigned to any region (no class).

To optimize the classification thresholds and maximize the classification performance, receiver operating characteristics (ROC) analysis was applied to PLS-DA models to set the thresholds of each class. Hence, ROC curves (one for each training model and each class) were built up with the PV obtained in the leave 10% out cross-validation and the real class for each sample in each model. The ROC curve plots the sensitivity and 1-specificity values obtained when the PV threshold that assigns samples to a class varies. Then, the selected threshold for each class and each model corresponds to the one that maximizes the sensitivity and specificity of the classification [36].

Models developed with the training sets were internally validated by the leave 10% out cross-validation method. The optimal number of latent variables and pre-processing were selected according to the lowest Root Mean Squared Error of Cross-Validation (RMSEcv) criteria. Model overfitting was assessed through the permutation test (n = 20 permutations) and the ANOVA of the cross-validated predictive residuals (*p*-value). No outliers were detected according to Hotteling's T^2 range and model residual parameters. For all models, the optimal pre-processing, according to the lowest RMSEcv, was mean-centering and scaling to the unit of variance. The permutation test and ANOVA *p*-value results showed that none of the training models were overfitted.

Finally, the external validation was carried out using each training model to predict the class of samples in the complementary validation set. The performance of the PLS-DA models was evaluated by the Q^2 values and efficiency in external validation, which was expressed as the percentage of correct classification in each class.

3. Results and Discussion

With the aim of developing an efficient multi-isotopic tool for the geographical authentication of hazelnuts, the isotopic markers that demonstrated the highest discriminatory power were selected in the previous prospective study [35], which include bulk $\delta^{18}O$, as well as $\delta^{2}H$ and $\delta^{13}C$ of the main hazelnut FAMEs. By focusing exclusively on the most discriminating markers, this approach streamlines workload optimization and facilitates the method's application in routine controls. To gain insight into the behavior of the different isotopic markers prior to constructing the multi-isotopic model, Table 2 presents their mean and standard deviation results across the four origins and the significant differences identified with the corresponding statistical tests (ANOVA and Kruskal–Wallis). Noteworthy, differences between means according to the origin, particularly for bulk $\delta^{18}O$ and $\delta^{2}H$ of the FAMEs, highlight the discriminatory potential of these markers. However, although the means were significantly different between two or more countries, no individual markers could independently distinguish between all origins. This reinforces the validity of our approach, which relies on a multi-isotopic analysis combined with multivariate techniques to comprehensively assess the origin of hazelnuts. In addition, significant differences were

found between the means of the isotopic ratios of different years for the same geographical region (Table S2 of the Supplementary Materials), confirming the need to include multiple harvest years in the model building in order to consider the interannual variability and increase the robustness of the models in the prediction of future samples.

Table 2. Mean and standard deviation results and differences in isotopic ratios across the various origins.

Isotopic Signatures	Chile (n = 40)	Spain (n = 91)	Georgia (n = 40)	Italy (n = 36)
$\delta^{18}O$ (‰)	23.7 ± 0.7^{b}	23.6 ± 0.8^{b}	19.8 ± 0.5^{c}	24.8 ± 0.8^{a}
$\delta^{13}C_{Palmitic}$ (‰)	-29.9 ± 0.7^{a}	-29.9 ± 0.5^{a}	-31.8 ± 1.2^{b}	-29.9 ± 0.6^{a}
$\delta^{13}C_{Stearic}$ (‰)	-31.0 ± 1.2^{ab}	-30.8 ± 1.1^{a}	-32.8 ± 1.1^{c}	-31.3 ± 1.2^{b}
$\delta^{13}C_{Oleic}$ (‰)	-28.4 ± 0.6^{a}	-28.4 ± 0.5^{a}	-30.3 ± 1.3^{b}	-28.3 ± 0.6^{a}
$\delta^{13}C_{Linoleic}$ (‰)	-30.5 ± 0.7^{a}	-30.5 ± 0.7^{a}	-32.7 ± 1.6^{b}	-30.5 ± 0.9^{a}
$\delta^{2}H_{Palmitic}$ (‰)	-179.0 ± 6.8^{b}	-163.0 ± 8.2^{a}	-186.6 ± 6.9^{c}	-166.5 ± 8.2^{a}
$\delta^{2}H_{Oleic}$ (‰)	-191.3 ± 5.3^{c}	-174.8 ± 5.4^{a}	-197.5 ± 5.3^{d}	-179.4 ± 6.9^{b}
$\delta^{2}H_{Linoleic}$ (‰)	-215.3 ± 5.8^{b}	-199.5 ± 6.7^{a}	-219.4 ± 5.0^{c}	-200.7 ± 7.7^{a}

In all cases, *p*-values were <0.05 (according to ANOVA or Kruskal–Wallis). Significant differences between origins were noted with different superscript letters in the same row as a > b > c > d.

An examination of the score plot of the exploratory PCA (Figure 1) revealed a clear separation between the GEO and CHL clusters, while ESP and ITA samples formed an overlapping cluster. This initial exploration suggested that the isotopic profiles of the hazelnuts from regions in Spain and Italy displayed minimal differences and were unlikely to be easily distinguishable by the PLS-DA classification model. Nevertheless, a PLS-DA to discriminate between the four origins was attempted.

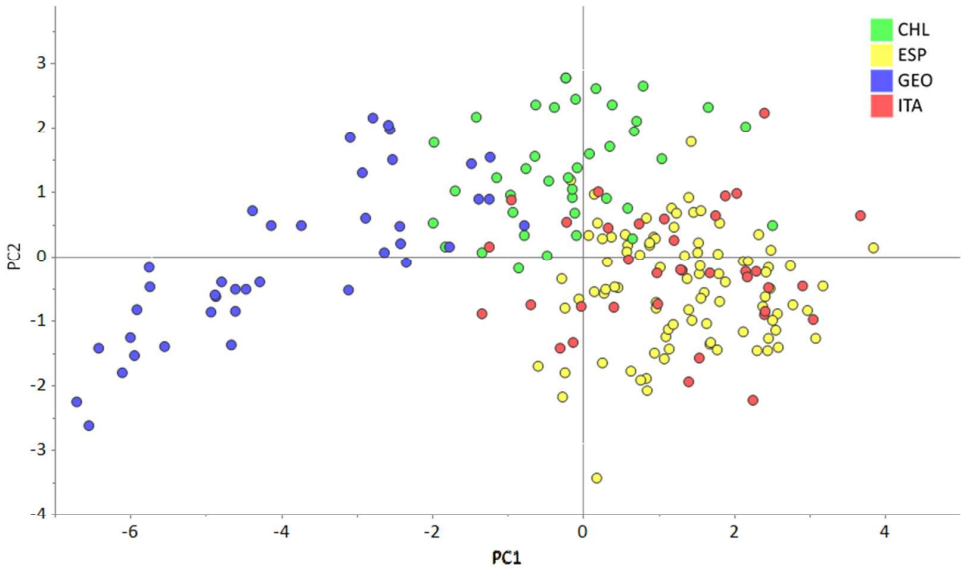


Figure 1. Score plot of the first and second principal components of the PCA model developed using bulk $\delta^{18}O$ data and $\delta^{2}H$, $\delta^{13}C$ data of the main fatty acid methyl esters.

3.1. Four-Class PLS-DA Model

The internal validation results obtained from seven iterations revealed that the four-class PLS-DA model was unable to effectively discriminate among all the tested origins, with correct classification rates of 66.1%, 86.9%, 100%, and 19.7% for CHL, ESP, GEO, and ITA samples, respectively (Table S3 of the Supplementary Materials). In particular,

Foods **2024**, *13*, 3399

the model failed at classifying ITA hazelnuts, predominantly misclassifying them as ESP samples or failing to assign them to any class. This outcome could be due to the similarities between the Spanish and the Italian regions studied. Indeed, both regions share similar climatic and environmental conditions, including proximity to the coast, low altitude, and Mediterranean climate. These similarities are further evidenced by the isotopic ratios of the precipitations in these particular areas: ITA ($\delta^2H_{water} = -38‰$, $\delta^{18}O_{water} = -6.2‰$) and ESP ($\delta^2H_{water} = -34‰$, $\delta^{18}O_{water} = -5.6‰$). However, the values in the GEO regions ($\delta^2H_{water} = -53‰$, $\delta^{18}O_{water} = -8.4‰$) and CHL ($\delta^2H_{water} = -41‰$, $\delta^{18}O_{water} = -6.4‰$) differ considerably [37–39]. The above-mentioned analogies were reflected in the ITA and ESP isotopic profiles, making them indistinguishable from the PLS-DA model. Although stable isotope analysis is a useful tool to discriminate samples from different geographical regions, it becomes challenging when these regions are closely situated and share similar pedoclimatic conditions, latitude, altitude, or proximity to the sea. Despite the fact that distinguishing between similar regions may be challenging, it is valuable to differentiate them from other regions with distinct geographical and environmental conditions. Specifically, counterfeiting Spanish and Italian hazelnuts among themselves, given their similar market prices, would not be as profitable as counterfeiting them with significantly cheaper hazelnuts, such as those from Chile and especially Georgia. Thus, it remains crucial to distinguish between highly priced hazelnuts (ESP and ITA) and hazelnuts from these other regions. For this reason, we merged ESP and ITA classes into one class (ESP + ITA) and built PLS-DA models with three classes to classify samples into GEO, CHL, and ESP + ITA. The optimal thresholds for each class are summarized in Table S4 of the Supplementary Materials. Alternatively, if distinguishing the Spanish region from the Italian region was specifically required, other stable isotopes with strong geogenic connections, such as Strontium (Sr), could additionally be considered according to the promising results from our previous study [35].

3.2. Three-Class PLS-DA Model

The mean results of the leave 10% out cross-validation of the seven training sets are reported in Table S5 of the Supplementary Materials. The internal validation presented promising outcomes, achieving a 95.9% global correct classification and percentages equal to or higher than 87.5% for all the individual classes. Figure 2 displays the score plot of the PLS-DA model. However, the proper validation of authentication models requires external validation, assessed through the prediction of samples that have not been used to build the models, such as a validation set.

External validation results, expressed as the mean and standard deviation obtained from the seven iteration sets, are summarized in Table 3. The model was successful at classifying the samples into three classes: CHL, GEO, and ESP + ITA. A 93.7% global correct classification rate was achieved with a low standard deviation, indicating the good discriminant capacity of the model, with minimal dependency on the composition of the validation set. No annual effects were observed in the misclassified samples as there were wrongly classified samples from all years (Table S6 of Supplementary Materials) despite the interannual differences observed for some isotopic markers and origins (Table S2 of Supplementary Materials).

The model was especially successful in classifying GEO samples (100%), reflecting the significant differences in all isotopic ratios compared to other origins (Table 2), particularly for lower bulk $\delta^{18}O$ values, consistent with the $\delta^{18}O$ precipitation values across the studied areas (GEO $\delta^{18}O_{water} = -8.4‰$; CHL $\delta^{18}O_{water} = -6.4‰$; ITA $\delta^{18}O_{water} = -6.2‰$; ESP $\delta^{18}O_{water} = -5.6‰$) [37–39]. However, while none of these differences alone were sufficient to distinguish GEO from other origins, the combination of all isotopic markers into a PLS-DA model enabled a complete differentiation, further ratifying the hypothesis that a multi-isotopic chemometric approach is crucial for geographic classification.

Foods **2024**, *13*, 3399

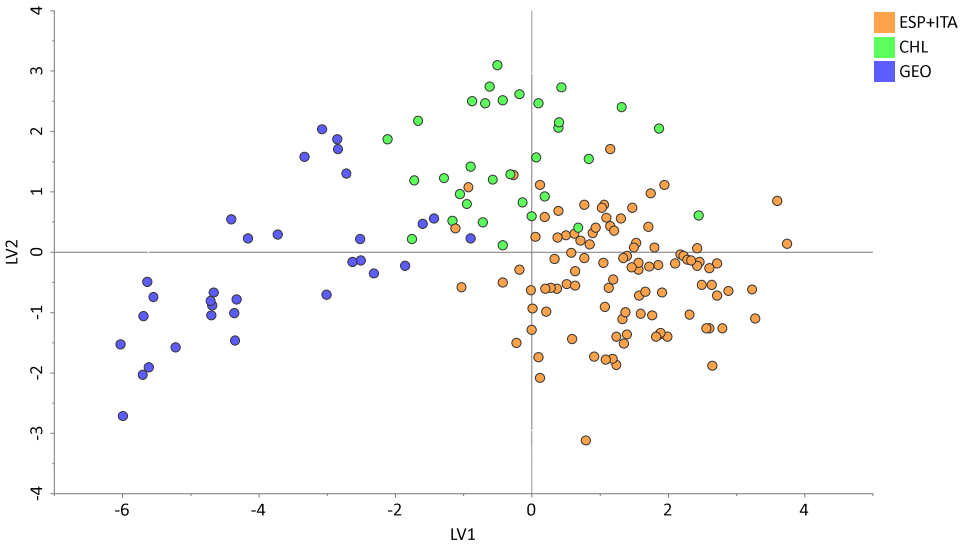


Figure 2. Score plot of the first and second latent variables of the PLS-DA model developed using bulk $\delta^{18}O$ data and $\delta^{2}H$, $\delta^{13}C$ data of the main fatty acid methyl esters.

Table 3. Results of the external validation of three-class PLS-DA models. Results are mean values (± standard deviation) obtained from seven iterations.

True Origins		Origins Assigned by the Model				
	n	CHL (n)	ESP + ITA (n)	GEO (n)	Not Assigned (n)	Correct Classification (%)
CHL	8	6.6 ± 1.1	1.3 ± 1.0	0.0 ± 0.0	0.1 ± 0.5	82.1 ± 14.2
ESP + ITA	25	0.9 ± 0.7	23.9 ± 0.9	0.0 ± 0.0	0.3 ± 0.5	95.4 ± 3.6
GEO	8	0.0 ± 0.0	0.0 ± 0.0	8.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0
Total	41					93.7 ± 4.2

Model parameters are the mean values obtained with the training sets (N = 166) from 7 iterations: 6 LVs, Q^2 = 0.648, and RMSEcv = 0.296. For all models, ANOVA *p*-value < 0.05. CHL: Chile; ESP: Spain; GEO: Georgia; ITA: Italy.

Moreover, the model successfully classified 82.1% of the CHL samples despite the similar $\delta^{18}O$ and $\delta^{13}C$ isotopic ratios of the main FAMEs in the ESP + ITA class, leading to the occasional misclassification of CHL hazelnuts into this class. The successful discrimination of CHL and ESP + ITA samples may be primarily attributed to the notable differences observed in the mean $\delta^{2}H$ values of the FAMEs as all three isotopic ratios ($\delta^{2}H_{Palmitic}$, $\delta^{2}H_{Oleic}$ and $\delta^{2}H_{Linoleic}$) showed significant differences between CHL and ESP + ITA (Table 2). Specifically, and considering associated errors normally <6‰, the $\delta^{2}H_{Oleic}$ values were more strongly enriched (around −175 and −180‰) for the ESP + ITA class than for the CHL class (around −190‰). The same was true for $\delta^{2}H_{Palmitic}$ and $\delta^{2}H_{Linoleic}$ values, approximately −165‰ and −200 ‰ for the ESP + ITA, and −179‰ and −215‰ for the CHL, respectively. These findings align with the $\delta^{2}H$ precipitation values registered in each region, which showed slight variations. CHL exhibited the most depleted $\delta^{2}H$ values ($\delta^{2}H_{water}$ = −41‰) compared to ITA ($\delta^{2}H_{water}$ = −38‰) and ESP ($\delta^{2}H_{water}$ = −34‰) [37–39]. However, besides precipitation, other factors can influence the $\delta^{18}O$ and $\delta^{2}H$ of FAME isotopic ratios. The $\delta^{18}O$ values of the samples are not only derived from water in the hydrosphere but can originate from multiple sources, such as photosynthesized CO_2 or atmospheric O_2. Additionally, although the only source of organic hydrogen for plants comes from water in the hydrosphere, this hydrogen is metab-

olized by different synthetic routes to synthesize fatty acids, resulting in different ratios for each fatty acid [8,40,41].

Although the isotopic ratios of bio-elements might be influenced by many factors, these results show their potential as a tool for verifying the geographical origin of hazelnuts across regions with diverse geological and environmental conditions.

4. Conclusions

In this study, we evaluated the suitability of a multi-isotopic approach combined with chemometrics as a tool for authenticating hazelnut geographical origin. This assessment involved the analysis of bulk and compound-specific isotopic markers (bulk $\delta^{18}O$, together with $\delta^{13}C$ and $\delta^{2}H$ of the FAMEs) using a set of 207 samples from regions located in four different countries and across different harvest seasons.

The results showed that the differences in $\delta^{18}O$ and $\delta^{2}H$ of the FAMEs seemed coherent with the precipitation data of the four territories, which proved to be useful for discriminating hazelnuts according to origin. When the considered regions were sufficiently distinct in terms of climate and geography, this approach proved valuable for distinguishing hazelnuts originating from various sources. This was demonstrated by the promising results obtained in classifying regions located in the Mediterranean basin (ESP + ITA) from those in CHL and GEO. These regions could be accurately classified with a 93.7 $\pm$ 4.2% of the global correct classification rate, proving the method to be reliable and not set-dependent.

On the other hand, it is important to note that regions with similar pedoclimatic characteristics may present challenges in distinguishing them based on their isotopic ratios of light elements. This was evident in the regions studied in this research, situated in Spain and Italy, where the differences in analyzed isotopic ratios were not significant enough for differentiation.

These findings suggest that, provided the regions are sufficiently distinct in climate and geography, the further development of a reliable and universally applicable method for identifying the geographical origin of hazelnuts would be feasible and could effectively detect counterfeit products from different sources. Nevertheless, this shall require further evaluation of the model by including samples from additional production areas. Also, as the isotopic ratios depend on the climatic conditions, it would be advisable to regularly update the models by including future harvest years. This step is essential to implement the model as a reliable authentication tool to support inspections verifying the geographical origin of hazelnuts.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/foods13213399/s1. Table S1: Sample information and composition of the seven training (train) and validation (test) sets; Table S2: Mean $\pm$ standard deviation and statistical tests (normality test of Shapiro–Wilk, independent samples *t*-test, one-way analysis of variance (ANOVA), U Mann–Whitney and Kruskal–Wallis) with *p*-values to assess the interannual variability of the isotopic ratios for each geographical region; Table S3: Results of the leave 10% out cross-validation of the PLS-DA training models developed on $\delta^{18}O$ and the $\delta^{2}H$, $\delta^{13}C$ of the main FAMEs data to discriminate samples from three regions of origin. Results are the mean values ($\pm$ standard deviation) obtained from seven iterations; Table S4: Thresholds of the PLS-DA models (7 training sets) for each class optimized with ROC curves; Table S5: Results of the leave 10% out cross-validation of the PLS-DA training models developed on $\delta^{18}O$, and the $\delta^{2}H$, $\delta^{13}C$ of the main FAMEs data to discriminate samples from three regions of origin. Results are the mean values ($\pm$standard deviation) obtained from seven iterations; and Table S6: Misclassified and not assigned sample information from each test set.

Author Contributions: Conceptualization, M.R. (Mònica Rosell), M.R. (Mercè Rovira), A.R., S.V., and A.T.; methodology, B.T.-C., M.R. (Mònica Rosell), S.V. and A.T.; validation, B.T.-C.; formal analysis, B.T.-C.; investigation, B.T.-C. and M.R. (Mònica Rosell); resources, M.R. (Mercè Rovira), A.R. and S.V.; data curation, B.T.-C. and M.R. (Mònica Rosell); writing—original draft preparation, B.T.-C.; writing—review and editing, M.R. (Mònica Rosell), A.S., M.R. (Mercè Rovira), A.R., F.G., S.V., and

A.T.; supervision, M.R. (Mònica Rosell), A.S., F.G., S.V., and A.T.; project administration, S.V.; funding acquisition, S.V. All authors have read and agreed to the published version of the manuscript.

Funding: This work was developed in the context of the project TRACENUTS, PID2020-117701RB-I00 financed by MICIU/AEI/10.13039/501100011033. The MAiMA group is part of the MAGH project (2021-SGR-00308), financed by the AGAUR (Catalan Government). B. Torres-Cobos thanks the Spanish Ministry of Universities predoctoral fellowships FPU20/01454.

Data Availability Statement: The original data presented in the study are available at CORA. Repositori de Dades de Recerca: Torres Cobos, Berta; Rosell Linares, Mònica; Soler, Albert; Rovira, Mercè; Romero, Agustí; Guardiola, Francesc; Vichi, Stefania; Tres, Alba, 2024, "Stable isotope ratio values (bulk $\delta^{18}O$, and $\delta^{13}C$ and $\delta^{2}H$ of various fatty acids) of hazelnuts", https://doi.org/10.34810/data1724.

Acknowledgments: INSA-UB Maria de Maeztu Unit of Excellence (Grant CEX2021-001234-M) funded by MICIU/AEI/FEDER, UE. INSA-UB authors are part of the LiBiFOOD Consolidated Research Group (2021-SGR-00854) recognized by AGAUR (Catalan Government). The authors would like to express their gratitude to Ferrero Hazelnut Company and Tuscia University (Department of Agriculture and Forest Science) for providing the hazelnut samples from Chile, Georgia, and Italy, respectively. They also thank the technicians at MAiMA (Roger Puig, Pedro Valenzuela, and Tatiana Palomeque) and CCiT-UB (Raul Carrey, Anna Navarro, Pilar Rubio, and Pilar Teixidor) for their collaboration and technical support in this project.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Camin, F.; Boner, M.; Bontempo, L.; Fauhl-Hassek, C.; Kelly, S.D.; Riedl, J.; Rossmann, A. Stable isotope techniques for verifying the declared geographical origin of food in legal cases. *Trends Food Sci. Technol.* **2017**, *61*, 176–187. [CrossRef]
2. Hrastar, R.; Petrišič, M.G.; Ogrinc, N.; Košir, I.J. Fatty acid and stable carbon isotope characterization of camelina sativa oil: Implications for authentication. *J. Agric. Food Chem.* **2009**, *57*, 579–585. [CrossRef] [PubMed]
3. 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]
4. Laursen, K.H.; Bontempo, L.; Camin, F.; Roßmann, A. Advances in isotopic analysis for food authenticity testing. In *Advances in Food Authenticity Testing*; Downey, G., Ed.; Woodhead Publishing: Cambridge, UK, 2016; pp. 227–252. [CrossRef]
5. Korenaga, T.; Suzuki, Y.; Chikaraishi, Y. Biochemical stable isotope analysis in food authenticity. In *Engineering Tools in the Beverage Industry*; Grumezescu, A.M., Holban, A.M., Eds.; Woodhead Publishing: Cambridge, UK, 2019; pp. 209–227. [CrossRef]
6. 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]
7. Podio, N.S.; Baroni, M.V.; Badini, R.G.; Inga, M.; Ostera, H.A.; Cagnoni, M.; Gautier, E.A.; García, P.P.; Hoogewerff, J.; Wunderlin, D.A. Elemental and isotopic fingerprint of argentinean wheat. Matching soil, water, and crop composition to differentiate provenance. *J. Agric. Food Chem.* **2013**, *61*, 3763–3773. [CrossRef]
8. De Rijke, E.; Schoorl, J.C.; Cerli, C.; Vonhof, H.B.; Verdegaal, S.J.A.; Vivó-Truyols, G.; Lopatka, M.; Dekter, R.; Bakker, D.; Sjerps, M.J.; et al. The use of $\delta^{2}H$ and $\delta^{18}O$ isotopic analyses combined with chemometrics as a traceability tool for the geographical origin of bell peppers. *Food Chem.* **2016**, *204*, 122–128. [CrossRef]
9. Bertoldi, D.; Cossignani, L.; Blasi, F.; Perini, M.; Barbero, A.; Pianezze, S.; Montesano, D. Characterization and geographical traceability of Italian goji berries. *Food Chem.* **2019**, *275*, 585–593. [CrossRef]
10. Bontempo, L.; Paolini, M.; Franceschi, P.; Ziller, L.; García-González, D.L.; Camin, F. Characterisation and attempted differentiation of European and extra-European olive oils using stable isotope ratio analysis. *Food Chem.* **2019**, *276*, 782–789. [CrossRef]
11. 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]
12. Krauß, S.; Vieweg, A.; Vetter, W. Stable isotope signatures ($\delta^{2}H$-, $\delta^{13}C$-, $\delta^{15}N$-values) of walnuts (*Juglans regia* L.) from different regions in Germany. *J. Sci. Food Agric.* **2020**, *100*, 1625–1634. [CrossRef]
13. Smith, B.N.; Epstein, S. Two categories of $^{13}C/^{12}C$ ratios for higher plants. *Plant Physiol.* **1971**, *47*, 380–384. [CrossRef] [PubMed]
14. Yoneyama, T.; Fujiwara, H.; Engelaar, W.M.H.G. Weather and nodule mediated variations in $\delta^{13}C$ and $\delta^{15}N$ values in field-grown soybean (*Glycine max* L.) with special interest in the analyses of xylem fluids. *J. Exp. Bot.* **2000**, *51*, 559–566. [CrossRef]
15. Martin, G.J.; Martin, M.L. Climatic significance of isotope ratios. *Phytochem. Rev.* **2003**, *2*, 179–190. [CrossRef]
16. Scartazza, A.; Mata, C.; Matteucci, G.; Yakir, D.; Moscatello, S.; Brugnoli, E. Comparisons of delta^{13}C of photosynthetic products and ecosystem respiratory CO_2 and their responses to seasonal climate variability. *Oecologia* **2004**, *140*, 340–351. [CrossRef] [PubMed]

17. Camin, F.; Larcher, R.; Perini, M.; Bontempo, L.; Bertoldi, D.; Gagliano, G.; Nicolini, G.; Versini, G. Characterisation of authentic Italian extra-virgin olive oils by stable isotope ratios of C, O, and H and mineral composition. *Food Chem.* **2010**, *118*, 901–909. [CrossRef]
18. Schellenberg, A.; Chmielus, S.; Schlicht, C.; Camin, F.; Perini, M.; Bontempo, L.; Heinrich, K.; Kelly, S.D.; Rossmann, A.; Thomas, F.; et al. Multielement stable isotope ratios (H, C, N, S) of honey from different European regions. *Food Chem.* **2010**, *121*, 770–777. [CrossRef]
19. Faberi, A.; Marianella, R.M.; Fuselli, F.; La Mantia, A.; Ciardiello, F.; Montesano, C.; Mascini, M.; Sergi, M.; Compagnone, D. Fatty acid composition and $\delta^{13}C$ of bulk and individual fatty acids as marker for authenticating Italian PDO/PGI extra virgin olive oils by means of isotopic ratio mass spectrometry. *J. Mass Spectrom.* **2014**, *49*, 840–849. [CrossRef]
20. Yoshida, M.; Tabata, A.; Niino, T.; Chiku, K.; Nakashita, R.; Suzuki, Y. Potential application of light element stable isotope ratio in crude fiber for geographical origin verification of raw and cooked kabocha pumpkin (*Cucurbita maxima*). *Food Chem.* **2022**, *373*, 131462. [CrossRef]
21. Chung, I.M.; Kim, Y.J.; Moon, H.S.; Han, J.G.; Kong, W.S.; Yarnes, C.T.; Kim, S.H. Improved accuracy of geographical origin identification of shiitake grown in sawdust medium: A compound-specific isotope model-based pilot study. *Food Chem.* **2022**, *369*, 130955. [CrossRef]
22. Luykx, D.M.A.M.; van Ruth, S.M. An overview of analytical methods for determining the geographical origin of food products. *Food Chem.* **2008**, *107*, 897–911. [CrossRef]
23. Drivelos, S.A.; Georgiou, C.A. Multi-element and multi-isotope-ratio analysis to determine the geographical origin of foods in the European Union. *TRAC—Trend Anal Chem.* **2012**, *40*, 38–51. [CrossRef]
24. Muñoz-Redondo, J.M.; Bertoldi, D.; Tonon, A.; Ziller, L.; Camin, F.; Moreno-Rojas, J.M. Multi-element and stable isotopes characterization of commercial avocado fruit (*Persea americana* Mill) with origin authentication purposes. *Food Contr.* **2022**, *137*, 108975. [CrossRef]
25. Nie, J.; Yang, J.; Liu, C.; Li, C.; Shao, S.; Yao, C.; Chen, B.; Tao, Y.; Wang, F.; Zhang, Y.; et al. Stable isotope and elemental profiles determine geographical origin of saffron from China and Iran. *Food Chem.* **2023**, *405*, 134733. [CrossRef]
26. Król, K.; Gantner, M. Morphological traits and chemical composition of hazelnut from different geographical origins: A review. *Agriculture* **2020**, *10*, 375. [CrossRef]
27. FAOSTAT, Food and Agriculture Organization of the United Nations. Agricultural Data 2022. Available online: http://www.fao.org/faostat/en/ (accessed on 14 March 2024).
28. GIview Database. Search for Geographical Indications Across the European Union and Beyond, Data 2023. Available online: https://www.tmdn.org/giview/gi/search (accessed on 14 March 2024).
29. Anderson, K.A.; Smith, B.W. Effect of season and variety on the differentiation of geographic growing origin of pistachios by stable isotope profiling. *J. Agric. Food Chem.* **2006**, *54*, 1747–1752. [CrossRef]
30. Di Pierro, E.A.; Ziller, L.; Tonon, A.; Bianco, L.; Franceschi, P.; Troggio, M.; Camin, F. Characterization and valorisation of the Italian walnut (*Juglans regia* L.): A first application of stable isotope ratio analysis to determine walnut geographical origin. In Proceedings of the 2nd Isotope Ratio MS Day, Messina, Italy, 27–29 June 2018. Available online: http://hdl.handle.net/10449/49708 (accessed on 20 June 2024).
31. Popescu, R.; Ionete, R.E.; Botoran, O.R.; Costinel, D.; Bucura, F.; Geana, E.I.; Alabedallat, Y.F.J.; Botu, M. ^{1}H-NMR profiling and carbon isotope discrimination as tools for the comparative assessment of walnut (*Juglans regia* L.) cultivars with various geographical and genetic origins—A preliminary study. *Molecules* **2019**, *24*, 1378. [CrossRef]
32. Wadood, S.A.; Nie, J.; Li, C.; Rogers, K.M.; Zhang, Y.; Yuan, Y. Geographical origin classification of peanuts and processed fractions using stable isotopes. *Food Chem. X* **2022**, *16*, 100456. [CrossRef]
33. Mahalovich, M.F.; Kimsey, M.J.; Fortin-Noreus, J.K.; Robbins, C.T. Isotopic heterogeneity in whitebark pine (*Pinus albicaulis* Engelm.) nuts across geographic, edaphic and climatic gradients in the Northern Rockies (USA). *For. Ecol. Manag.* **2016**, *359*, 174–189. [CrossRef]
34. Sammarco, G.; Rossi, M.; Summan, M.; Cavanna, D.; Viotto, L.; Pettenà, P.; Dall'Asta, C.; Iacumin, P. Hazelnut products traceability through combined isotope ratio mass spectrometry and multi-elemental analysis. *JSFA Rep.* **2023**, *3*, 633–645. [CrossRef]
35. Torres-Cobos, B.; Rosell, M.; Soler, A.; Rovira, M.; Romero, A.; Guardiola, F.; Vichi, S.; Tres, A. Investigating isotopic markers for hazelnut geographical authentication: Promising variables and potential applications. *Food Chem.* **2024**, *449*, 139083. [CrossRef]
36. Quintanilla-Casas, B.; Marin, M.; Guardiola, F.; García-González, D.L.; Barbieri, S.; Bendini, A.; Gallina Toschi, T.; Vichi, S.; Tres, A. Supporting the sensory panel to grade virgin olive oils: An in-house-validated screening tool by volatile fingerprinting and Chemometrics. *Foods* **2020**, *9*, 1509. [CrossRef] [PubMed]
37. Bowen, G.J.; Revenaugh, J. Interpolating the isotopic composition of modern meteoric precipitation. *Water Resour. Res.* **2003**, *39*, 1299. [CrossRef]
38. IAEA/WMO. Global Network of Isotopes in Precipitation. The GNIP Database. 2015. Available online: https://nucleus.iaea.org/wiser (accessed on 22 January 2024).
39. Bowen, G.J. The Online Isotopes in Precipitation Calculator, Version 3.2. 2024. Available online: http://www.waterisotopes.org (accessed on 15 June 2024).

40. Ehtesham, E.; Hayman, A.R.; McComb, K.A.; Van Hale, R.; Frew, R.D. Correlation of geographical location with stable isotope values of hydrogen and carbon of fatty acids from New Zealand milk and bulk milk powder. *J. Agric. Food Chem.* **2013**, *61*, 8914–8923. [CrossRef]
41. Paolini, M.; Bontempo, L.; Camin, F. Compound-specific $\delta^{13}C$ and $\delta^{2}H$ analysis of olive oil fatty acids. *Talanta* **2017**, *174*, 38–43. [CrossRef]

5.3 Development of metabolic methods for hazelnut geographical and varietal authentication: Proof-of-concept studies

5.3.1 Sesquiterpene fingerprinting

Considering the genetic and environmental effects on the SH composition of plants [143, 168, 214], and the fact that several studies have identified these compounds in hazelnut kernels and oil [233-236], it was worth to consider them as potential markers for hazelnut geographical and varietal authentication. Consequently, SH were explored through HS-SPME-GC-MS in whole hazelnut kernels, ground hazelnuts, and hazelnut oil extracted by mechanical pressing or by solvent extraction. However, results revealed a minimal presence of SH in hazelnuts (data not shown) and ultimately this led to discard these compounds as possible authenticity markers.

5.3.2 Triacylglycerol fingerprinting

As previously mentioned (section 1.3.4), TAG analysis through HT-GC combined with MS detection and fingerprinting approaches, could serve as a suitable screening tool for verifying hazelnut cultivar and provenance.

To develop a rapid analytical method for TAG fingerprinting through HT-GC-MS, a preliminary study was carried out using OO as the sample. OO was chosen due to several factors: its lipid composition is similar to that of hazelnuts [44,150], it is easy to handle in liquid form, eliminating the need for extraction steps and established methods for TAG analysis in OO already exist, allowing for direct comparison with our results.

The method was developed and applied on a case study regarding the detection of VOO adulteration. The sample set consisted in 143 oils, including:

- 28 genuine VOO, encompassing a variety of geographical production areas, olive cultivars, and technological conditions.
- their blends (n= 115) with 36 adulterant oils of other botanical origins (sunflower, high-oleic sunflower, soybean and hazelnut). Among the adulterants there were high linoleic (HL) seed oils (refined sunflower and refined soybean

oils), and high oleic (HO) vegetable oils (high-oleic refined sunflower, and virgin and refined hazelnut oils). Their blends with genuine oils were prepared in-house, VOO samples were adulterated at 2% and 5% with HL oils (n = 58), and at 5% and 10% with HO oils (n= 57).

The sample set was analysed by four methods: the official method to detect OO adulteration based on ΔECN42 [163], and three untargeted methods: a high high-throughput instrumental method based on TAG untargeted profiling by flow injection analysis-heated electrospray-high resolution mass spectrometry (FIA-HESI-HRMS) [257], a variation of the official method based on the TAG fingerprinting by HPLC coupled to refractive index detector (HPLC-RID), and the TAG fingerprinting by HT-GC-MS.

The data obtained from official method were processed following the directives of the Regulation (EC) No. 2022/2105 [163]. The data obtained by the other three untargeted methods were used to build independent PLS-DA classification models (**Figure 11**). For each method, two types of PLS-DA models were developed:

1) HL adulteration model: A binary PLS-DA to discriminate between genuine and HL adulterated samples.

2) HO adulteration model: A binary PLS-DA to discriminate between genuine and HO adulterated samples.

Models were built with 80% of the samples and were externally validated by predicting the class of the samples from the validation set (20% of the whole sample set). Samples were considered as genuine (i.e. not-adulterated with HL oils >2% or with HO oils >5%) only when classified as such by both models. An exploration of the regression coefficient of the three untargeted methods was carried out to study the variables that contributed the most to the discrimination between genuine and adulterated samples. The performances of each method were compared in terms of accuracy of prediction, sensitivity and specificity. The outcomes of this preliminary study are available in Publication 6.

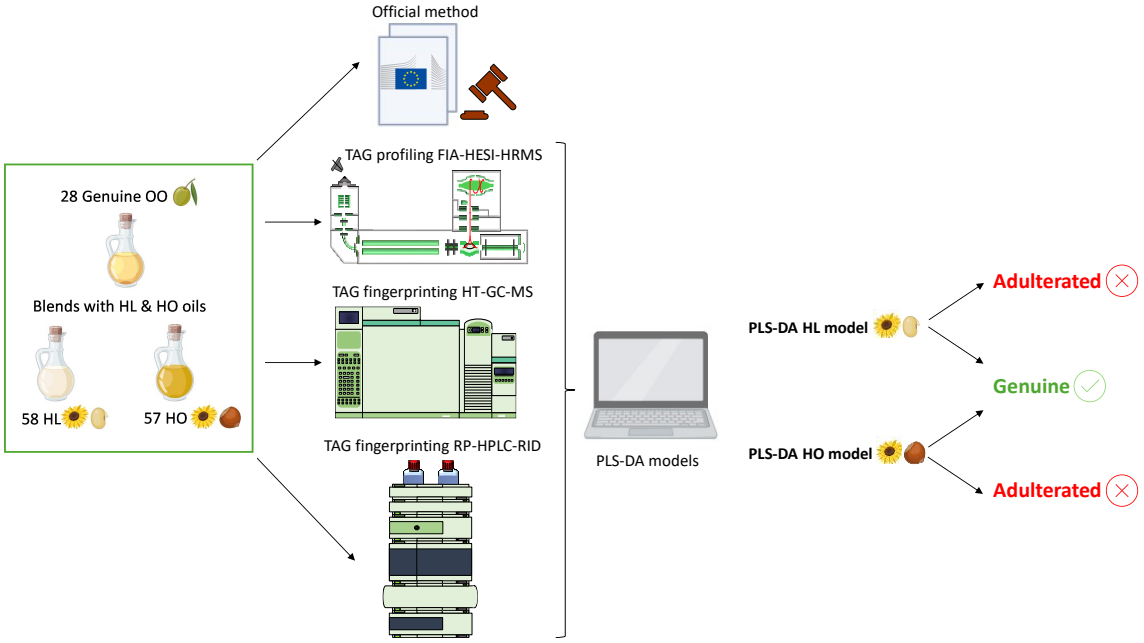


Figure 11. Graphical abstract of Publication 6. OO: olive oil, HL: high linoleic, HO: high oleic, TAG: triacylglycerol, FIA-HESI-HRMS: flow injection analysis-heated electrospray-high resolution mass spectrometry, HT-GC-MS: high-temperature-gas chromatography-mass spectrometry, RP-HPLC-RID: reverse phase-high performance liquid chromatography-refractive index detector, PLS-DA: partial least square-discriminant analysis.

5.3.3 Publication 6

Revealing adulterated olive oils by triacylglycerol screening methods: Beyond the official method

Berta Torres-Cobos, Beatriz Quintanilla-Casas, Giulia Vicario, Francesc Guardiola, Alba Tres, Stefania Vichi

*Food Chemistry, **2023**, 409, 135256, https://doi.org/10.1016/j.foodchem.2022.135256*

Supplementary material available in **Annex 5**

Food Chemistry 409 (2023) 135256

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem

Revealing adulterated olive oils by triacylglycerol screening methods: Beyond the official method

Berta Torres-Cobos [a,b], Beatriz Quintanilla-Casas [a,b,*], Giulia Vicario [a,1], Francesc Guardiola [a,b], Alba Tres [a,b], Stefania Vichi [a,b]

[a] Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Campus De l'Alimentació Torribera, Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona, Santa Coloma de Gramenet, Spain
[b] Institut de Recerca en Nutrició i Seguretat Alimentària (INSA-UB), Universitat de Barcelona, Av Prat de la Riba, 171, 08921 Santa Coloma de Gramenet, Spain

ARTICLE INFO

Keywords:
Olive oil
Adulteration
High Resolution Mass Spectrometry
High Temperature Gas Chromatography
Triacylglycerols
Fingerprinting
ECN42
Screening

ABSTRACT

Official control methods to detect olive oil (OO) adulteration fail to provide satisfactory consumer protection. Thus, faster and more sensitive screening tools are needed to increase their effectiveness. Here, the official method for adulterant detection in OO was compared with three untargeted screening methods based on triacylglycerol analysis using high-throughput (FIA-HESI-HRMS; HT-GC–MS; HPLC-RID) and pattern recognition techniques (PLS-DA). They were assayed on a set of genuine and adulterated samples with a high natural variability (n = 143). The sensitivity of the official method was 1 for high linoleic (HL) blends at ≥2 % but only 0.39 for high oleic (HO) blends at ≥5 %, while specificity was 0.96. The sensitivity of the screening methods in external validation was 0.90–0.99 for the detection of HL and 0.82–0.88 for HO blends. Among them, HT-GC–MS offered the highest sensitivity (0.94) and specificity (0.76), proving to be the most suitable screening tool for OO authentication.

1. Introduction

According to the latest report from the EU Food Fraud Network (European Union, 2021), olive oil (OO) tops the list of reported adulterated food products. One of the most common frauds is mixing OO with cheaper vegetable oils (Casadei et al., 2021; The Food Integrity Project, 2018). The official methods to assess OO purity and detect the presence of extraneous vegetable oils include the analysis of fatty acids (FAs), triacylglycerols (TAGs) and sterols (Regulation (EU) No 2568/91 and its amendments). Adulterants may be masked by the removal of minor compounds, as occurs with desterolised seed oils, but major constituents such as FAs and/or the corresponding TAGs appear to be more robust parameters. As not only the amounts of FAs but also their arrangement to form TAGs are genetically determined, the official method based on TAGs rather than FA composition has proved to be more effective in detecting low levels of vegetable oils in OO (Christopoulou et al., 2004). Nonetheless, the current official method for TAG analysis is time- and reagent-consuming because it entails a double analysis of the samples. Firstly, High Performance Liquid Chromatography coupled to a Refractive Index Detector (HPLC-RID) is used to determine the experimental amount of TAGs and calculate the equivalent carbon number 42 (ECN42). Then, the theoretical ECN42 is determined from the FA content analysed by Gas Chromatography coupled to a Flame Ionization Detector (GC-FID), and both values are compared to calculate the ΔECN42 parameter. The lengthiness of this procedure limits the annual conformity checks to one per thousand tons of the OO marketed in European Union member states (Regulation (EU) No 2568/91 and its amendments), which does not guarantee a satisfactory level of consumer protection (European Commission, 2020). Additionally, this method shows low sensitivity when high linoleic (HL) seed oils are present at low levels or when the adulterants have a similar TAG composition to OO, as is the case of high oleic (HO) vegetable oils (Conte et al., 2020; Mailer & Gafner, 2020). Considering these drawbacks, and that the illegal blending of OO with both HL and HO vegetable oils is reported to be a common practice (Casadei et al., 2021; Everstine et al., 2013; Mailer & Gafner, 2020; Yan et al., 2020), more efficient methods

* Corresponding author at: Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Campus De l'Alimentació Torribera, Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona, Av Prat de la Riba, 171, 08921 Santa Coloma de Gramenet, Spain.
E-mail address: beatrizquintanilla@ub.edu (B. Quintanilla-Casas).
[1] Present address: Crop Science Research Centre, Scuola Superiore Sant'Anna, Piazza Martiri della Libertà, 56127, Pisa, Italy.

https://doi.org/10.1016/j.foodchem.2022.135256
Received 27 July 2022; Received in revised form 1 December 2022; Accepted 17 December 2022
Available online 20 December 2022

are needed to detect extraneous vegetable oils in OO.

In this context, the development of fit-for-purpose screening tools would allow a more rapid assessment of a higher number of samples for a more efficient purity control of OO (Aparicio et al., 2013; Ruiz-Samblás et al., 2015). Screening methods featuring a binary qualitative output (compliant/non-compliant) obtained by time-efficient analytical methods, and combining different analytical and classification techniques, have been developed to overcome limitations in food authentication (López et al., 2014). Their advantages include immediacy in decision-making and reducing the time and cost of routine analysis (Muñoz-Olivas, 2004). The main goal of these screening methods is to achieve high sensitivity values for an efficient food fraud alert. Thus, qualitative screening can identify all suspicious samples, and subsequent control checks can confirm whether they have been adulterated or not (Magnusson & Örnemark, 2014). This is of particular importance when the fraudulent products pose a threat to food safety (López et al., 2014), in which case adulteration not only implies a loss of traceability due to blending with substances from unknown sources, but also carries the risk of allergenic or toxic effects related to the nature of the adulterants (Arlorio et al., 2010; Gelpí et al., 2002).

Most of the screening methods proposed as alternatives to the official method for adulterant detection in OO are also based on the measurement of the TAG fraction, due to its advantages as an authenticity marker (Meenu et al., 2019). Approaches based on high-throughput analytical techniques coupled to multivariate classification methods have proved to be efficient in fraud detection and have a short analysis time (Cavanna et al., 2018; Esslinger et al., 2014). Among them, untargeted profiling and fingerprinting methods have been applied as an alternative to the classic targeted approach of the current official method (Ballin & Laursen, 2019). Besides involving an effortless data treatment, untargeted approaches can consider more analytical information than conventional targeted methods, which may be valuable for authentication (Quintanilla-Casas et al., 2020). Although some of the alternative methods have provided promising results, comparison with the official method is difficult, as they differ in sample sets, variability, and analytical conditions. A further challenge to assessing the performance of the new methods is that multivariate validation is not as well defined as univariate validation (López et al., 2014).

In the present study, three promising screening methods for adulterant detection in OO based on TAG analysis and the official method based on ΔECN42 (Regulation (EU) No 2568/91 and its amendments) were compared. The selected screening methods are all based on an untargeted approach. One involves TAG profiling by Flow Injection Analysis-Heated Electrospray-High Resolution Mass Spectrometry (FIA-HESI-HRMS), an extremely fast high-throughput instrumental method requiring minimal sample preparation, which had already been tested for the analysis of OO adulteration with satisfactory results (Quintanilla-Casas et al., 2021). The other two methods were developed in the present study and follow a fingerprinting approach, a cutting-edge strategy that uses the whole analytical signal to find specific patterns characteristic of a given food category. Thus, TAG fingerprinting by High-Temperature Gas Chromatography coupled to Mass Spectrometry (HT-GC–MS) was developed by adapting the chromatographic conditions of (Ruiz-Samblás et al., 2012) to MS detection, which allowed us to build an unfolded matrix from Selected Ion Monitoring (SIM) chromatograms (Torres-Cobos et al., 2021). The third method, TAG fingerprinting by High Performance Liquid Chromatography coupled to Refractive Index Detector (HPLC-RID), was developed by applying chemometrics to the raw HPLC-RID data obtained by the official method, but without the need for any further TAG identification or quantitation. Finally, unlike in previous studies, the performance of each method could be directly compared because the same sample set (genuine samples and their blends with HO and HL adulterants) was analysed in each case.

2. Material and methods

2.1. Samples and experimental design

A set of 150 samples was analysed by the four methods described above. The set was composed of 30 traceable genuine extra virgin olive oils (EVOO) and their blends (n = 120) with 36 adulterant oils of different botanical origin, including HL oils (refined sunflower oil [SFO, n = 14] and refined soybean oil [SBO, n = 10]), as well as HO oils (virgin and refined hazelnut oils [HZN, n = 10] and high-oleic refined sunflower oils [HOSFO, n = 12]). The blends were in-house prepared at 2 % and 5 % for the HL (n = 60) and at 5 % and 10 % for the HO (n = 60) adulterants, following a balanced incomplete Latin squares experimental design. Seven of the chromatographic or FIA runs had to be discarded due to analytical problems, and these samples were eliminated from the datasets of the four methods to avoid any bias between the different models tested. The final dataset (n = 143) is provided in Table S1 of Supplementary Information.

All samples were obtained directly from reliable producers in the framework of the Autenfood project (ACCIÓ-Programa Operatiu FEDER Catalunya 2014–2020). The EVOO samples represented a range of geographical production regions, olive cultivars and technological conditions. The natural variability of the samples was further ensured by including several adulterant oils of each type in the sampling.

2.2. Method I. Official method to determine ΔECN42 according to Regulation (EU) No 2568/91

The difference between the experimental and theoretical ECN42 values of TAGs was calculated and compared with the limit established for EVOO, i.e., $\Delta ECN42 < |0.20|$.

2.2.1. Material and reagents

Methanol (>99.5 %), HPLC grade hexane (99 %), HPLC grade heptane and diethyl ether stabilized with 7 ppm of BHT were purchased from Scharlau (Sentmenat, Spain). Potassium hydroxide was purchased from Panreac (Castellar del Vallès, Spain), and propionitrile (>99 %) from Sigma-Aldrich (St. Louis, USA). Methanolic potassium hydroxide solution (2 M) was prepared according to the official method. Extra Bond® solid phase extraction cartridges (6 mL), packed with 1 g of silica phase, and 13 mm-0.45 μm nylon filters, both from Scharlau (Sentmenat, Spain), were used for sample purification and preparation.

2.2.2. TAG analysis by reverse phase (RP)-HPLC-RID and assessment of experimental ECN42

Once the samples were prepared and purified according to Annex XVIII of Regulation (EU) No 2568/91 and its amendments, TAGs were determined by HPLC-RID using an Agilent Technologies 1100 instrument (Agilent Technologies, Santa Clara, California, USA). Analytes were separated on a Luna C18 column (250 × 4.6 mm, I.D., 5 μm) from Phenomenex (Torrance, California, USA) at 40℃. The analysis was performed in isocratic mode with propionitrile as the mobile phase at 1 mL/min. The injection volume was 20 μL. The peak areas of TAGs with an ECN between 42 and 52 were integrated and normalized, obtaining the experimental TAG content with ECN42.

2.2.3. Analysis of FA methyl esters (FAME) by GC-FID and calculation of theoretical ECN42

After obtaining FAME by transmethylation according to Annex X of Regulation (EU) No 2568/91 and its amendments, they were analysed by GC-FID using an Agilent 4890D chromatograph, coupled to an Agilent Technologies 7683B automatic sampler (Agilent Technologies, Santa Clara, California, USA). The injection volume was 1 μL, with a split injection ratio of 1:100. The temperature of the injector and detector was 250 ℃. Analytes were separated on a SP-2380 capillary column (60 m × 0.25 mm I.D., 0.2 μm) (Supelco, St. Louis, USA). The initial column

temperature was 165 ℃ and it was held for 8 min; it was then increased to 210 C at 2 ℃ /min and held for 15 min. Hydrogen (99.995 %) was the carrier gas, and the flow rate was 1 mL/min. The areas of the FAME with 16 and 18 carbon atoms were integrated and normalized. The theoretical content of ECN42 TAGs was calculated according to Annex XVIII of Regulation (EU) No 2568/91 and its amendments.

2.3. Method II. Screening method based on TAG profiling by FIA-HESI-HRMS

2.3.1. Material and reagents

Dichloromethane (SupraSolv® for GC-ECD/FID), methanol (SupraSolv® for gas chromatography) and NaCl (ACS reagent, ≥99.0 %) were purchased from Merck (Darmstadt, Germany). Nitrogen (Alphagaz, purity 99.999 %, Air Liquid) was used in the Orbitrap-Exactive as nebulization gas.

2.3.2. Sample preparation

As described by Vichi et al. (2012), 30 mg of oil sample was dissolved in dichloromethane:methanol (70:30, v/v) and diluted 1:100 using the same solvent mixture. The solution was saturated with NaCl as the cationization agent and vortex-mixed for 30 s. The supernatant was then further diluted to 1:50 with the same solvent mixture and analysed.

2.3.3. TAG profile by FIA-HESI-HRMS

Untargeted profiling of TAGs was done according to Vichi et al. (2012), using an Orbitrap Exactive instrument (Thermo Fisher Scientific, Bremen, Germany) equipped with an electrospray source (H-ESI II) and coupled to a Surveyor MS pump with an Accela Open automatic sampler (Thermo Fisher Scientific, San Jose, USA).

The selected mass peaks were single positive-charged molecular ions forming adducts with sodium. These peaks were exported to peak lists and feasible elemental formulae attributable to TAGs were generated. Different restrictive criteria were set to generate reliable elemental formulae: C ≤ 200, H ≤ 400, O ≤ 15, Na = 1 and RDB: 2.5–19.5 (Quintanilla-Casas et al., 2021). Mass error tolerance was set at 5 ppm. Signal intensities of TAGs were expressed as a relative intensity, the total TAG profile being equal to 100 %. The molecular formulae calculation was performed with Xcalibur 2.1 (Thermo Fisher Scientific, Bremen, Germany), and the posterior data analysis was done using excel files, as shown in Fig. 1a.

2.4. Method III. Screening method based on TAG fingerprinting by HT-GC–MS

2.4.1. Material and reagents

Dichloromethane (SupraSolv® for GC-ECD/FID) was purchased from Merck (Darmstadt, Germany).

2.4.2. Sample preparation

A small amount (60 mg) of oil sample was dissolved in 3 mL of dichloromethane to a final concentration of 2 % (w/v).

2.4.3. TAG fingerprinting by selected ion Monitoring (SIM) HT-GC -MS

HT-GC–MS analysis was carried out on an Agilent Technologies 6890 N gas chromatograph coupled to an Agilent 5973 Network quadrupolar mass selective analyser (Agilent Technologies, Santa Clara, California, USA). A total of 2 μL of sample was injected with a split ratio of 1:20. Analytes were separated on a Rtx®-65TG column (Restek, Bellefonte, PA, USA) (30 × 0.25 mm I.D., 0.10 μm). The initial column temperature was 315 ℃, which was increased to 350 ℃ at 2 ℃ /min and held for 15 min. Helium was the carrier gas and the flow rate was 1.5 mL/min. The temperature of the injector was 370C and the transfer line, 350C. Electron impact mass spectra were recorded at 70 eV ionization energy. Mass spectra were acquired in SIM mode.

A fingerprinting approach was followed using extracted ion chromatograms (EIC) obtained for *m/z* signals attributable to 11 TAG fragment ions, according to Barber, Merren & Kelly (1964): acyl ions corresponding to a FFAA (Po: palmitoleic acid, C16:1; P: palmitic acid, C:16:0; O: oleic acid, C18:1; L: linoleic acid, C18:2; Ln: linolenic acid, C18:3) molecule (RCOOH) with a loss of an OH group $[RCOOH - OH]^+$ (*m/z* 237, $[Po - OH]^+$; *m/z* 239, $[P - OH]^+$; *m/z* 260, $[Ln - OH]^+$; *m/z* 262, $[L - OH]^+$; *m/z* 264, 265, $[O - OH]^+$) and to a FFAA acyl ions attached to a residual of the glycerol skeleton $[RCOOH - C_3H_5O]^+$ (*m/z* 311, $[Po - C_3H_5O]^+$; *m/z* 313, $[P - C_3H_5O]^+$; *m/z* 335, $[Ln - C_3H_5O]^+$; *m/z* 337, $[L - C_3H_5O]^+$; *m/z* 339, $[O - C_3H_5O]^+$). A data matrix was built for each ion with the 3159 scan intensities of each EIC (columns) for all samples (rows) (143 samples × 3159 variables) and it was aligned by the COW algorithm in Matlab® (Nielsen et al., 1998) to solve the retention time shifts among samples. The optimal COW parameters for each case

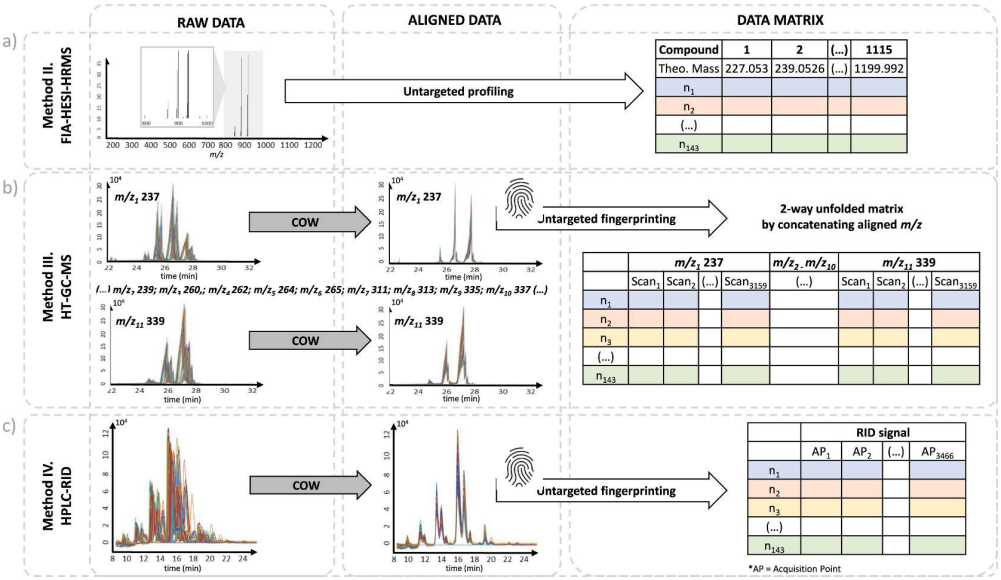


Fig. 1. Scheme of the data matrices building for each screening method: a) Method II: FIA-HESI-HRMS, b) Method III: HT-GC–MS fingerprinting and c) Method IV: HPLC-RID fingerprinting.

were calculated upon the starting parameters provided in Table S2, Supplementary information.

Then, the 11 aligned matrices – one for each EIC - were concatenated to obtain a two-way unfolded matrix (143 samples × 34749 variables) (Fig. 1b).

2.5. Method IV. Screening method based on TAG fingerprinting by RP-HPLC-RID

2.5.1. Sample preparation

Samples were prepared as described for Method I – Official method (section 2.2.2).

2.5.2. TAG fingerprinting by RP-HPLC-RID

Separation and detection of compounds were as described in section 2.2.3 for the official method (Method I). Using a fingerprinting approach, a data matrix was built with the chromatographic intensities obtained for each acquisition point from minute 8 until minute 25 for each sample (143 samples × 3466 variables). Chromatograms were aligned by the COW algorithm in Matlab® to solve the retention time shifts among samples (Nielsen et al., 1998) (Fig. 1c). The optimal COW parameters for each case were calculated upon the starting parameters provided in Table S2, Supplementary information.

2.6. Data processing and chemometrics

Data obtained by the official method (method I) were processed according to Regulation (EU) No 2568/91 and its amendments. Regarding the data matrices provided by each of the tested alternative screening methods (II-IV), classification models were developed and validated with SIMCA v13.0 (Umetrics AB, Sweden).

First, a Principal Component Analysis (PCA) was performed to explore the data and identify potential outliers. In no case were outlier samples detected by PCA, according to the Hotelling's T^2 range and distance to the model parameters.

Then, the whole sample set (n = 143) (Table S1, Supplementary information) was randomly split 7 times (7 iterations) into: i) a training set (80 % of the samples, n = 114; genuine EVOOs, n = 22; HL blends, n = 46; HO blends, n = 46) and ii) a validation set (20 % of the samples, n = 29; genuine EVOOs, n = 6; HL blends, n = 12; HO blends, n = 11). For each screening method and each iteration, two independent binary Partial Least Squares Discriminant Analysis (PLS-DA) models were built: one to discriminate between genuine and HL adulterated samples (n = 68; HL blends, n = 46; genuine EVOOs, n = 22) and the other to discriminate between genuine and HO adulterated samples (n = 68; HO blends, n = 46; genuine EVOOs, n = 22). In PLS-DA binary models, classes are expressed as PLS dummy variables (here 0 for the non-adulterated and 1 for the adulterated class). Then, the PLS predicted value of each sample was used for its classification into one class or the other according to a classification threshold (predicted value = 0.5). PLS-DA models were calibrated by leave 10 %-out cross-validation, selecting the optimal number of Latent Variables (LV) according to the lowest Root Mean Squared Error of Cross Validation (RMSEcv). A permutation test and ANOVA were carried out on the cross-validated predictive residuals (p-value) to assess overfitting. The Q^2 values and efficiency, expressed as the % of correct classification, were assessed to evaluate the suitability of each PLS-DA model.

After testing multiple pre-processing treatments, the optimal one for the HL and HO models based on FIA-HESI-HRMS profiling data was found to be logarithm 10, mean centring and scaling to unit of variance; for the HL model based on HT-GC–MS fingerprinting, it was mean centring and scaling to unit of variance, and for the HO model based on HT-GC–MS fingerprinting, logarithm 10. Finally, the pre-processing applied for the HL model based on RP-HPLC-RID fingerprinting was logarithm 10, mean centring and scaling to unit of variance and for the HO model, a first derivative was also needed. The optimal pre-processing treatments for the HL and HO models were applied to training sets of each iteration.

2.6.1. External validation of the authentication strategy

Samples in the abovementioned validation sets (20 %, n = 29 for each of the 7 iterations) were not included in the previous calibrated models but used to externally validate the authentication tool following a combined strategy according to Quintanilla-Casas et al. (2021) (Fig. 2). Thus, validation samples were classified as non-adulterated (with HL oils at ≥ 2 % or HO oils at ≥ 5 %) only when identified as such by both models, whereas they were considered as adulterated when identified as such by at least one of the authentication models (HL or HO).

External validation results for each screening method were compared according to the % of correct classification of each class, and the sensitivity (true positives/[true positives + false negatives]) and specificity (true negatives/[true negatives + false positives]) values, positive samples being those containing the adulterant.

2.6.2. Exploration of regression coefficients

For the three screening methods, the regression coefficients of the binary HL and HO PLS-DA models developed with the full sample set were explored to tentatively identify the variables that most contributed to the discrimination between adulterated and genuine samples. Regression coefficients were considered as significant when a jack-knife standard error of cross-validation (SEcv) was lower than the given coefficient value.

3. Results

3.1. Method I. Official method to determine ΔECN42 according to Regulation (EU) No 2568/91

The results obtained with the official method based on ΔECN42 showed high percentages of correct classification for genuine (96.4 %) and HL adulterated samples (100 %), whereas only 38.6 % of the blends with HO oils were classified as adulterated (Table 1).

3.2. Screening methods (II-IV) based on untargeted TAG profiling and fingerprinting

For all the screening methods, the cross-validation results of models built upon training sets (7 iterations) were successful. Mean overall sensitivity and specificity were 0.96–1 and 0.95–1, respectively. Each of the PLS-DA models (7 per screening method) were then used to predict the class of the samples in the corresponding validation set. The process was run seven times to evaluate the effect of the sample set composition and to increase the robustness of the external validation. The results achieved by the external validation of each screening method (sensitivity and specificity) (Table S3, Supplementary information) were expressed as mean values of the seven iterations (Table 2) and showed a sensitivity above 0.82 for all the developed methods, the HT-GC–MS method standing out with an average sensitivity for HL and HO adulterated samples of 0.94. Remarkably, the detection performance for HO blends was not far below the results obtained for the HL adulterants (Table 2). The three tested methods had a sensitivity close to 0.9 (mean value of the 7 iterations) and a specificity between 0.50 and 0.76 (mean values of the 7 iterations).

As explained in section 2.6.2, the coefficients of models were studied to tentatively identify the variables that contributed most to the discrimination between HL adulterated and genuine samples, or between HO adulterated and genuine samples. For FIA-HESI-HRMS, the relevant coefficients agreed with those reported by Quintanilla-Casas et al. (2021). In both models, the highest regression coefficients corresponded to several minor TAG species distributed within the entire experimental *m/z* range. Specifically, for the HL model, the most

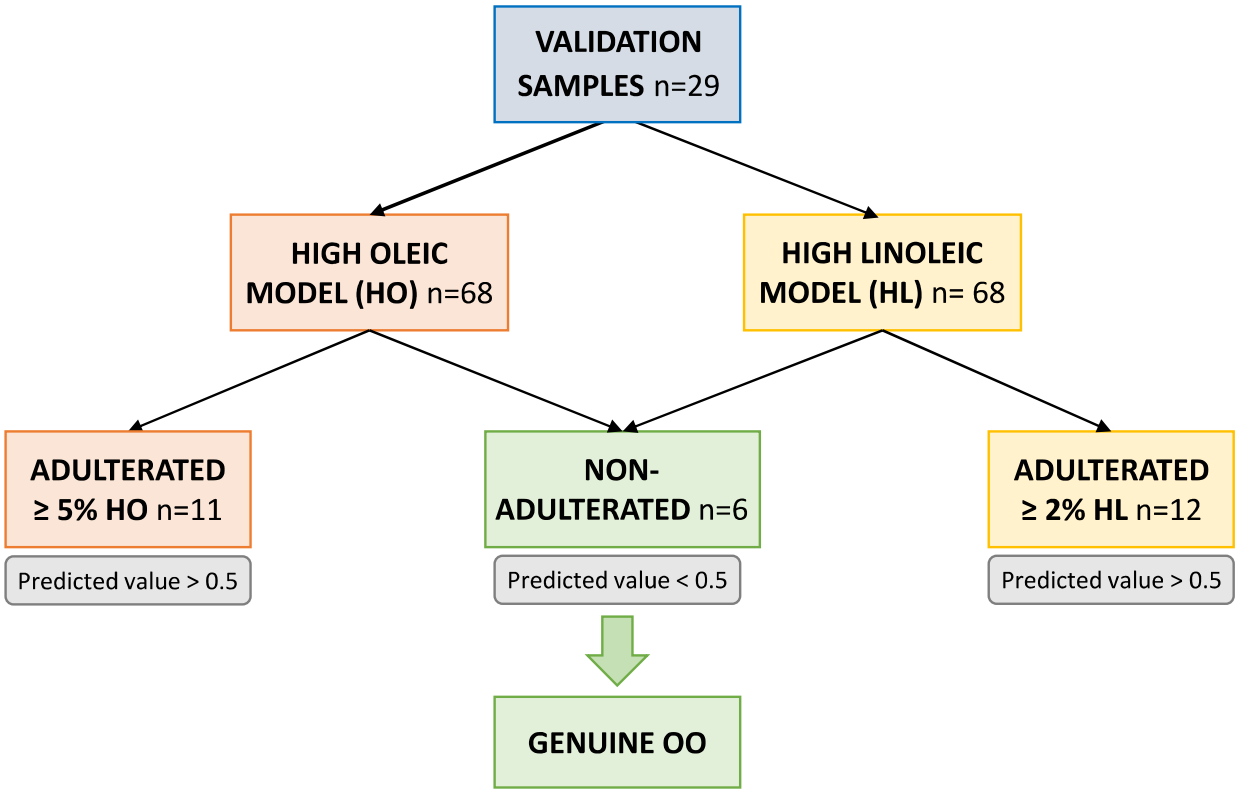


Fig. 2. Classification diagram based on two combined binary PLS-DA models to predict the presence of high linoleic (HL) (≥2%) and high oleic (HO) (≥5%) oil adulterants in olive oil (OO) using the tested untargeted screening methods.

Table 1
Classification results according to the official method for OO adulteration detection.

	n	Adulterated (n)	Genuine (n)	Correct classification (%)	Sensitivity	Specificity
Adulterated	115	80	35	**69.6**	**0.70**	
HL blends	58	58	0	**100**	**1**	
HO blends	57	22	35	**38.6**	**0.39**	
Genuine	28	1	27	**96.4**		**0.96**
Total	143					

relevant coefficients for the HL adulterated class belonged to TAGs containing L and Ln acids, such as C54 TAGs LLL and LLLn, and C52 TAGs such as PoLL/PLLn. On the other hand, for the HO model, the most discriminant variables associated with the HO adulterated class were signals corresponding to long-chain TAGs, C54 and C58, which could contain arachidic (A), behenic (B) or lignoceric acids (Li). Additionally, some TAGs containing L and short-chain TAGs C23, C24 and C26 also played a significant role in detecting HO adulterated samples (Quintanilla-Casas et al., 2021).

The study of the coefficients of models based on the HPLC-RID method (Fig. 3) produced results consistent with the FIA-HESI-HRMS findings, as the most relevant coefficients also belonged to minor TAGs. Specifically, to detect the HL adulterated samples, the highest coefficients were found in the first part of the fingerprint, covering the region of TAG clusters with ECN40, ECN42 and ECN44. According to the elution order within each cluster (Regulation (EU) No 2568/91 and its amendments), the most discriminant TAGs could corresponded to species containing L and Ln acids, such as LLLn (ECN40), LLL, OLLn, PoLL, PLLn (ECN42), PLL, PoPoO, and POLn (ECN44). Although ECN42 TAGs (tentatively identified as LLL, OLLn, and PoLL) and ECN44 TAGs (tentatively identified as OOLn, PoOL, PoPoO, and PLL) were also relevant for the detection of HO adulterated samples, the highest coefficients of the HO model also include TAGs in the ECN46 region (OOL, PPLn, PPoO, PoSL, SOLn, PoPoS), or even higher (Fig. 3), which agree with the relevant coefficients found in the FIA-HESI-HRMS model corresponding to long chain TAGs.

Finally, for the HT-GC–MS method (Fig. 3), some of the most relevant ions for the detection of samples adulterated with HL oils were found in the C54 TAG cluster and corresponded to glyceryl fragments with acyl species such as Ln, L and O. According to the fragmentation (Barber et al., 1964) and the expected elution order (Ruiz-Samblás et al., 2010), these relevant ions would agree with TAGs such as OLLn (*m/z* 335), LLLn (*m/z* 260, *m/z* 262, *m/z* 335 and *m/z* 337) and OLnLn (*m/z* 260, *m/z* 335 and *m/z* 339). In addition, some relevant coefficients for the *m/z* 262 and *m/z* 337 were found in chromatographic regions where the eluted TAGs were of a higher carbon number, such as C56 TAG, and could contain L and A. Several high coefficients corresponded to other glyceryl species of lower molecular weight that eluted prior to the main TAG clusters, which could be C38 diglycerides (DAG) such as OA (*m/z* 264) and LA (*m/z* 262) and C40 diglycerides such as OB (*m/z* 264, *m/z* 265 and *m/z* 339), and LB (*m/z* 262 and *m/z* 337).

For the HO model, some of the highest regression coefficients to

Table 2
External validation of PLS-DA models (HL blends vs genuine olive oils; HO blends vs genuine olive oils) developed by FIA-HESI-HRMS, HT-GC–MS and RP-HPLC-RID screening methods. Results are mean values obtained from seven randomly selected validation sets. Results of individual sets are reported in Supplementary information (Table S3).

FIA-ESI-HRMS[a]						
	n	Adulterated (n)	Genuine (n)	Correct classification (%)	Sensitivity	Specificity
Adulterated	23	20.6 ± 1.7	2.4 ± 1.7	**89.4 ± 7.5**	0.89 ± 0.07	
HL blends	12	10.9 ± 0.7	1.1 ± 0.7	90.5 ± 5.8	0.90 ± 0.06	
HO blends	11	9.7 ± 1.4	1.3 ± 1.4	88.3 ± 12.6	0.88 ± 0.13	
Genuine	6	3.0 ± 0.8	3.0 ± 0.8	**50.0 ± 13.6**		0.50 ± 0.14
Total	29					
HT-GC-MS[b]						
	n	Adulterated (n)	Genuine (n)	Correct classification (%)	Sensitivity	Specificity
Adulterated	23	21.6 ± 0.5	1.4 ± 0.5	**93.8 ± 2.3**	0.94 ± 0.02	
HL blends	12	11.9 ± 0.4	0.1 ± 0.4	98.8 ± 3.2	0.99 ± 0.03	
HO blends	11	9.7 ± 0.5	1.3 ± 0.5	88.3 ± 4.4	0.88 ± 0.04	
Genuine	6	1.4 ± 0.8	4.6 ± 0.8	**76.2 ± 13.1**		0.76 ± 0.10
Total	29					
HPLC-RID[c]						
	n	Adulterated (n)	Genuine (n)	Correct classification (%)	Sensitivity	Specificity
Adulterated	23	20.6 ± 0.8	2.4 ± 0.8	**89.4 ± 3.4**	0.89 ± 0.03	
HL blends	12	11.6 ± 0.5	0.4 ± 0.5	96.4 ± 4.5	0.96 ± 0.04	
HO blends	11	9.0 ± 0.8	2.0 ± 0.8	81.8 ± 7.4	0.82 ± 0.07	
Genuine	6	2.4 ± 0.5	3.6 ± 0.5	**59.5 ± 8.9**		0.60 ± 0.09
Total	29					

For all models HL and HO models: n = 143 and ANOVA p-value < 0.05.
[a] HL model: mean LVs = 6, $Q^2 > 0.4104$, RMSEcv < 0.3626; HO model: mean LVs = 5, $Q^2 > 0.3594$, RMSEcv < 0.3972.
[b] HL model: mean LVs = 8, $Q^2 > 0.7404$, RMSEcv < 0.2898; HO model: mean LVs = 8, $Q^2 > 0.5679$, RMSEcv < 0.3292.
[c] HL model: mean LVs = 5, $Q^2 > 0.4551$, RMSEcv < 0.3473; HO model: mean LVs = 5, $Q^2 > 0.2674$, RMSEcv < 0.4069.

detect HO adulterated samples were also found in the C54 TAG cluster and included acyl species such as L, Ln, O and Po (Fig. 3). In agreement with the fragmentation and the expected elution order (Ruiz-Samblás et al., 2010), these glyceryl species would correspond to PoOA (*m/z* 237), OLL (*m/z* 262 and *m/z* 339), OOLn (*m/z* 335), LLL (*m/z* 262), OLLn (*m/z* 335), LLLn (*m/z* 335 and *m/z* 337), and OLnLn (*m/z* 335). Moreover, some other relevant coefficients were found in the C48 and C50 TAG clusters that included P and Po as acyl species, and that would agree with PPoPo (*m/z* 237) and PPoS (*m/z* 237) TAGs, respectively. In addition, some fragments that could correspond to DAG also appeared as relevant coefficients: PoL (*m/z* 311), SO (*m/z* 339), SLn and LLn (*m/z* 335), LA (*m/z* 262), LB (*m/z* 262) and OB (*m/z* 264 and *m/z* 265).

These relevant coefficients agree with those from the HPLC-RID and FIA-HESI-HRMS models, as all of them found C54 TAGs conformed by Ln and L (OLLn, LLLn), which are relevant for detecting HL adulterated samples. For the HO model, both HT-GC–MS and FIA-HESI-HRMS detected relevant variables corresponding to C54 TAGs conformed by longer chain FAs such as A, and for the HPLC-RID coefficients, C54 TAGs containing O and Ln (OOLn) were also significant to detect HO adulterated samples in the HT-GC–MS method.

4. Discussion

The high rate of correct classification achieved by ΔECN42 for genuine OO samples demonstrated the high specificity of the current reference method (Table 1), which guarantees that genuine samples will not be misclassified as adulterated. The specificity achieved by the official method in the present study was higher than that reported by Beccaria et al. (2016), who obtained between 16 and 19 % of false positives when analysing genuine EVOOs from very different geographical locations. This may denote a high dependence of the ΔECN42 parameter on OO characteristics or factors not included in the official analytical protocol. The official method also showed a high sensitivity for HL adulterants even when present in low amounts (≥2%). However, it proved unsuitable for the detection of adulteration with HO oils at 5–10 % as they have a similar TAG and FA composition to OO. This is a worrisome result, given that HO oils are emerging adulterants for the illegal blending of OO. While higher sensitivity values have been reported for the official method in HZO detection (García-González et al., 2007), our results agree with those of another study, that set the minimum detectable amount of HZO in OO at 20 % by this method (Moreda et al., 2003). The present results can be considered as representative of the efficacy of the method, as HO oil from different sources (HZO and HOSFO) and suppliers were included in the modelling and validation. Poor sensitivity towards HO adulterants, also reported by other authors (Beccaria et al., 2016), can thus be added to the other drawbacks of the official method. As well as being a time-consuming analytical procedure, requiring two different analyses per sample, it involves the use of highly contaminating organic solvents such as propionitrile, and lacks robustness due to the low resolution of the HPLC-RID peaks, specifically the very minor ECN42 cluster that includes the LLL peak (Beccaria et al., 2016).

In contrast, the new screening methods proved more suitable for the detection of adulterated samples, providing high sensitivity values regardless of the adulterant composition (Table 2). Among the developed screening tools, the HT-GC–MS method performed the best in terms of sensitivity (0.99 for HL and 0.88 for HO models). High sensitivity and a relatively short analysis time, allowing the rapid analysis of a high number of samples, are of paramount importance for a screening method (López et al., 2014). HT-GC–MS was also the screening tool that showed the highest specificity value (0.76). Although this value was lower than the one of the official method, it is an acceptable value for a screening method, given that confirmatory analyses or checks can be carried out afterwards to identify false positives (Magnusson & Örnemark, 2014). Moreover, it is important to highlight that these values were obtained with a sampling of a high natural variability as it included a high number of genuine olive oils of different varieties, regions and harvests, and various samples of various adulterant types, blended at low adulteration levels. These results demonstrate that a fingerprinting approach applied to multidimensional data such as

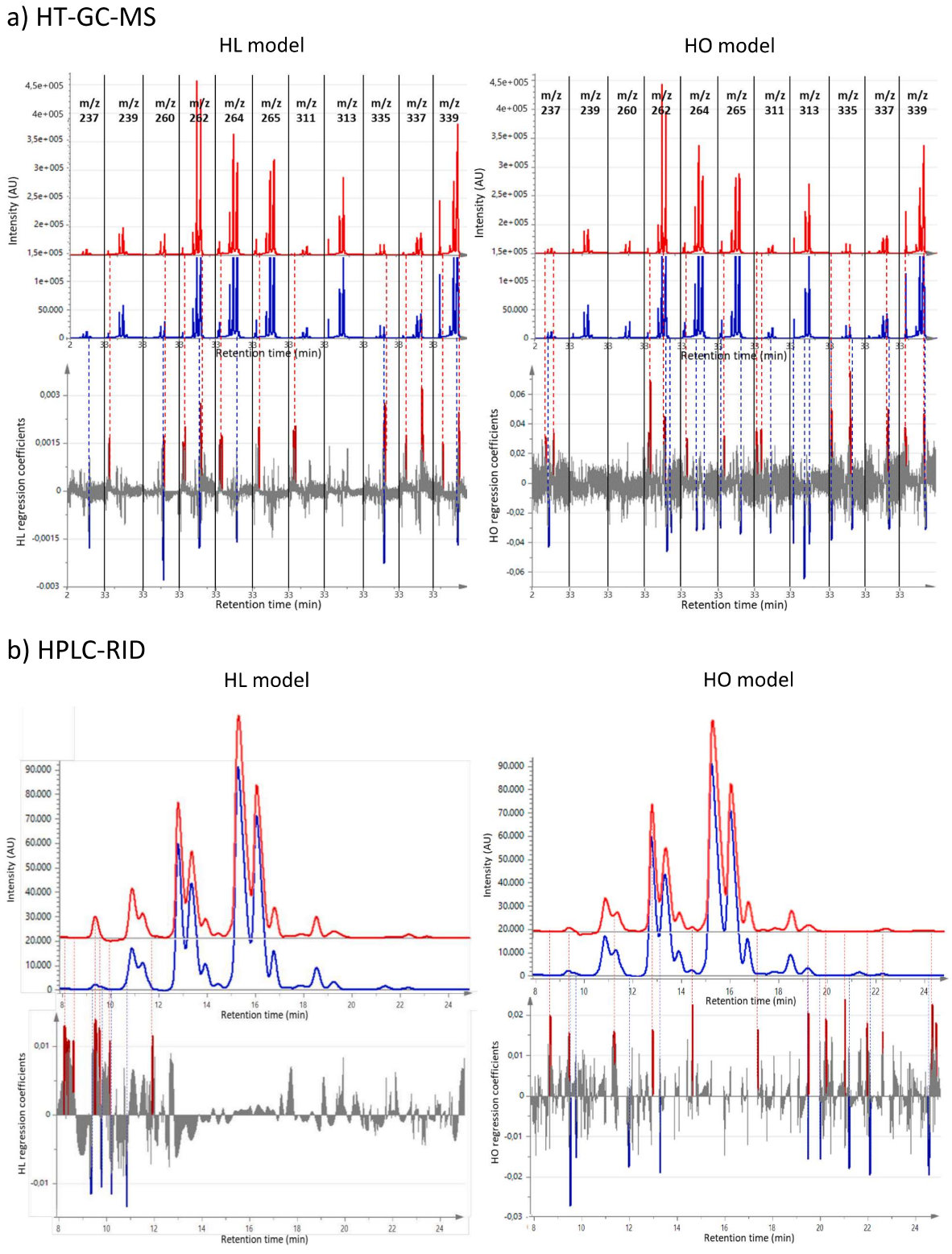


Fig. 3. Regression coefficients of the high linoleic (HL) and high oleic (HO) PLS-DA models on the a) HT-GC–MS fingerprint, plotted against the concatenated EICs of an adulterated sample (red EICs) and a genuine olive oil (blue EICs); and b) HPLC-RID fingerprint, plotted against the chromatographic profile of an adulterated sample (red chromatogram) and a genuine olive oil (blue chromatogram). For each model, relevant coefficients for the prediction of the adulterated class are highlighted in red (positive coefficient) and blue (negative coefficient).

GC–MS, can exploit the valuable specific information about authentication markers in the unfolded matrix built with specific TAG ions.

The method based on HPLC-RID also followed a fingerprinting approach, but its performance was lower compared to HT-GC–MS. This can be attributed to the RID-chromatographic data, which did not provide any specific information on these markers or distinguish unresolved or overlapping TAGs. Nonetheless, the HPLC-RID fingerprint effectively detected adulterated samples (sensitivity of 0.96 for HL and 0.82 for HO models) (Table 2) and is the most affordable and easily applicable of all the methods tested here. Considering that it is based on part of the data produced by the official method, the implementation of both in parallel could improve the detection of HO blends.

Finally, the FIA-ESI-HRMS method proved as successful as HPLC-RID in terms of overall sensitivity (0.89), showing a similar ability to detect both HL (0.90) and HO (0.88) adulteration adulterants (Table 2). The specificity value of this method (0.50) was low compared with that of previous models developed with the same method with a larger sample set (0.80) (Quintanilla-Casas et al., 2021). This fact indicates that the model developed upon this untargeted TAG profiling data seems to be highly dependent on the size of the sample set. Nonetheless, in view of the high sensitivity and extremely short analysis time (<2 min per sample) of this high-throughput method, its use with large-scale sampling could still be considered as a screening tool to be used in parallel with the current control practices.

Exploring the regression coefficients of the models developed for each screening method allowed us to tentatively ascertain the variables that contributed most to discriminate between adulterated and genuine samples. In all models, the highest regression coefficients corresponded to minor TAG species distributed within the entire experimental mass range. Although the structural elucidation of discriminant markers was not the scope of the present comparative study, some TAG species were tentatively identified in correspondence with the most relevant coefficients. Several of them were highly discriminant in models generated for different screening methods, demonstrating the consistency of the results. Particularly, TAGs that could correspond to PoLL/PLLn, OLLn, LLLn and LLL were highly discriminant for the HL models, whereas species that may match TAGs such as OOLn and OLLn and various other TAGs including Po, A or B were significant for the HO models.

In view of these results, the developed screening methods, especially HT-GC–MS TAG fingerprinting, could represent useful tools to facilitate inspections by official control bodies, improving the risk analysis on which they are currently based (Cugat & Biel, 2016). In particular, the high sensitivity of HT-GC–MS, an affordable and green analytical technique, indicates it has potential as a fit-for-purpose screening tool able to process a high number of samples and enhance the effectiveness of current official controls. Nonetheless, to confirm its high performance, this promising tool needs further testing with an enriched model, including a greater number of both genuine and adulterated samples.

5. Conclusion

Three different screening methods were tested as tools to support and improve official controls, being able to detect ≥2 % of HL and ≥5 % of HO adulterants in OO with sensitivity values >0.90 and >0.82, respectively. Among them, the HT-GC–MS method was the screening tool showing the best performance, with a 0.99 and a 0.88 sensitivity values for HL and for HO adulterants, respectively, and a specificity of 0.76, being the most promising screening tool tested for OO authentication.

The FIA-ESI-HRMS and HPLC-RID methods provided lower specificity values. However, these methods deserve to be further explored as possible supporting screening tools given that they still provide the high sensitivity desired for screening methods. FIA-ESI-HRMS is a fast technique proved to be suitable when assayed in large scale studies, and the HPLC-RID method can be run in parallel with the official OO analysis by using the same raw data.

The present study allowed to compare three screening methods among them and with the official method providing relevant information about their performance when applied to a dataset of genuine and adulterated samples with a high natural variability. On these bases, optimal models should be further developed and evaluated using a large scale dataset.

Funding

This work was developed in the context of the project AUTENFOOD, supported by ACCIÓ-Generalitat de Catalunya (Spain) and the European Union through the ERDF program 2014–2020 (Grant No COMRDI-15-1-0035). B. Torres-Cobos and B. Quintanilla-Casas thanks the Spanish Ministry of Universities predoctoral fellowships FPU20/01454 and FPU16/01744. A. Tres received a Ramon y Cajal grant (RYC-2017-23601) funded by MCIN/AEI/10.13039/501100011033 and by "ESF Investing in your future".

CRediT authorship contribution statement

Berta Torres-Cobos: Formal analysis, Investigation, Methodology, Validation, Data curation, Visualization, Writing – original draft. **Beatriz Quintanilla-Casas:** Formal analysis, Investigation, Methodology, Data curation, Validation, Writing – review & editing. **Giulia Vicario:** Formal analysis, Investigation. **Francesc Guardiola:** Supervision, Writing – review & editing. **Alba Tres:** Conceptualization, Methodology, Resources, Supervision, Project administration, Funding acquisition, Writing – review & editing. **Stefania Vichi:** Conceptualization, Methodology, Resources, Supervision, Project administration, Funding acquisition, Writing – review & editing.

Declaration of Competing 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

Data will be made available on request.

Acknowledgment

The authors are grateful to Víctor Casado Trujillo and Luis Sanchez-Peña for their help in extracting FIA-ESI-HRMS data and to Catalan virgin olive oil producers for providing traceable samples.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2022.135256.

References

Aparicio, R., Morales, M. T., Aparicio-Ruiz, R., Tena, N., & García-González, D. L. (2013). Authenticity of olive oil: Mapping and comparing official methods and promising alternatives. *Food Research International, 54*, 2025–2038. https://doi.org/10.1016/j.foodres.2013.07.039

Arlorio, M., Coisson, J. D., Bordiga, M., Travaglia, F., Garino, C., Zuidmeer, L., … Martelli, A. (2010). Olive oil adulterated with hazelnut oils: Simulation to identify possible risks to allergic consumers. *Food Additives & Contaminants: Part A, 27*(1), 11–18. https://doi.org/10.1080/02652030903225799

Ballin, N. Z., & Laursen, K. H. (2019). To target or not to target? Definitions and nomenclature for targeted versus non-targeted analytical food authentication. *Trends in Food Science and Technology, 86*, 537–543. https://doi.org/10.1016/j.tifs.2018.09.025

Barber, M., Merren, A. E., & Kelly, W. (1964). The mass spectrum of large molecules I. The triglycerides of straight chain fatty acids. *Tetrahedron Letters, 18*, 1063–1067.

Beccaria, M., Moret, E., Purcaro, G., Pizzale, L., Cotroneo, A., Dugo, P., … Conte, L. S. (2016). Reliability of the ΔeCN42 limit and global method for extra virgin olive oil purity assessment using different analytical approaches. *Food Chemistry, 190*, 216–225. https://doi.org/10.1016/j.foodchem.2015.05.075

Casadei, E., Valli, E., Panni, F., Donarski, J., Farrús Gubern, J., Lucci, P., … Gallina Toschi, T. (2021). Emerging trends in olive oil fraud and possible countermeasures. *Food Control, 124*, Article 107902. https://doi.org/10.1016/J.FOODCONT.2021.107902

Cavanna, D., Righetti, L., Elliott, C., & Suman, M. (2018). The scientific challenges in moving from targeted to non-targeted mass spectrometric methods for food fraud analysis: A proposed validation workflow to bring about a harmonized approach. *Trends in Food Science and Technology, 80*, 223–241. https://doi.org/10.1016/j.tifs.2018.08.007

Christopoulou, E., Lazaraki, M., Komaitis, M., & Kaselimis, K. (2004). Effectiveness of determinations of fatty acids and triglycerides for the detection of adulteration of olive oils with vegetable oils. *Food Chemistry, 84*, 463–474. https://doi.org/10.1016/S0308-8146(03)00273-5

Conte, L., Bendini, A., Valli, E., Lucci, P., Moret, S., Maquet, A., … Gallina Toschi, T. (2020). Olive oil quality and authenticity: A review of current EU legislation, standards, relevant methods of analyses, their drawbacks and recommendations for the future. *Trends in Food Science & Technology, 105*, 483–493. https://doi.org/10.1016/J.TIFS.2019.02.025

Cugat, G., & Biel, M. R. (2016). Official quality controls and the fight against fraud in the olive oil sector in Catalonia. Results 2011–2014. *Grasas y Aceites, 67*(4), e160.

Esslinger, S., Riedl, J., & Fauhl-Hassek, C. (2014). Potential and limitations of non-targeted fingerprinting for authentication of food in official control. *Food Research International, 60*, 189–204. https://doi.org/10.1016/j.foodres.2013.10.015

European Commission. (2020). *Study on the implementation of conformity checks in the olive oil sector throughout the European Union (Contract AGRI-2018-0485)*. www.areteonline.net.

European Union. (2021). *2020 Annual Report. The EU Agri-Food Fraud Network and the Administrative Assistance and Cooperation System*. doi: 10.2875/20163.

Everstine, K., Spink, J., & Kennedy, S. (2013). General interest economically motivated adulteration (EMA) of food: Common characteristics of EMA incidents. *Jornal of Food Protection, 76*(4), 723–735. https://doi.org/10.4315/0362-028X.JFP-12-399

García-González, D. L., Tena, N., & Aparicio, R. (2007). Characterization of olive paste volatiles to predict the sensory quality of virgin olive oil. *European Journal of Lipid Science and Technology, 109*(7), 663–672. https://doi.org/10.1002/EJLT.200700056

Gelpí, E., Posada de la Paz, M., Terracini, B., Abaitua, I., Gómez de la Cámara, A., Kilbourne, E. M., … Soldevilla, L. (2002). The Spanish toxic oil syndrome 20 years after its onset: A multidisciplinary review of scientific knowledge. *Environmental Health Perspectives, 110*(5), 457–464. http://ehpnet1.niehs.nih.gov/docs/2002/110p457-464gelpi/abstract.html.

López, M. I., Colomer, N., Ruisánchez, I., & Callao, M. P. (2014). Validation of multivariate screening methodology. Case study: Detection of food fraud. *Analytica Chimica Acta, 827*, 28–33. https://doi.org/10.1016/j.aca.2014.04.019

Magnusson, B., & Örnemark, U. (2014). *Eurachem guide: The Fitness for Purpose of Analytical Methods - A Laboratory Guide to Method Validation and Related Topics* (Eurachem, Ed.; 2nd ed.). www.eurachem.org.

Mailer, R., & Gafner, S. (2020). Adulteration of olive oil. *Botanical Adulterants Prevention Bulletin. Austin, TX: ABC-AHP-NCNPR Botanical Adulterants Prevention Program*, 1–14.

Meenu, M., Cai, Q., & Xu, B. (2019). A critical review on analytical techniques to detect adulteration of extra virgin olive oil. *Trends in Food Science & Technology, 91*, 391–408. https://doi.org/10.1016/j.tifs.2019.07.045

Moreda, W., Pérez-Camino, M. C., & Cert, A. (2003). Improved method for the determination of triacylglycerols in olive oils by high performance liquid chromatography. *Grasas y Aceites, 54*(2), 175–179.

Muñoz-Olivas, R. (2004). Screening analysis: An overview of methods applied to environmental, clinical and food analyses. *TrAC – Trends in Analytical Chemistry, 23*(3), 203–216. https://doi.org/10.1016/S0165-9936(04)00318-8

Quintanilla-Casas, B., Bertin, S., Leik, K., Bustamante, J., Guardiola, F., Valli, E., … Vichi, S. (2020). Profiling versus fingerprinting analysis of sesquiterpene hydrocarbons for the geographical authentication of extra virgin olive oils. *Food Chemistry, 307*, Article 125556. https://doi.org/10.1016/j.foodchem.2019.125556

Quintanilla-Casas, B., Strocchi, G., Bustamante, J., Torres-Cobos, B., Guardiola, F., Moreda, W., … Vichi, S. (2021). Large-scale evaluation of shotgun triacylglycerol profiling for the fast detection of olive oil adulteration. *Food Control, 123*, Article 107851. https://doi.org/10.1016/J.FOODCONT.2020.107851

Regulation No 2568/91 on the characteristics of olive oil and olive-residue oil and on the relevant methods of analysis, 1, (1991).

Ruiz-Samblás, C., González-Casado, A., & Cuadros-Rodríguez, L. (2015). Triacylglycerols determination by high-temperature gas chromatography in the analysis of vegetable oils and foods: A review of the past 10 years. *Critical Reviews in Food Science and Nutrition, 55*(11), 1618–1631. https://doi.org/10.1080/10408398.2012.713045

Ruiz-Samblás, C., González-Casado, A., Cuadros-Rodríguez, L., & García, F. P. R. (2010). Application of selected ion monitoring to the analysis of triacylglycerols in olive oil by high temperature-gas chromatography/mass spectrometry. *Talanta, 82*(1), 255–260. https://doi.org/10.1016/j.talanta.2010.04.030

Ruiz-Samblás, C., Marini, F., Cuadros-Rodríguez, L., & González-Casado, A. (2012). Quantification of blending of olive oils and edible vegetable oils by triacylglycerol fingerprint gas chromatography and chemometric tools. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences, 910*, 71–77. https://doi.org/10.1016/j.jchromb.2012.01.026

The food Integrity project. (2018). *FoodIntegrity Handbook - A guide to food authenticity issues and analytical solutions* (J.-F. Morin & M. Lees, Eds.). Eurofins Analytics France.

Torres-Cobos, B., Quintanilla-Casas, B., Romero, A., Ninot, A., Alonso-Salces, R. M., Gallina Toschi, T., … Vichi, S. (2021). Varietal authentication of virgin olive oil: Proving the efficiency of sesquiterpene fingerprinting for Mediterranean Arbequina oils. *Food Control, 128*, Article 108200. https://doi.org/10.1016/J.FOODCONT.2021.108200

Vichi, S., Cortés-Francisco, N., & Caixach, J. (2012). Ultrahigh resolution mass spectrometry and accurate mass measurements for high-throughput food lipids profiling. *Journal of Mass Spectrometry, 47*(9), 1177–1190. https://doi.org/10.1002/jms.3017

Yan, J., Erasmus, S. W., Aguilera Toro, M., Huang, H., & van Ruth, S. M. (2020). Food fraud: Assessing fraud vulnerability in the extra virgin olive oil supply chain. *Food Control, 111*, Article 107081. https://doi.org/10.1016/J.FOODCONT.2019.107081

5.3.4 Unsaponifiable fraction fingerprinting

The UF is a promising marker for verifying hazelnut varietal and geographical authenticity, as revised in section 1.3.4.

In the specific case of UF analysis by GC-MS, an untargeted approach may be highly advantageous. This approach allows for the extraction of all information from complex chromatograms, including minor species and poorly resolved peaks, without the time-consuming steps of compound identification and quantification.

Among untargeted approaches, two promising tools stand out for handling the high-dimensional data generated by GC-MS of the UF:

- Fingerprinting analysis of the unfolded matrix generated by concatenating the EICs.
- Untargeted profiling of the three-way GC–MS data using PARADISe, a deconvolution and identification tool based on PARAFAC2.

Consequently, a PoC study was conducted to evaluate the potential of the UF as a marker for hazelnut varietal and geographical authentication (Publication 7). In this PoC study, first, a new analytical method was developed to extract, purify and analyse the UF of the hazelnuts by GC-MS (**Figure 12**). Then, authentication models were developed by both approaches (fingerprinting and untargeted profiling) to compare their discrimination efficiency and to select the most promising approach for verifying hazelnut authenticity (Publication 7).

A set of 176 hazelnut samples harvested over two consecutive years (2019-2020) was used in this PoC study. This set included:

- 110 samples of the TG cultivar from three regions across different countries: Chile (CHL, n = 40), Italy (ITA, n = 24), and Spain (ESP, n = 46),

- 66 samples of different cultivars (non-TG) produced solely in Spain.

Two independent PLS-DA classification models were built following both untargeted approaches:

1) TG cultivar model: A binary PLS-DA to distinguish TG from other hazelnut cultivars (non-TG) grown in Spain. Only samples from the same origin (Spain) were considered (n = 112) to build and validate the model, avoiding influences other than the cultivar.

2) Geographical origin model: only samples of the same cultivar (TG) were used (n = 110). Two binary PLS-DA models were developed: one distinguishing between EU and non-EU (Chile) hazelnuts, and another discriminating EU samples by their region of origin, located in Italy or Spain.

Finally, to gain deeper understanding of the models and their chemical significance, the PLS-DA regression coefficients of the three models and from both approaches were explored (Publication 7).

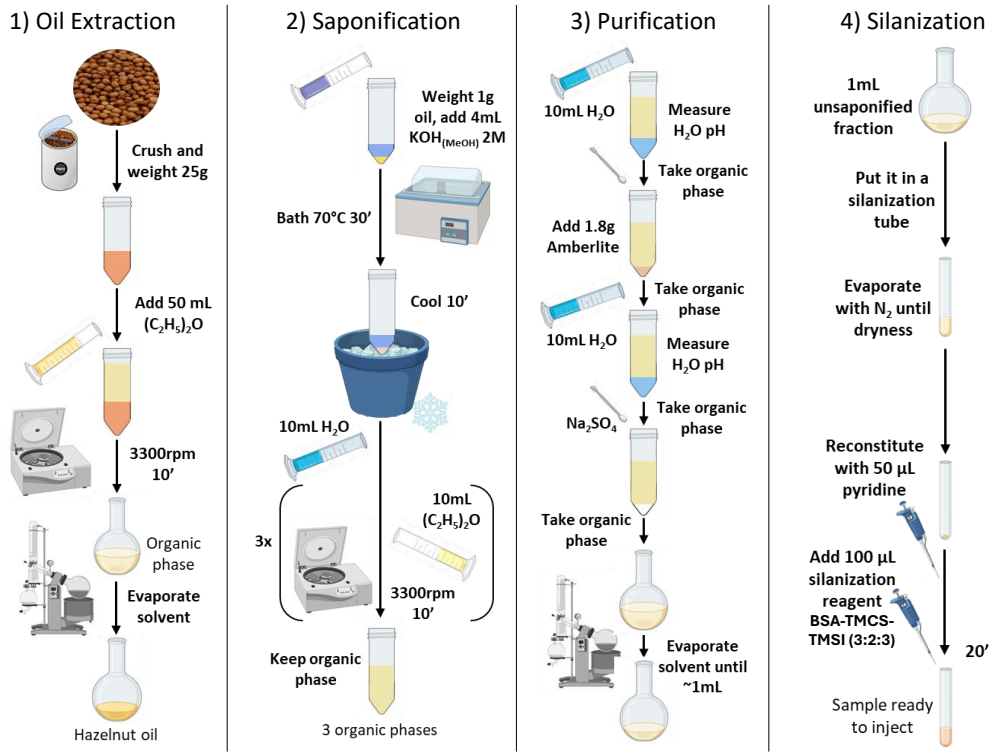


Figure 12. Graphical scheme of the process to obtain the UF of hazelnuts.

5.3.5 Publication 7

Prospective exploration of hazelnut's unsaponifiable fraction for geographical and varietal authentication: A comparative study of advanced fingerprinting and untargeted profiling techniques

Berta Torres-Cobos, Beatriz Quintanilla-Casas, Mercè Rovira, Agustí Romero, Francesc Guardiola, Stefania Vichi, Alba Tres

Food Chemistry, ***2024****, 441, 139083. https://doi.org/10.1016/j.foodchem.2023.138294*

Supplementary material available in **Annex 6**

Food Chemistry 441 (2024) 138294

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem

Prospective exploration of hazelnut's unsaponifiable fraction for geographical and varietal authentication: A comparative study of advanced fingerprinting and untargeted profiling techniques

B. Torres-Cobos [a,b], B. Quintanilla-Casas [c], M. Rovira [d], A. Romero [d], F. Guardiola [a,b], S. Vichi [a,b,*], A. Tres [a,b]

[a] University of Barcelona, Department of Nutrition, Food Sciences and Gastronomy, Prat de la Riba 171, Santa Coloma de Gramenet 08921, Spain
[b] University of Barcelona, Institute of Research on Food Nutrition and Safety (INSA-UB), Prat de la Riba 171, Santa Coloma de Gramenet 08921, Spain
[c] University of Copenhagen, Department of Food Science, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark
[d] Institute of Agrifood Research and Technology (IRTA), Ctra. de Reus – El Morell Km 3.8, Constantí 43120, Spain

ARTICLE INFO

Keywords:
Hazelnut
Geographical and varietal authentication
Unsaponifiable fraction
Fingerprinting
Untargeted profiling
PLS-DA

ABSTRACT

This study compares two data processing techniques (fingerprinting and untargeted profiling) to authenticate hazelnut cultivar and provenance based on its unsaponifiable fraction by GC–MS. PLS-DA classification models were developed on a selected sample set (n = 176). As test cases, cultivar models were developed for "Tonda di Giffoni" vs other cultivars, whereas provenance models were developed for three origins (Chile, Italy or Spain). Both fingerprinting and untargeted profiling successfully classified hazelnuts by cultivar or provenance, revealing the potential of the unsaponifiable fraction. External validation provided over 90 % correct classification, with fingerprinting slightly outperforming. Analysing PLS-DA models' regression coefficients and tentatively identifying compounds corresponding to highly relevant variables showed consistent agreement in key discriminant compounds across both approaches. However, fingerprinting in selected ion mode extracted slightly more information from chromatographic data, including minor discriminant species. Conversely, untargeted profiling acquired in full scan mode, provided pure spectra, facilitating chemical interpretability.

1. Introduction

Hazelnuts are widely used raw, roasted and as a key ingredient in food and confectionery products, adding flavour and texture to various sweet and savoury products. They rank third in the global nut market with a production volume higher than one million tons per year (FAOstat, 2021). The main hazelnut producing countries are Turkey (63.5 %), Italy (7.9 %), United States (6.5 %), Azerbaijan (6.3 %), Georgia (4.3 %) and Chile (3.3 %), followed by China, Iran, France and Spain (<3% each) (FAOstat, 2021). Among the most prominent cultivars are 'Tombul', 'Palaz', 'Çakildak' (Turkey); 'Tonda di Giffoni', 'Tonda Romana', 'Tonda Gentile delle Langhe' (Italy); 'Negret' and 'Pauetet' (Spain) (Król & Gantner, 2020). Sensory and qualitative attributes of hazelnuts are strongly influenced by varietal and geographical factors (Amaral et al., 2006; Król & Gantner, 2020; Parcerisa, Richardson, Rafecas, Codony, & Boatella, 1998). Their prices can also vary greatly based on their cultivar and geographical origin (FAOstat, 2021), being higher for hazelnuts with special geographical indications such as Protected Designation of Origin (PDO) or Protected Geographical Indication registered in the European Union. The great value of hazelnuts makes them susceptible to economically motivated fraud, which is further aggravated by the growth of emerging nut producing countries, the expansion of markets and the lack of effective fraud detection methods. All these factors contribute to a growing vulnerability that counterfeiters can exploit. Hence, having suitable tools to verify the cultivar and origin of hazelnuts is crucial to guarantee their authenticity and to protect the consumer.

In this regard, phenotypic observations based on physical characteristics are currently used for this purpose. However, the fact that they are susceptible to external influences and can only be used on whole kernels (Ciarmiello et al., 2014; Król & Gantner, 2020) limits their efficiency. For this reason, several studies to explore more suitable tools for hazelnut authentication have been carried out in the last decade.

* Corresponding author at: Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Campus De l'Alimentació Torribera, Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona, Av Prat de la Riba, 171, 08921 Santa Coloma de Gramenet, Spain.
E-mail address: stefaniavichi@ub.edu (S. Vichi).

https://doi.org/10.1016/j.foodchem.2023.138294
Received 30 June 2023; Received in revised form 22 December 2023; Accepted 26 December 2023
Available online 4 January 2024

DNA-based methods offer high accuracy (Lang et al., 2021), but their complex and expensive procedures make them unsuitable for routine analysis. Furthermore, both phenotypic and DNA-based methods are limited to varietal authentication only. In contrast, methods for geographical authentication of nuts often involve analysing the mineral composition (Inaudi et al., 2020; Oddone, Aceto, Baldizzone, Musso, & Osella, 2009) or isotope ratios of light or heavy elements (Krauß, Vieweg, & Vetter, 2019; Zannella et al., 2017) associated with the growing area. Unfortunately, these methods are not suitable for determining the cultivar. Similarly, models based on near infrared spectrometry (Biancolillo et al., 2018; Sammarco & Dall'Asta, & Suman, 2023) have been suggested to verify the geographical origin of some Italian PDOs. Alternatively, methods based on the analysis of hazelnut metabolites have been proposed to verify their varietal or geographical origin, given that metabolomics is a state-of-the-art approach for food authentication. These approaches analyse protein/peptide compounds, phenolic profiles and components of the lipid fraction, measured by chromatographic techniques, such as gas (Parcerisa et al., 1998; Tüfekci and Karataş, 2018) and liquid chromatography (Ciarmiello et al., 2014; Ghisoni et al., 2020; Klockmann, Reiner, Bachmann, Hackl, & Fischer, 2016) as well as proton nuclear magnetic resonance (^{1}H NMR) (Bachmann, Klockmann, Haerdter, Fischer, & Hackl, 2018). Nevertheless, some of these markers such as phenols may be unstable under certain conditions (light, temperature, time) and none of the methods developed so far have been tested for their suitability in verifying both the hazelnut's cultivar and origin. This underscores the need reliable methods that can fulfil this objective.

In the pursuit of appropriate candidates for hazelnut cultivar and geographical markers, the unsaponifiable lipid fraction stands out for presenting relatively stable metabolites under storage conditions, which are known to be influenced by both genetic (Amaral et al., 2006; Matthäus & Özcan, 2012; Parcerisa et al., 1998) and environmental factors (Benitez-Sanchez, León-Camacho, & Aparicio, 2003; Ghisoni et al., 2020; Matthäus & Özcan, 2012). This rich fraction contains several families of secondary metabolites such as linear and terpene alcohols and hydrocarbons, sterols, methylsterols and dimethylsterols, among others (Benitez-Sanchez et al., 2003; Goriainov et al., 2021). This makes the unsaponifiable lipid fraction a promising candidate for the geographical and varietal authentication of hazelnuts. The most appropriate technique for its analysis is gas chromatography-mass spectrometry (GC–MS) (Goriainov et al., 2021; Phillips, Ruggio, & Ashraf-Khorassani, 2005), as it provides comprehensive molecular-level information through three-way data (an array sized of intensity × retention time × *m/z*, for each sample), together with a high sensitivity and widespread availability in routine labs.

In addition to selecting the appropriate authentication markers and analytical technique, an essential aspect of the authentication strategy is the data analysis approach, especially when dealing with complex chromatograms such as those from nut's unsaponifiable lipid fraction. Untargeted approaches are an advantageous alternative to conventional targeted methods, as they provide more information and overcome the difficulties of identifying and quantifying analytical compounds in complex chromatograms (Quintanilla-Casas et al., 2020a). In fact, untargeted methods coupled to chemometric pattern recognition techniques, such as partial least squares discriminant analysis (PLS-DA), proved to be efficient tools for authentication purposes (Quintanilla-Casas et al., 2020a,b; Riedl, Esslinger, & Fauhl-Hassek, 2015; Torres-Cobos et al., 2021).

Among untargeted methods, fingerprinting operates on high dimensional data (i.e. two-way or three-way data) such as spectra or chromatograms and consists on finding specific patterns, known as fingerprints, which are unique to a specific characteristic of the food sample, such as cultivar or geographical origin (Ballin & Laursen, 2019; Bosque-Sendra, Cuadros-Rodríguez, Ruiz-Samblás, & de la Mata, 2012). Fingerprinting methods have been widely tested for food authentication, proving to be successful (Quintanilla-Casas et al., 2020a,b; Torres-Cobos et al., 2021). Fingerprinting of three-way data, such as GC–MS data (intensity × retention time × *m/z*, for each sample), typically entails complex multi-way chemometric algorithms, but a recently introduced approach simplifies the process by transforming the data into a manageable two-way format (retention time × intensity, for each ion and sample). This process involves the creation of models using the unfolded matrix of extracted chromatograms of specific ions and has proven successful for authentication purposes (Quintanilla-Casas et al., 2020a,b; Torres-Cobos et al., 2021, 2023). Another alternative for analysing three-way GC–MS data is using advanced untargeted profiling techniques, such as powerful deconvolution tools, to extract the maximum information from the samples (Rinnan, Amigo, & Skov, 2014). Among them, a deconvolution and identification tool called PARADISe has emerged for GC–MS, which is based on PARAllel FACtor analysis 2 (PARAFAC2). PARAFAC2 models provide estimates for each mode – relative concentration, elution profile and pure mass spectra – for each analyte, while handling common issues in chromatographic data, including co-elution, baseline variations and retention time shifts (Johnsen, Skou, Khakimov, & Bro, 2017; Baccolo, Quintanilla-Casas, Vichi, Augustijn, & Bro, 2021). This user-friendly software allows an efficient untargeted analysis of large GC–MS datasets, while minimizing inter-user variability. Several studies have reported its usefulness for chromatographic datasets in different fields (Johnsen et al., 2017; Baccolo et al., 2021; Ríos-Reina, Aparicio-Ruiz, Morales, & García-Gonzalez, 2023; Sales, Portolés, Johnsen, Danielsen, & Beltran, 2019). Baccolo et al. (2021) evidenced its advantages in time-saving, comprehensiveness of the chromatographic results and tentative identification over manual profiling. To our knowledge, no comparisons have been made to evaluate the efficiency of fingerprinting and untargeted profiling approaches, along with other deconvolution methods, in extracting information from chromatographic data for authentication purposes.

This study aims to explore the potential of the unsaponifiable fraction for the cultivar and geographical authentication of hazelnuts and to compare two different data processing techniques (fingerprinting and untargeted profiling) to determine the best method for developing efficient authentication models. For this purpose, PLS-DA classification models based on GC–MS data of hazelnut unsaponifiable fraction were developed using both approaches. These were applied to the same test cases: a cultivar model to distinguish 'Tonda di Giffoni' (TG), one of the most widespread cultivars in the world, from other hazelnut cultivars (non-TG); a provenance model to discriminate three different countries of origin (Spain, Chile and Italy). Finally, the regression coefficients of the models have been explored with the sole purpose of gaining a deeper understanding of the models and their chemical significance, to ensure that models are based on genuine chemical information rather than arbitrary randomness.

2. Material and methods

2.1. Material and reagents

Diethyl ether stabilized with 7 mg/L of BHT, anhydrous sodium sulphate and anhydrous pyridine 99.5 % were purchased from Scharlau (Sentmenat, Spain). Methanol for gas chromatography ECD and FID SupraSolv® and Horning's silylating mixture II (N,O-bis(trimethylsilyl) acetamide/chlorotrimethylsilane/1-(trimethylsilyl)imidazole, 3:2:3, v/v/v) were purchased from Merck (Darmstadt, Germany). Potassium hydroxide 85 % for analysis in pellets form was purchased from Thermo Scientific (Waltham, Massachusetts, USA) and amberlite IRN78 OH hydroxide form from Supelco (Bellefonte, Pennsylvania, USA).

2.2. Sampling

The sample set consisted of 176 traceable hazelnuts collected over two consecutive harvest years, 2019 and 2020 (Supplementary material, Table S1). They were obtained in the framework of the TRACENUTS

project (PID2020-117701RB100), directly from producers. Out of these samples, 110 were of the TG cultivar from Chile (n = 40), Italy (n = 24) and Spain (n = 46), while 66 were from different cultivars (non-TG) produced in Spain. Samples were stored vacuum-packed at 4 °C until analysis.

2.2.1. Sample preparation

Around 30 g of hazelnuts were grinded and their lipid fraction was extracted using 50 mL of diethyl ether. The mixture was centrifuged at 1220 *g* for 10 min, the liquid phase was taken and the organic solvent was evaporated with a rotatory evaporator until only the hazelnut oil was left. Then, an aliquot of 1 g of the hazelnut oil was saponified by adding 4 mL of 2 M methanolic potassium hydroxide solution and heated for 30 min at 70 °C in a water bath. The reaction was quenched with ice for 10 min and 10 mL of water were added. Once the sample reached room temperature, the unsaponifiable fraction was extracted with 3 x 10 mL of diethyl ether, centrifuging each time (1220 *g*; 10 min) to separate the organic phase from the aqueous phase. The organic extracts were subsequently pooled and washed with 10 mL of distilled water. Following this, 2 g of amberlite adsorbent were added to remove the excess of dissolved free fatty acids. After removing the adsorbent, the organic phase was washed again with 10 mL of water and anhydrous sodium sulphate was added to remove any remaining moisture. Once the extract was cleaned, purified and any residual water was removed, the solvent was evaporated using a rotatory evaporator until the volume was reduced to approximately 1 mL. The resulting solution was transferred into a silylation tube, and the remaining solvent was evaporated to dryness by applying a stream of N_2. The dry unsaponifiable fraction was reconstituted with 50 µL of pyridine. Finally, 100 µL of silylating reagent were added and allowed to react for 20 min at room temperature prior to injection.

2.3. Gas chromatography-mass spectrometry (GC–MS)

The samples were analysed by an Agilent 6890 N Network GC system equipped with a Combi-pal autosampler (CTC Analytics, Zwingen, Switzerland) and coupled to an Agilent 5975C Inert MSD quadrupolar mass selective analyser (Agilent Technologies, Santa Clara, California, USA). Helium was the carrier gas at a flow rate of 1.5 mL/min. Analytes were separated on a ZB-5 ms capillary column (60 m × 0.25 mm i.d., 0.25 µm film thickness) from Phenomenex (Torrance, California, USA). Column temperature was initially held at 150 °C for 2 min, then increased to 260 °C at a rate of 10 °C/min, held for 2 min and then increased to 320 °C at 2 °C/min, holding the last temperature for 13 min. The ion source and the transfer line were set at 230 and 300 °C, respectively. Mass spectra were acquired at 1.9 scan/s and the electron energy was set at 70 eV. For the untargeted profiling approach, data acquisition was performed in the full scan mode within the 50–500 *m/z* range. For the fingerprinting approach, data were acquired using selected ion monitoring (SIM) of 15 ions that resulted characteristic of several compound families of the unsaponifiable fraction: *m/z* 57 (linear hydrocarbons); *m/z* 69, 81, 93 (terpene alcohols and hydrocarbons); *m/z* 73 (silylated compounds, i.e. any compound with a hydroxy group); *m/z* 75, 103 (linear alcohols); *m/z* 83, 117 (fatty acids); *m/z* 117, 129, 189, 199, 204, 218, 393 (sterols) and *m/z* 189, 204, 218 (4-methylsterols and 4,4-dimethylsterols) (Li, Beveridge, & Drover, 2007; Xu et al., 2018). Ions with *m/z* 75, 81, 83, 93, 103 and 117 were acquired from 14.85 to 42.5 min; *m/z* 129, 57, 69 and 73 were acquired from 14.85 to 58 min; *m/z* 189, 199, 204, 218 and 393 were acquired from 42.5 to 58 min.

2.4. Chemometrics

2.4.1. SIM fingerprinting approach

A fingerprinting approach was followed using the extracted ion chromatograms (EIC) of the 15 selected ions. The intensities of the scans from minute 14.85 to 58 (4903 scans per selected ion over 43 min) were considered for ions *m/z* 129, 57, 69 and 73; from 14.85 to 42.5 min for ions *m/z* 75, 81, 83, 93, 103 and 117; and from 42.5 to 58 min for ions *m/z* 189, 199, 204, 218 and 393 (4903 scans × 4 ions + 3171 scans × 6 ions + 1785 scans × 5 ions = 47563 variables per sample). A data matrix was built for each ion, with the scan intensities of each EIC (columns) for all samples (rows). Then, the EICs of each ion matrix were aligned using the Correlation Optimized Warping (COW) algorithm in Matlab® (Nielsen, Carstensen, & Smedsgaard, 1998) to correct the retention time shifts among samples. Finally, the 15 aligned EIC matrices were concatenated conforming a two-way unfolded matrix (176 samples × 47563 variables).

2.4.2. Untargeted profiling approach using PARADISe

The GC–MS raw dataset, acquired in full scan mode from 14.85 to 58 min, was imported in PARADISe and aligned to solve peak shifts using the automatic alignment tool that applies icoshift (Larsen, Van den Berg, & Engelsen, 2006) and COW algorithms (Tomasi, Van den Berg, & Andersson, 2004). Even if raw data alignment is not required, it eases the subsequent interval selection for each of the peaks. The optimal number of components of PARAFAC2 models for each interval was determined based on the combination of the following parameters: high number of true peaks based on the deep learning tool, high model fitting, high core consistency indicating better model adequacy, as well as low and random model residuals (Quintanilla-Casas, Bro, Hinrich, & Davie-Martin, 2023). After excluding baseline noise and other interferences, components corresponding to actual chemical compounds (Fig. 1) were selected and exported to a peak table (Supplementary material, Table S2), listing the intervals (Supplementary material, Table S3) and the peak areas (relative concentration) for all exported peaks in each sample, together with the resolved mass spectra. The data matrix used for further analysis only contained the peak areas for all exported peaks (155 columns) for each sample (176 rows).

2.4.3. Partial least squares discriminant analysis (PLS-DA) models

Principal Component Analysis (PCA) was performed on each dataset [fingerprinting (47563 columns × 176 rows) and untargeted profiling through PARADISe (155 columns × 176 rows)] to explore the data and to identify any potential outliers according to the Hotelling's T^2 range and model residuals.

Then, each data matrix (from fingerprinting and from untargeted profiling through PARADISe) was used to calibrate and validate independent PLS-DA classification models with SIMCA v13.0© (Umetrics AB, Sweden) to discriminate between: i) cultivars of Spanish hazelnuts; and ii) geographical origins of TG hazelnuts.

For the cultivar models, a binary PLS-DA model was applied to discriminate TG samples from those of other cultivars (non-TG). To eliminate any potential influences from other factors, only samples from the same origin (Spain) were considered (n = 112; 46 Spanish TG hazelnuts and 66 Spanish non-TG hazelnuts).

For the classification according to the geographical origin of TG hazelnuts, two discrimination approaches were developed, using hazelnut samples from the same cultivar (TG) obtained from different geographical origins (n = 110). The first geographical approach aimed to discriminate between 'European' (EU) (n = 70) and 'non-European' (non-EU: Chile) (n = 40) TG samples; whereas the second model aimed to classify the 70 TG EU samples into their specific country of origin: 'Spain (n = 46) or 'Italy' (n = 24).

For each type of authentication model, each sample set was randomly divided into training (80 % of the samples, TG/non-TG model n = 90, EU/non-EU model n = 88, Spain/Italy model n = 56) and validation set (20 % of the samples, TG/non-TG model n = 22, EU/non-EU model n = 22, Spain/Italy model n = 14). This splitting was run seven times (7 iteration for authentication model type) to evaluate the effect of the sample set composition and to increase the robustness of the external validation.

In PLS-DA binary models, classes are expressed as PLS dummy

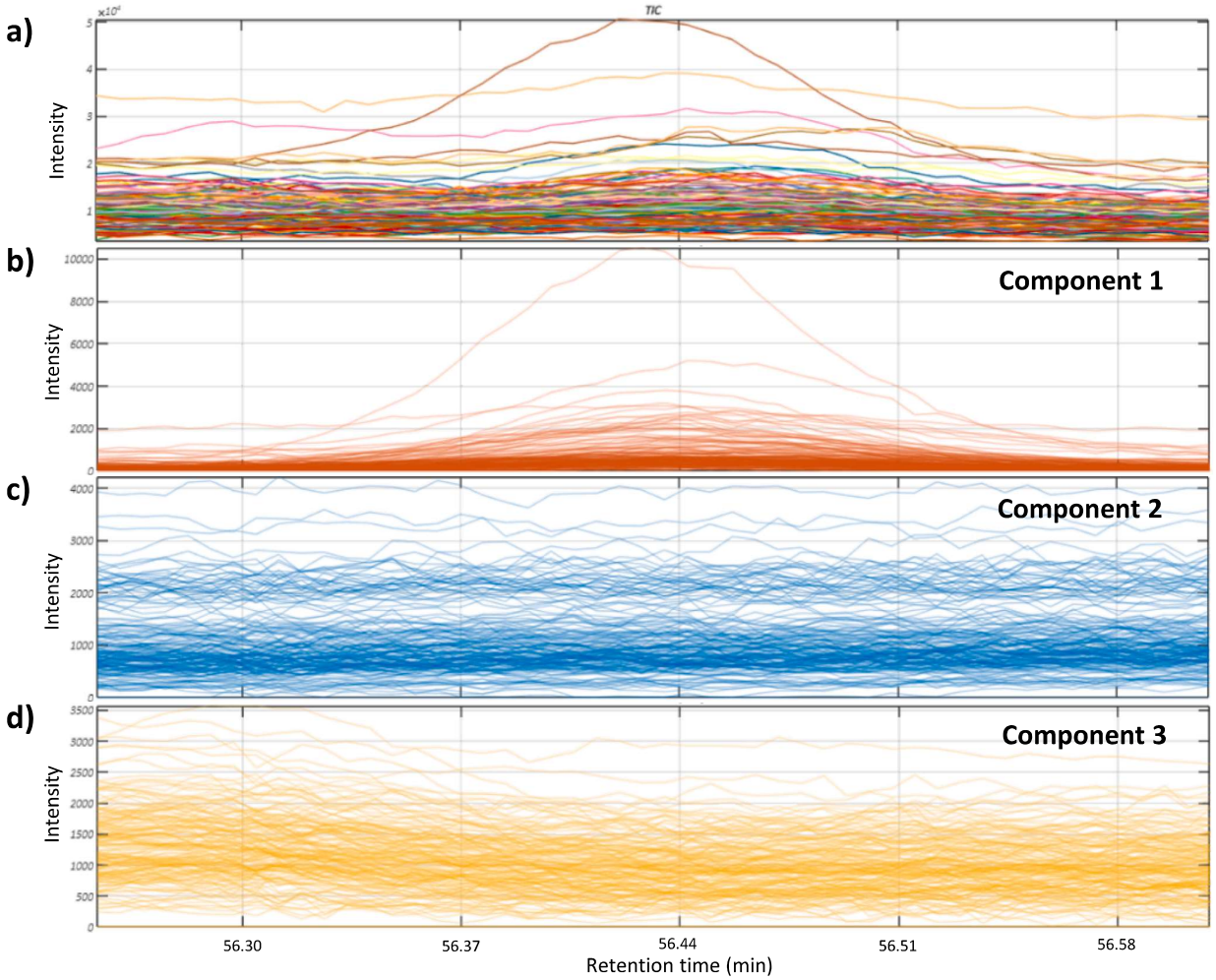


Fig. 1. Plot of the TIC interval between 56.2 and 56.6 min against the PARADISe extracted component plots. a) TIC plot, b) component 1 (orange) corresponding to chemical compound 31, c) component 2 (blue) is baseline noise, d) component 3 (yellow) is baseline noise. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

variables (here, 1 for 'non-TG', 'non-EU' and 'Italy' classes and 0 for 'TG', 'EU' and 'Spain' classes). Then, the PLS predicted value of each sample is used for its classification into one class or the other according to a classification threshold (predicted value = 0.5). In each iteration, models were first internally validated using the training set of samples through leave 10 %-out cross-validation, and the optimal number of latent variables (LV) was selected according to the lowest Root Mean Squared Error of Cross Validation (RMSEcv) criteria. The optimal pre-processing, according to the criteria below, for all the models was mean centring and scaling to the unit of variance. Permutation test (n = 20 permutations) and ANOVA on the cross-validated predictive residuals (p-value) were carried out to assess the models' overfitting (Supplementary material, Table S4). Then, the models were externally validated by predicting the class of the samples in the corresponding validation set, which had not been used to build the models. The suitability of each PLS-DA model was evaluated by the Q^2 values and efficiency, which was expressed as the percentage of correct classification of each class, and the sensitivity (true positives/ [true positives + false negatives]) and specificity (true negatives/ [true negatives + false positives]) values, positive samples being the non-TG, non-EU and Italian samples for the corresponding models. The performance of models from each data processing approach (fingerprint and untargeted profiling through PARADISe) was compared to determine the most suitable one for authentication.

2.4.4. Evaluation of PLS-DA regression coefficients

The regression coefficients of the PLS-DA models developed with all samples in the corresponding sample sets (cultivar n = 112, or origin n = 110) with both the fingerprinting and untargeted profiling approaches were compared to tentatively identify the key variables that contributed to the discrimination between classes. This comparison aimed to reveal the variables that were relevant for both approaches or for only one of the approaches. The jack-knife standard error of cross-validation (SEcv) was used to evaluate the significance of the regression coefficients, considering significant those with values higher than their corresponding SEcv (Torres-Cobos et al., 2021). Out of the significant variables, only the ones with the highest absolute values (25 % higher than the coefficient media) were considered and the corresponding compounds were tentatively identified based on their mass spectra and elution order.

Table 1
External validation of PLS-DA models ('Tonda di Giffoni vs non-Tonda di Giffoni; 'European' vs 'non-European' (Chilean samples) and 'Spanish' vs 'Italian') developed on the fingerprinting and untargeted profiling through PARADISe. Results are mean values (±standard deviation) obtained from seven iterations.

Cultivar model: TG/non-TG						
Fingerprinting (LVs = 7, $Q^2 > 0.64$, RMSEcv < 0.30) [a]						
	n	non-TG (n)	TG (n)	Correct classification (%)	Sensitivity	Specificity
non-TG	13	11.9 ± 0.7	1.1 ± 0.7	91.2 ± 5.3	0.91 ± 0.05	
TG	9	1.0 ± 1.0	8.0 ± 1.0	88.9 ± 11.1		0.89 ± 0.11
Total	22			**90.3 ± 6.7**		
PARADISe (LVs = 6–7, $Q^2 > 0.40$, RMSEcv < 0.36) [a]						
	n	**non-TG (n)**	**TG (n)**	**Correct classification (%)**	**Sensitivity**	**Specificity**
non-TG	13	10.7 ± 1.1	2.3 ± 1.1	82.4 ± 8.6	0.82 ± 0.09	
TG	9	0.9 ± 0.9	8.1 ± 0.9	90.5 ± 10.0		0.91 ± 0.10
Total	22			**85.7 ± 4.9**		
Geographical origin model: EU/non-EU						
Fingerprinting (LVs = 6–7, $Q^2 > 0.72$, RMSEcv < 0.29) [a]						
	n	**non-EU (n)**	**EU (n)**	**Correct classification (%)**	**Sensitivity**	**Specificity**
non-EU	8	8.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	1.00 ± 0.00	
EU	14	0.1 ± 0.4	13.9 ± 0.4	99.0 ± 2.7		0.99 ± 0.03
Total	22			**99.4 ± 1.7**		
PARADISe (LVs = 5–6, $Q^2 > 0.60$, RMSEcv < 0.29) [a]						
	n	**non-EU (n)**	**non-EU (n)**	**Correct classification (%)**	**Sensitivity**	**Specificity**
non-EU	8	7.6 ± 0.8	0.4 ± 0.8	94.6 ± 9.8	0.95 ± 0.10	
EU	14	0.1 ± 0.4	13.9 ± 0.4	99.0 ± 2.7		0.99 ± 0.03
Total	22			**97.4 ± 3.6**		
Geographical origin model: Spanish/Italian						
Fingerprinting (LVs = 5–6, $Q^2 > 0.62$, RMSEcv < 0.31) [a]						
	n	**ITA (n)**	**ESP (n)**	**Correct classification (%)**	**Sensitivity**	**Specificity**
ITA	5	4.6 ± 0.5	0.4 ± 0.5	91.4 ± 10.7	0.91 ± 0.11	
ESP	9	0.3 ± 0.5	8.7 ± 0.5	96.8 ± 5.4		0.97 ± 0.05
Total	14			**94.9 ± 3.5**		
PARADISe (LVs = 4–5, $Q^2 > 0.67$, RMSEcv < 0.28) [a]						
	n	**ITA (n)**	**ESP (n)**	**Correct classification (%)**	**Sensitivity**	**Specificity**
ITA	5	4.9 ± 0.4	0.1 ± 0.4	97.1 ± 7.6	0.97 ± 0.08	
ESP	9	0.3 ± 0.5	8.7 ± 0.5	96.8 ± 5.4		0.97 ± 0.05
Total	14			**96.9 ± 3.8**		

For all models, ANOVA p-value < 0.05. [a] Model parameters: mean values obtained with the training sets from 7 iterations. TG: 'Tonda di Giffoni'; non-TG: other cultivars; EU: European (Spanish and Italian); non-EU: non-European (Chilean); ESP: Spanish hazelnuts; ITA: Italian hazelnuts.

3. Results

3.1. Performance of PLS-DA classification models: fingerprinting vs. untargeted profiling data

All models built on training sets (7 iterations per authentication model type) from both approaches achieved 100 % of correct classification in leave 10 %-out cross validation, which corresponds to the maximum value of sensitivity and specificity (sensitivity and specificity = 1) (Supplementary material, Figure S1) (Supplementary material, Table S4). Subsequently, the PLS-DA models developed on each approach were used to predict the class of the samples conforming the corresponding validation sets. Table 1 presents the mean values of the seven iterations obtained from the external validation of each authentication model type (Cultivar: TG/non-TG; Geographical origin: EU/non-EU and Spain/Italy) developed on fingerprinting and untargeted profiling approaches. No outliers were detected according to the Hotelling's T^2 range and model residuals.

In the case of cultivar authentication, the fingerprinting model outperformed the untargeted profiling one with higher sensitivity (0.91 vs 0.82) and total correct classification percentage (90.3 % vs 85.7 %). Although the specificity of the untargeted profiling model was slightly higher (0.91 vs 0.89), the fingerprinting model performed better overall.

Fingerprinting models were also more efficient in distinguishing between EU and non-EU samples, because even if both approaches achieved the same specificity (0.99), the fingerprinting model exhibited a higher sensitivity (1 vs 0.95) and overall correct classification (99.4 % vs 97.4 %).

Finally, in terms of classification by EU country, both approaches presented the same specificity (0.97) but untargeted profiling models discriminated better the Italian samples, with a higher sensitivity (0.97 vs 0.91, by arbitrarily considering Italian hazelnuts as the positive samples) and overall correct classification (96.9 % vs 94.9 %).

Regarding the standard deviation of mean external validation results calculated on 7 iterations, it ranged from 1.7 to 6.7 % and from 3.6 to 4.9 % in global correct classification by fingerprinting and untargeted profiling models, respectively, being the TG/non-TG and the EU/non-EU the models with the highest and lowest standard deviations, respectively, in both approaches.

3.2. Exploring models through regression coefficient analysis

To study and compare the most informative variables in PLS-DA models developed on the fingerprinting or untargeted profiling (PARADISe) data, we assessed the corresponding regression coefficients. The most relevant coefficients of PLS-DA models based on both approaches corresponded to variables distributed throughout the entire chromatogram. For the fingerprinting approach, they were present in the EICs of all the ions considered (Fig. 2) and several of them corresponded to minor components (Fig. 3).

Although a targeted approach was not the aim of the present study, we tentatively identified some of the most discriminant compounds

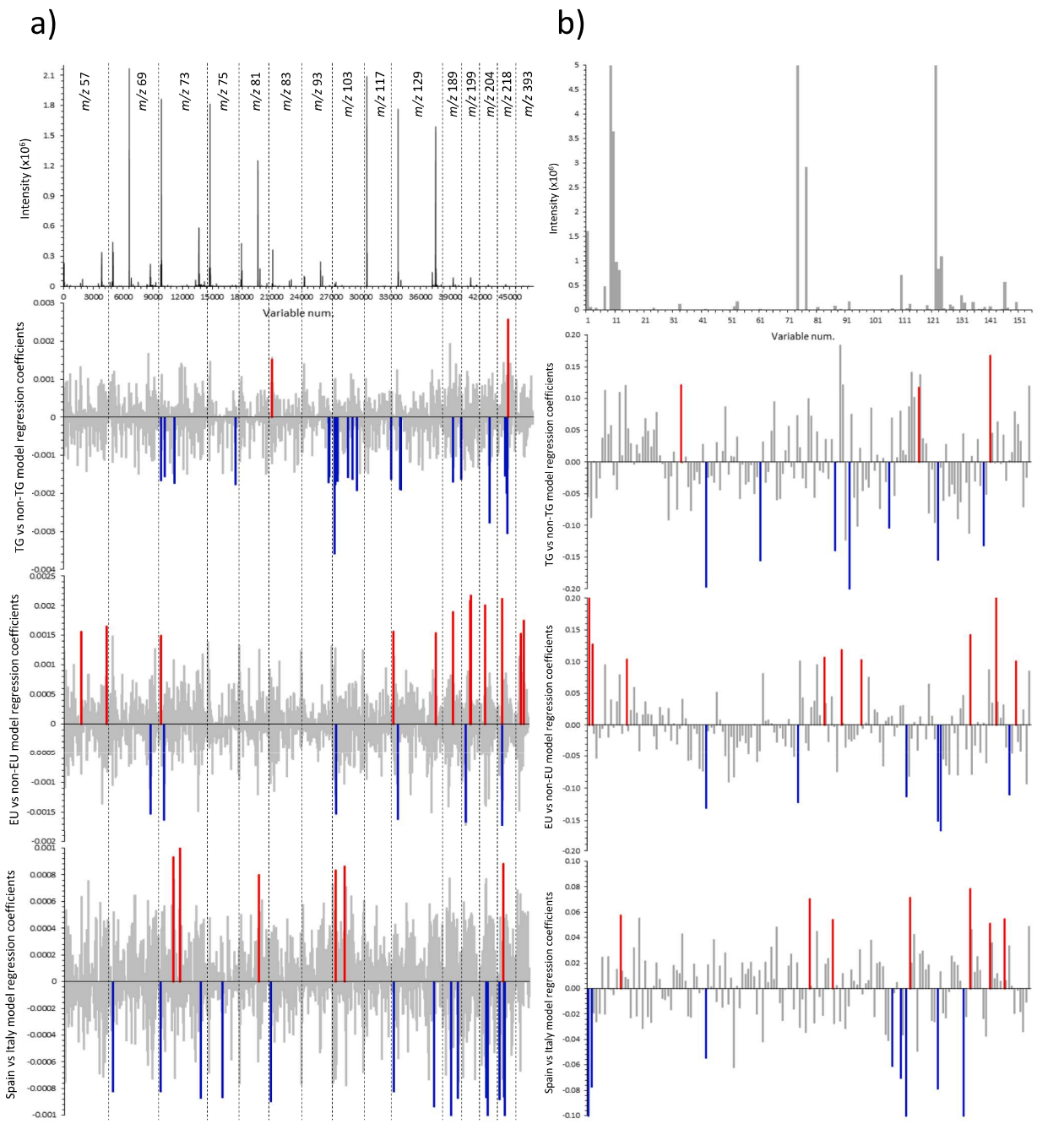
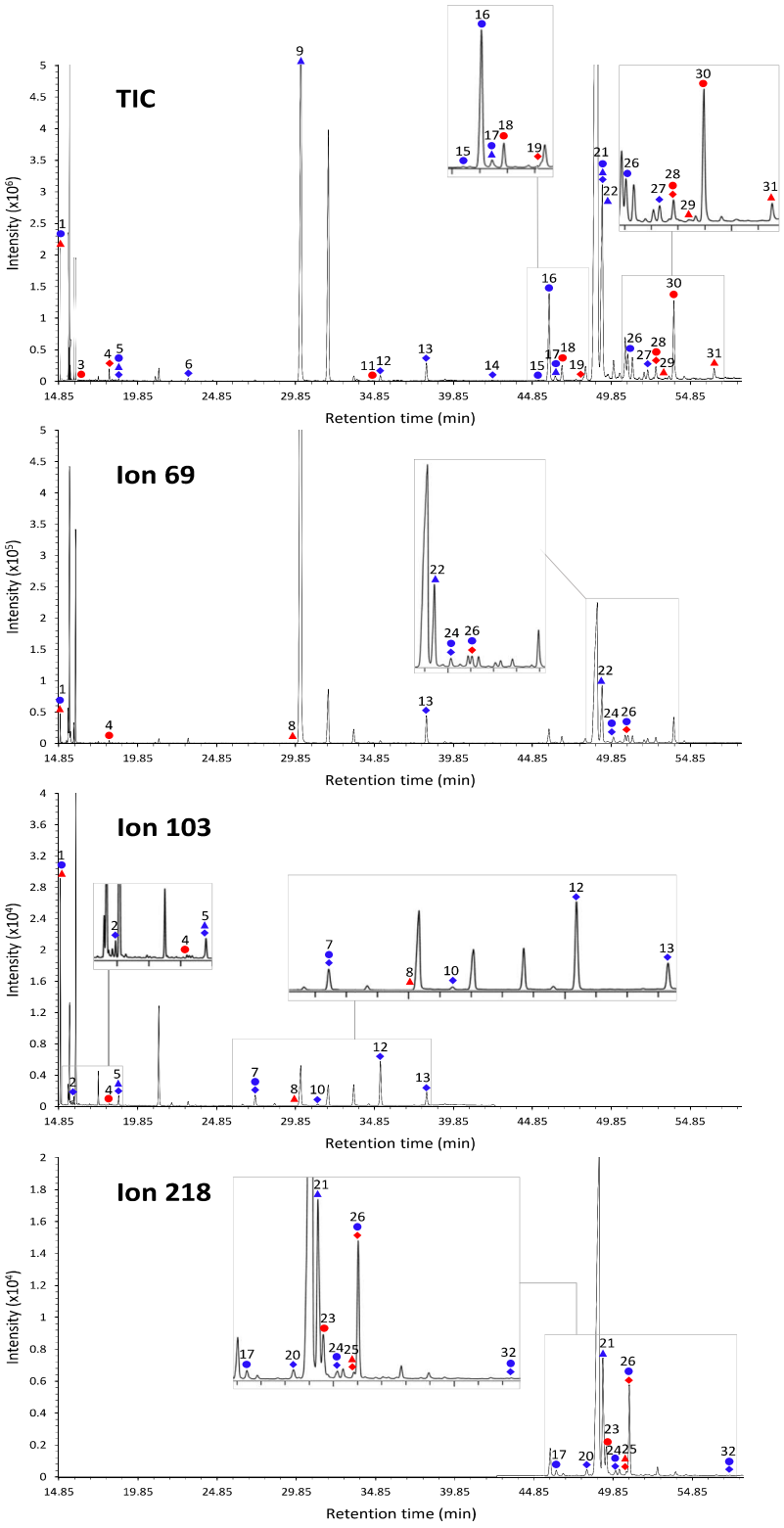


Fig. 2. Regression coefficients of the PLS-DA models ('Tonda di Giffoni -TG vs 'other cultivars' - non-TG; 'European' - EU vs 'non-European' – non-EU and 'Spain' – ESP vs 'Italy' - ITA) developed by a) fingerprinting, plotted against the variables (acquisition points) of concatenated EICs of a Tonda di Giffoni Spanish sample and b) PARADISe, plotted against the variables (detected compounds) of the TIC of the same sample. For each model, the most relevant coefficients for the prediction of the TG, EU and ESP classes are highlighted in blue (negative coefficients) and those relevant for non-TG, non-EU and ITA in red (positive coefficients). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

according to their elution order and the MS spectra (identification levels 2a and 3 according to Schymanski et al., 2014) obtained from full scan MS data (Li et al., 2007; Xu et al., 2018). Specifically, we tentatively identified the variables that were more relevant in classifying the samples within each category of the three authentication models (TG/non-TGs; EU/non-EU; Spain/Italy) (Table 2). The mass spectra of these compounds matched with those of a fatty acid (*FA*), linear (*LA*) and terpene alcohols (*TA*), sterols (*S*), methylsterols (*MS*) and dimethylsterols (*DMS*) previously described in hazelnut oil (Table 2). Additionally, some relevant coefficients corresponded to compounds that could not be linked to any specific structure or chemical family with sufficient confidence and are reported as unknown compounds.

3.2.1. *Cultivar model: TG/non-TG*

Concerning the cultivar model, the regression coefficients with higher absolute value for the TG class in fingerprinting model belonged



(caption on next page)

Fig. 3. Gas chromatograms of the MS response of the unsaponifiable fraction of hazelnuts: Total Ion Chromatogram (TIC) and three representative extracted ions (69, 113, 218). 1) phytol (*TA1*), 2) lineal alcohol C17 (*LA1*), 3) terpene alcohol (*TA2*), 4) unknown (*UNK1*), 5) unknown (*UNK2*), 6) terpene alcohol (*TA3*), 7) lineal alcohol C24 (*LA2*), 8) unknown (*UNK4*), 9) squalene (*TT1*), 10) lineal alcohol C25 (*LA3*), 11) terpene alcohol (*TA4*), 12) lineal alcohol C26 (*LA4*), 13) *cis*-farnesol (*TA5*), 14) lineal alcohol C28 (*LA5*), 15) unknown (*UNK6*), 16) campesterol (*S1*), 17) campestanol (*S2*), 18) stigmasterol (*S3*), 19) sterol (*S4*), 20) unknown (*UNK7*), 21) sitostanol (*S5*), 22) Δ5-avenasterol (*S6*), 23) unknown (*UNK8*), 24) dimethylsterol (*DMS1*), 25) Δ7-stigmastenol (*S7*), 26) lupeol (*DMS2*), 27) 28-methylobtusifoliol (*MS1*), 28) 24-methylenecycloartanol (*DMS3*), 29) dimethyl sterol (*DMS4*), 30) citrostadienol (*MS2*), 31) sterol (*S8*), 32) unknown (*UNK9*). Diamond blue: TG, red: non-TG; Triangle blue: EU, red: non-EU; Circle blue: SPA, red: ITA. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

to the EIC of *m/z* 103, 204 and 218, with the first exhibiting the greatest number of relevant coefficients within this category (Fig. 2). These coefficients may correspond to linear alcohols (*m/z* 103: C17, *LA1*; C25, *LA3;* C26, *LA4*), a terpene alcohol (*m/z* 69, 103: *cis*-farnesol, *TA5*), a not identified dimethylsterol at 50.0 min (*m/z* 218: dimethylsterol 1, *DMS1*) and three unknown compounds eluting at 18.6 min (*m/z* 73, 103 and 129: Unknown 2, *UNK2*), 48.4 min (*m/z* 218: Unknown 7, *UNK7*), and 57.2 min (*m/z* 189: Unknown 9, *UNK9*), respectively (Blue diamonds in EICs from Fig. 3; Table 2). The non-TG class showed the most relevant coefficients in the EIC of *m/z* 83 and 218 (Fig. 2), which could correspond to a fatty acid (*m/z* 83: an oleic acid isomer, *FA1*), a sterol (*m/z* 218; Δ7-stigmastenol, *S7*) and a 4,4-dimethylsterol (*m/z* 218: lupeol, *DMS2*) (Red diamonds in EICs from Fig. 3; Table 2).

In the case of untargeted profiling cultivar model, pure spectra revealed that the most significant coefficients for the TG class corresponded to: the same unknown compound eluting at 18.6 min (Unknown 2, *UNK2*), terpene alcohols (a not identified one eluting at 23.1 min: terpene alcohol 3, *TA3*) and *cis*-farnesol *TA5*), linear alcohols (C26, *LA4*; C28, *LA5*), a sterol (sitostanol, *S5*) and a 4-methylsterol (28-methylobtusifoliol, *MS1*) (Blue diamonds in TIC from Fig. 3, Table 2). Some of these compounds presented *m/z* that had also shown to be relevant for the fingerprint model. For the non-TG class, the most relevant coefficients corresponded to an unknown compound eluting at 18.1 min (Unknown 1, *UNK1*), a not identified sterol eluting at 47.7 min (Sterol, 4 *S4*) and a 4,4-dimethylsterol (24-methylenecycloartanol, *DMS3*) (Red diamonds in TIC from Fig. 3, Table 2).

3.2.2. Geographical origin model: EU/non-EU

For the EU/non-EU model based on fingerprinting data, the most relevant regression coefficients of for the EU category were in the EIC of *m/z* 69, 73, 103, 129, 199 and 218 (Fig. 2), and corresponded to the above-mentioned unknown compound eluting at 18.6 min (*m/z* 73, 103, 129: Unknown 2, *UNK2*), and to sterols (*m/z* 218: sitostanol, *S5*; *m/z* 69, 199: Δ5-avenasterol, *S6*)(Blue triangles in EICs from Fig. 3, Table 2). The most relevant regression coefficients for the non-EU class were distributed in several EIC (Fig. 2). Some of these discriminant variables corresponded to: a terpene alcohol (*m/z* 69, 103, 129: phytol, *TA1*), an unknown specie eluting at 30.0 min (*m/z*: 57: Unknown 4, *UNK4*) a sterol (*m/z* 393: Δ7-stigmastenol, *S7*) and a not identified 4,4-dimethylsterol eluting at 53.3 min (*m/z* 57, 129, 189, 199, 393: dimethylsterol 4; *DMS4*) (Red triangles in EICs from Fig. 3, Table 2).

For the untargeted profiling model, the relevant coefficients for the EU class corresponded to the previously mentioned unknown compound eluting at 18.6 min (Unknown 2, *UNK2*), a triterpene hydrocarbon (squalene, *TT1*) and three sterols (campestanol, *S2*; sitostanol, *S5;* Δ5-avenasterol, *S6*) (Blue triangles in the TIC of Fig. 3, Table 2). The important coefficients for the non-EU class, in this case, corresponded to: a terpene alcohol (phytol, *TA1*), the above-mentioned not identified dimethylsterol 4 (*DMS4*) and a not identified sterol eluting at 56.4 min (Sterol 8, *S8*) (Red triangles in the TIC of Fig. 3, Table 2).

3.2.3. Geographical origin model for TG samples: Spain/Italy

Regarding the discrimination of TG hazelnuts by their origin from Spain or Italy, for the model based on fingerprinting approach the most significant coefficients for the Spain class were detected in EIC of *m/z* 69, 73, 75, 83, 129, 189, 204 and 218 (Fig. 2). They could correspond to: a terpene alcohol (*m/z* 69, 73, 83, 103, 129: phytol, *TA1*), a lineal alcohol (*m/z* 75, 103: C24, *LA2*), a sterol (*m/z* 218: campestanol, *S2*), dimethylsterols (*m/z* 218: not identified dimethylsterol 1, *DMS1*; *m/z* 73, 129, 189, 204, 218: lupeol *DMS2*,) and the previously mentioned unknown compound eluting at 57.2 min (*m/z* 189: Unknown 9, *UNK9*) (Blue circles in the EICs of Fig. 3, Table 2). The most important coefficients for the Italian class corresponded to unidentified species eluting at 18.1 min (*m/z* 103: Unknown 1, *UNK1*), 25.9 min (*m/z* 73: Unknown 3, *UNK3*), 32 min (*m/z* 73, 81: Unknown 5, *UNK5*) and 49.5 min (*m/z* 218: Unknown 8, *UNK8*) (Red circles in the EICs of Fig. 3, Table 2).

Finally, the relevant coefficients for the Spain class in the untargeted profiling model related to: a terpene alcohol (phytol, *TA1*), two unknown compounds, eluting at 18.6 min (Unknown 2, *UNK2*) and at 45.5 min (Unknown 6, *UNK6*), three sterols (campesterol, *S1*; campestanol, *S2*; sitostanol, *S5*) and a 4,4-dimethylsterol (lupeol, *DMS2*) (Blue circles in the TIC of Fig. 3, Table 2). The Italian relevant coefficients corresponded to two not identified terpene alcohols eluting at 15.9 min (Terpene alcohol 2, *TA2*) and 34.8 min (Terpene alcohol 4, *TA4*), a sterol (stigmasterol, *S3*), a 4,4-dimethylsterol (24-methylenecycloartanol, *DMS3*) and a 4-methylsterol (citrostadienol, *MS2*) (Fig. 3, Table 2) (Red circles in the TIC of Fig. 3, Table 2).

4. Discussion

Both untargeted profiling and fingerprinting approaches successfully classified samples according to their cultivar or geographical origin depending on the variable selected for supervising the analysis, achieving percentages of correct classification in external validation higher than 90 % in almost all cases. The results confirm our hypothesis that the unsaponifiable fraction's secondary metabolites, which depend on genetic and environmental factors, have great potential for hazelnut varietal and geographical authentication. Although specificity values were similar for both approaches, fingerprinting outperformed untargeted profiling in two of the three models, providing higher sensitivity and overall correct classification for cultivar and provenance from EU or non-EU areas. This agreed with a previous study on spectroscopic data (Quintanilla-Casas et al., 2022) reporting slightly better prediction results using fingerprinting compared to untargeted profiling approach. On the other hand, the untargeted profiling model demonstrated higher sensitivity in classifying hazelnuts based on their country of origin (Spain or Italy, by arbitrarily considering Italian hazelnuts as the positive samples). In view of these results, we can affirm that both untargeted approaches applied to hazelnut unsaponifiable GC–MS data proved to be highly effective in extracting valuable sample information for the development of efficient authentication models, with the fingerprinting approach achieving slightly higher classification efficiency than untargeted profiling approach.

The standard deviation of the external validation results obtained from the randomly selected sample sets (7 iterations) can provide valuable insights into the dependence of the models on the sample set composition. This metric can be considered as an indicator of the robustness of the models and can help describe their performance in various scenarios. In this regard, the models generated by both approaches exhibited a remarkable low standard deviation, which implies that both showed low dependency on the composition of the validation sample set, indicating a high degree of reliability. Nevertheless, it should be considered that this study was designed to compare these two

Table 2
Tentative identification of compounds (based on the MS spectra from full scan acquisition at the same retention time for the fingerprinting approach and on the MS spectra of compounds extracted by PARADISe from full scan chromatograms for the untargeted profiling) corresponding to the variables with the highest regression coefficients for each class in binary PLS-DA models developed by fingerprinting and PARADISe approaches. The compounds that were relevant in the models developed by both approaches are evidenced in bold.

			TG/non-TG				EU/non-EU				Spanish/Italian			
			Fingerprinting		PARADISe		Fingerprinting		PARADISe		Fingerprinting		PARADISe	
N[a]	Chemical family	Tentative identification and level of annotation[b]	TG	non-TG	TG	non-TG	EU	non-EU	EU	non-EU	ESP	ITA	ESP	ITA
-[c]	Fatty acid 1	NI FA, 3	–	FA1	–	–	–	–	–	–	–	–	–	–
2	Linear alcohol 1	C17, 2a	LA1	–	–	–	–	–	–	–	–	–	–	–
7	Linear alcohol 2	C24, 2a	–	–	–	–	–	–	–	–	LA2	–	–	–
10	Linear alcohol 3	C25, 2a	LA3	–	–	–	–	–	–	–	–	–	–	–
12	Linear alcohol 4	C26, 2a	**LA4**	–	**LA4**	–	–	–	–	–	–	–	–	–
14	Linear alcohol 5	C28, 2a	–	–	LA5	–	–	–	–	–	–	–	–	–
1	Terpene alcohol 1	phytol, 2a	–	–	–	–	–	**TA1**	–	**TA1**	**TA1**	–	**TA1**	–
3	Terpene alcohol 2	NI TA (15.9 min), 3	–	–	–	–	–	–	–	–	–	–	–	TA2
6	Terpene alcohol 3	NI TA (23.1 min), 3	–	–	TA3	–	–	–	–	–	–	–	–	–
11	Terpene alcohol 4	NI TA (34.8 min) 3	–	–	–	–	–	–	–	–	–	–	–	TA4
13	Terpene alcohol 5	*cis*-farnesol, 2a	**TA5**	–	**TA5**	–	–	–	–	–	–	–	–	–
9	Triterpenoid 1	squalene, 2a	–	–	–	–	–	–	TT1	–	–	–	–	–
16	Sterol 1	campesterol, 2a	–	–	–	–	–	–	–	–	–	–	S1	–
17	Sterol 2	campestanol, 2a	–	–	–	–	–	–	S2	–	**S2**	–	**S2**	–
18	Sterol 3	stigmasterol, 2a	–	–	–	–	–	–	–	–	–	–	–	S3
19	Sterol 4	NI S (47.7 min), 3	–	–	–	S4	–	–	–	–	–	–	–	–
21	Sterol 5	sitostanol, 2a	–	–	S5	–	**S5**	–	**S5**	–	–	–	S5	–
22	Sterol 6	Δ5-avenasterol, 2a	–	–	–	–	**S6**	–	**S6**	–	–	–	–	–
25	Sterol 7	Δ7-stigmastenol, 2a	–	S7	–	–	–	S7	–	–	–	–	–	–
31	Sterol 8	NI S (56.4 min), 3	–	–	–	–	–	–	–	S8	–	–	–	–
27	4-methylsterol 1	28-methylobtusifoliol, 2a	–	–	MS1	–	–	–	–	–	–	–	–	–
30	4-methylsterol 2	citrostadienol, 2a	–	–	–	–	–	–	–	–	–	–	–	MS2
24	4,4-dimethylsterol 1	NI DMS (50.0 min), 3	DMS1	–	–	–	–	–	–	–	DMS1	–	–	–
26	4,4-dimethylsterol 2	lupeol, 2a	–	DMS2	–	–	–	–	–	–	**DMS2**	–	**DMS2**	–
28	4,4-dimethylsterol 3	24-methylenecycloartanol, 2a	–	–	–	DMS3	–	–	–	–	–	–	–	DMS3
29	4,4-dimethylsterol 4	NI DMS (53.3 min), 3	–	–	–	–	–	**DMS4**	–	**DMS4**	–	–	–	–
4	Unknown 1	UNK 18.1 min	–	–	–	UNK1	–	–	–	–	–	UNK1	–	–
5	Unknown 2	UNK 18.6 min	**UNK2**	–	**UNK2**	–	**UNK2**	–	**UNK2**	–	–	–	UNK2	–
-[c]	Unknown 3	UNK 25.9 min	–	–	–	–	–	–	–	–	–	UNK3	–	–
8	Unknown 4	UNK 30.0 min	–	–	–	–	–	UNK4	–	–	–	–	–	–
-[c]	Unknown 5	UNK 32.0 min	–	–	–	–	–	–	–	–	–	UNK5	–	–
15	Unknown 6	UNK 45.5 min	–	–	–	–	–	–	–	–	–	–	UNK6	–
20	Unknown 7	UNK 48.4 min	UNK7	–	–	–	–	–	–	–	–	–	–	–
23	Unknown 8	UNK 49.5 min	–	–	–	–	–	–	–	–	–	UNK8	–	–
32	Unknown 9	UNK 57.2 min	UNK9	–	–	–	–	–	–	–	UNK9	–	–	–

[a] : Compound code according to Fig. 3. TG: Tonda di Giffoni class; non-TG: other cultivars class; EU: European class (Spanish and Italian hazelnuts); non-EU: non-EU class (Chilean hazelnuts); ESP: Spanish hazelnuts class; ITA: Italian hazelnuts class; NI: not identified; UNK: unknown compound.

[b] : tentative molecular structure identification and level of annotation according to Schymanski et al., 2014 (2a: probable structure by library spectrum match; 3: tentative candidate, evidence exists for possible structure, but insufficient information for one exact structure only).

[c] : Compounds not shown in Fig. 3.

approaches, and to preliminarily evaluate the usefulness of the unsaponifiable fraction for these authentication purposes. Therefore, the sampling set included a limited number of regions and cultivars, which implies that these models are not representative of the real hazelnut production, and therefore their applicability is limited to this specific purpose. Further sample collecting including a wider natural variability (i.e. other main producing countries) is needed to develop models that can be applied in a real scenario.

Examining the regression coefficients of models generated by both fingerprinting and untargeted profiling it was evident that the distribution of most relevant compounds for the classification was throughout the entire chromatogram. This, combined with the fact that several of these compounds were present in low concentrations (Fig. 3), highlights the necessity for methods that enable comprehensive utilization of sample information such as the untargeted approaches evaluated in the present study. The tentative identification of the compounds corresponding to the most relevant variables, provided insights into the chemical families that played a crucial role in the classification process. This analysis also enabled to determine whether there were any differences in the key discriminant compounds according to the untargeted approach applied. It is worth clarifying that our intention was not to provide an exhaustive exploration of the discriminant variables, but to focus on some the most relevant variables to acquire an understanding of the type of compounds on which the classification was based on. Regarding the chemical families that mainly drove the classification in models obtained from both untargeted approaches, steroid compounds tentatively identified as sterols, 4-methylsterols, and 4,4-dimethylsterols; linear and terpene alcohols; and some unknown compounds were found to be the key discriminators, with the steroid compounds playing a crucial role in classification (Table 2). Previous studies demonstrated the influence of both genetic (Parcerisa et al., 1998; Amaral et al., 2006; Matthäus & Özcan, 2012) and environmental factors (Benitez-Sanchez et al., 2003; Matthäus & Özcan, 2012; Ghisoni et al., 2020) on the steroid fraction of hazelnuts, which supports the present findings.

The comparison of the key discriminant compounds between fingerprinting and untargeted profiling authentication models revealed partial agreement in relevant variables (Table 2). Compounds tentatively identified as unknown compound 2 (*UNK2*), C26 linear alcohol (*LA4*) and *cis*-farnesol (*TA5*) were significant to classify TG samples for both approaches. Likewise, unknown compound 2 (*UNK2*), sitostanol (*S5*) and Δ5-avenasterol (*S6*) were found to be relevant in discriminating EU samples, while phytol (*TA1*) and the not identified dimethylsterol 4 (*DMS4*) were characteristic of the non-EU samples. In addition, phytol (*TA1*), along with campestanol (*S2*) and lupeol (*DMS2*), were useful in discriminating Spanish from the Italian samples in both approaches.

However, in addition to the matching discriminant markers, it is worth noting that each approach identified distinct relevant variables in each of the authentication models (Table 2). This can be attributed to the fact that the information provided by the unfolded matrix-based fingerprinting and untargeted profiling approach varies in terms of both quantity and type, due to their differing mode of operation. On the one hand, the higher sensitivity of SIM acquisition in unfolded matrix-based fingerprinting can detect even minor compounds that may significantly contribute to sample categorization, that may be overlooked by untargeted profiling's full scan acquisition. For instance, minor compounds like the not identified minor fatty acid (*FA1*), linear alcohols C24 (*LA2*) and C25 (*LA3*), the not identified dimethylsterol 1 (*DMS1*) and Δ7-stigmastenol (*S7*) were found to be significant for classification in fingerprinting models but were not detected as chromatographic peaks by full scan untargeted profiling.

On the other hand, the selection of specific ions for acquisition in SIM mode might hinder the detection of other significant compounds characterized by different ions, which can, in turn, be detected by untargeted profiling in full scan mode. Nevertheless, in this case, compounds found as relevant only in untargeted profiling models (Table 2), such as those tentatively identified as linear alcohol C28 (*LA5*), terpene alcohols *TA2*, *TA3* and *TA4*, campesterol (*S1*), stigmasterol (*S3*), sterols *S4* and *S8,* 28-methylobtusifoliol (*MS1*), 24-methylenecycloartanol (*DMS3*) and citrostadienol (*MS2*), as well as some unknown compounds (*UNK6*), were also detected by SIM and thus, they were included in the unfolded matrix. However, they resulted to be less relevant for the classification in fingerprinting model compared to other minor compounds.

Therefore, if the representative ions of the chemical families being analysed are selected properly, the information obtained from the fingerprinting method using SIM acquisition appears to be greater than the information contained in the untargeted profiling matrix based on full scan acquisition. However, this assumption requires a general prior knowledge of the chemical families of compounds present in the samples, which is satisfied in the case of the unsaponifiable fraction of hazelnut but may present a challenge in other authentication scenarios. In this sense, one of the main advantages of untargeted profiling is its ability to provide chemically interpretable results, making it suitable for analysing samples with unknown compositions and allowing for easy identification of the markers of interest. It represents a straightforward way to identify the most relevant compounds as the pure mass spectra are provided, unlike SIM fingerprinting that does not allow for clear identification and requires further full scan analysis to properly assess compounds' mass spectra and chemical structure.

One final consideration that should be addressed concerns the applicability, ease of implementation and level of prior knowledge required by the user, and transferability for each of the untargeted approaches compared. PARADISe is a user-friendly interface to utilize PARAFAC2, but it does require a certain level of know-how for interval selection and optimization of PARAFAC2 models, which is not necessary for building the unfolded matrix in the fingerprinting approach. This issue may be resolved in future versions of PARADISe by enabling automatic interval selection, but at present, the fingerprinting unfolded matrix approach is easier to use and apply. On the other hand, transferring untargeted analytical methods to other laboratories or instruments can be a challenging task, especially for fingerprinting methods. In fact, while conventional strategies for target methods can be adapted to assess the performance of untargeted profiling results, thereby enabling easy in-house and inter-laboratory validations, a lack of precise guidelines regarding the validation procedure for fingerprinting methods make it even more challenging to transfer these methods, despite ongoing efforts to establish them (Quintanilla-Casas et al., 2020b).

5. Conclusions

In conclusion, the unsaponifiable fraction of the hazelnut oil has proven to be a promising tool for their geographical and varietal authentication. Even if it is not a fast-screening technique, the study has proved that GC–MS coupled with untargeted methods such as fingerprinting and advanced profiling techniques like untargeted profiling can provide high-dimensional molecular-level information for hazelnut authentication. Both untargeted profiling and fingerprinting proved to be successful in the authentication of hazelnuts, although fingerprinting provided slightly better prediction results. As revealed by the examination of the regression coefficients of the PLS-DA models, this may be due to the greater information extracted by the fingerprinting method from chromatographic data, which enabled considering even very minor discriminant species. However, untargeted profiling enables easier chemical interpretability than fingerprinting based on SIM data, providing the pure spectra of the relevant compounds. It is remarkable that these results were obtained in a challenging scenario in which the origin was discriminated between samples of the same cultivar, and in turn, the cultivar was discriminated between samples from the same origin. This positions the analytical strategy as a suitable candidate to verify challenging samples as a support to fast-screening tools. Nevertheless, optimal models should be further developed and evaluated using a large-scale dataset, that would include the natural heterogeneity

of the samples, the main producing regions and their principal cultivars in addition to several harvest years.

Funding

This work was developed in the context of the project TRACENUTS, PID2020-117701RB100 financed by MCIN/AEI/https://doi.org/10.13039/501100011033. B. Torres-Cobos thanks the Spanish Ministry of Universities predoctoral fellowships FPU20/014540. B. Quintanilla-Casas thanks the Fundación Alfonso Martín Escudero for the research grant for universities and centers abroad 2022. A. Tres received a Ramon y Cajal grant (RYC-2017-23601) funded by MCIN/AEI/https://doi.org/10.13039/501100011033 and by "ESF Investing in your future".

CRediT authorship contribution statement

B. Torres-Cobos: Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft. **B. Quintanilla-Casas:** Data curation, Investigation, Methodology, Validation, Writing – review & editing. **M. Rovira:** Conceptualization, Resources, Writing – review & editing. **A. Romero:** Conceptualization, Resources, Writing – review & editing. **F. Guardiola:** Supervision, Writing – review & editing. **S. Vichi:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **A. Tres:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of competing 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

Data will be made available on request.

Acknowledgements

INSA-UB Maria de Maeztu Unit of Excellence (Grant CEX2021-001234-M) funded by MICIN/AEI/FEDER, UE. The authors would like to express their gratitude to Ferrero Hazelnut Company and Tuscia University (Department of Agriculture and Forest Science) for providing the hazelnut samples from Chile and Italy, respectively.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2023.138294.

References

Amaral, J. S., Casal, S., Citova, I., Santos, A., Seabra, R. M., & Oliveira, B. P. P. (2006). Characterization of several hazelnut (Corylus avellana L.) cultivars based in chemical, fatty acid and sterol composition. *European Food Research and Technology, 222*, 274–280. https://doi-org.sire.ub.edu/10.1007/s00217-005-0068-0.

Baccolo, G., Quintanilla-Casas, B., Vichi, S., Augustijn, D., & Bro, R. (2021). From untargeted chemical profiling to peak tables – A fully automated AI driven approach to untargeted GC-MS. *TrAC Trends in Analytical Chemistry, 145*, Article 116451. https://doi.org/10.1016/j.trac.2021.116451

Bachmann, R., Klockmann, S., Haerdter, J., Fischer, M., & Hackl, T. (2018). 1H NMR Spectroscopy for Determination of the Geographical Origin of Hazelnuts. *Journal of Agricultural and Food Chemistry, 66*, 11873–11879. https://doi.org/10.1021/acs.jafc.8b03724

Ballin, N. Z., & Laursen, K. H. (2019). To target or not to target? Definitions and nomenclature for targeted versus non-targeted analytical food authentication. *Trends in Food Science & Technology, 86*, 537–543. https://doi.org/10.1016/j.tifs.2018.09.025

Benitez-Sanchez, P. L., León-Camacho, M. L., & Aparicio, R. (2003). A comprehensive study of hazelnut oil composition with comparisons to other vegetable oils, particularly olive oil. *European Food Research and Technology, 218*, 13–19. https://doi.org/10.1007/s00217-003-0766-4

Biancolillo, A., De Luca, S., Bassi, S., Roudier, L., Bucci, R., Magrì, A. D., & Marini, F. (2018). Authentication of an Italian PDO hazelnut ("Nocciola Romana") by NIR spectroscopy. *Environmental Science and Pollution Research, 25*, 28780–28786. https://doi.org/10.1007/s11356-018-1755-2

Bosque-Sendra, J. M., Cuadros-Rodríguez, L., Ruiz-Samblás, C., & de la Mata, A. P. (2012). Combining chromatography and chemometrics for the characterization and authentication of fats and oils from triacylglycerol compositional data—A review. *Analytica Chimica Acta, 724*, 1–11. https://doi.org/10.1016/j.aca.2012.02.041

Ciarmiello, L. F., Mazzeo, M. F., Minasi, P., Peluso, A., De Luca, A., Piccirillo, P., Siciliano, R. A., & Carbone, V. (2014). Analysis of Different European Hazelnut (Corylus avellana L.) Cultivars: Authentication, Phenotypic Features, and Phenolic Profiles. *Journal of Agricultural and Food Chemistry, 62*, 6236–6246. https://doi.org/10.1021/jf5018324

FAOstat Food and Agriculture Organization of the United Nations, agricultural data 2021; https://www.fao.org/faostat/es/#data/PP (accessed Oct 15, 2023).

Ghisoni, S., Lucini, L., Rocchetti, G., Chiodelli, G., Farinelli, D., Tombesi, S., & Trevisan, M. (2020). Untargeted metabolomics with multivariate analysis to discriminate hazelnut (Corylus avellana L.) cultivars and their geographical origin. *Journal of the Science of Food and Agriculture, 100*, 500–508. https://doi.org/10.1002/jsfa.9998

Goriainov, S. V., Esparza, C. A., Borisova, A. R., Orlova, S. V., Vandyshev, V. V., Hajjar, F., Platonov, E. A., Chromchenkova, E. P., Novikov, O. O., Borisov, R. S., & Kalabin, G. A. (2021). Phytochemical Study of the Composition of the Unsaponifiable Fraction of Various Vegetable Oils by Gas Chromatography-Mass Spectrometry. *Journal of Analytical Chemistry, 76*, 1635–1644. https://doi.org/10.1134/S1061934821140045

Inaudi, P., Giacomino, A., Malandrino, M., La Gioia, C., Conca, E., Karak, T., & Abollino, O. (2020). The Inorganic Component as a Possible Marker for Quality and for Authentication of the Hazelnut's Origin. *International Journal of Environmental Research and Public Health, 17*, 447. https://doi.org/10.3390/ijerph17020447

Johnsen, L. G., Skou, P. B., Khakimov, B., & Bro, R. (2017). Gas chromatography – mass spectrometry data processing made easy. *Journal of Chromatography A, 1503*, 57–64. https://doi.org/10.1016/j.chroma.2017.04.052

Klockmann, S., Reiner, E., Bachmann, R., Hackl, T., & Fischer, M. (2016). Food Fingerprinting: Metabolomic Approaches for Geographical Origin Discrimination of Hazelnuts (Corylus avellana) by UPLC-QTOF-MS. *Journal of Agricultural and Food Chemistry, 64*, 9253–9262. https://doi.org/10.1021/acs.jafc.6b04433

Krauß, S., Vieweg, A., & Vetter, W. (2019). Stable isotope signatures (δ2H-, δ13C-, δ15N-values) of walnuts (Juglans regia L.) from different regions in Germany. *Journal of the Science of Food and Agriculture, 100*, 1625–1634. https://doi.org/10.1002/jsfa.10174

Król, K., & Gantner, M. (2020). Morphological Traits and Chemical Composition of Hazelnut from Different Geographical Origins: A Review. *Agriculture, 10*, 375. https://doi.org/10.3390/agriculture10090375

Lang, C., Weber, N., Möller, M., Schramm, L., Schelm, S., Kohlbacher, O., & Fischer, M. (2021). Genetic authentication: Differentiation of hazelnut cultivars using polymorphic sites of the chloroplast genome. *Food Control, 130*, Article 108344. https://doi.org/10.1016/j.foodcont.2021.108344

Larsen, F. H., Van den Berg, F., & Engelsen, S. B. (2006). An exploratory chemometric study of 1H NMR spectra of table wines. *Journal of Chemometrics, 20*, 198–208. https://doi.org/10.1002/cem.991

Li, T. S. C., Beveridge, T. H. J., & Drover, J. C. G. (2007). Phytosterol content of sea buckthorn (Hippophae rhamnoides L.) seed oil: Extraction and identification. *Food Chemistry, 101*, 1633–1639. https://doi.org/10.1016/j.foodchem.2006.04.033

Matthäus, B., & Özcan, M. M. (2012). The comparison of properties of the oil and kernels of various hazelnuts from Germany and Turkey. *European Journal of Lipid Science and Technology, 114*, 801–806. https://doi.org/10.1002/ejlt.201100299

Nielsen, N. P. V., Carstensen, J. M., & Smedsgaard, J. (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A, 805*, 17–35. https://doi.org/10.1016/S0021-9673(98)00021-1

Oddone, M., Aceto, M., Baldizzone, M., Musso, D., & Osella, D. (2009). Authentication and Traceability Study of Hazelnuts from Piedmont, Italy. *Journal of Agricultural and Food Chemistry, 57*, 3404–3408. https://doi.org/10.1021/jf900312p

Parcerisa, J., Richardson, D. G., Rafecas, M., Codony, R., & Boatella, J. (1998). Fatty acid, tocopherol and sterol content of some hazelnut varieties (Corylus avellana L.) harvested in Oregon (USA). *Journal of Chromatography A, 805*, 259–268. https://doi.org/10.1016/S0021-9673(98)00049-1

Phillips, K. M., Ruggio, D. M., & Ashraf-Khorassani, M. (2005). Phytosterol Composition of Nuts and Seeds Commonly Consumed in the United States. *Journal of Agricultural and Food Chemistry, 53*, 9436–9445. https://doi.org/10.1021/jf051505h

Quintanilla-Casas, B., Bro, R., Hinrich, J. L., Davie-Martin, C. L. (2023). Tutorial on PARADISe: PARAFAC2-based Deconvolution and Identification System for processing GC–MS data, PROTOCOL (Version 1), Protocol Exchange.

Quintanilla-Casas, B., Bertin, S., Leik, K., Bustamante, J., Guardiola, F., Valli, E., … Vichi, S. (2020a). Profiling versus fingerprinting analysis of sesquiterpene hydrocarbons for the geographical authentication of extra virgin olive oils. *Food Chemistry, 307*, Article 125556. https://doi.org/10.1016/j.foodchem.2019.125556

Quintanilla-Casas, B., Marin, M., Guardiola, F., García-González, D. L., Barbieri, S., Bendini, A., … Tres, A. (2020b). Supporting the Sensory Panel to Grade Virgin Olive Oils: An In-House-Validated Screening Tool by Volatile Fingerprinting and Chemometrics. *Foods, 9*, 1509. https://doi.org/10.3390/foods9101509

Quintanilla-Casas, B., Rinnan, Å., Romero, A., Guardiola, F., Tres, A., Vichi, S., & Bro, R. (2022). Using fluorescence excitation-emission matrices to predict bitterness and

pungency of virgin olive oil: A feasibility study. *Food Chemistry, 395*, Article 133602. https://doi.org/10.1016/j.foodchem.2022.133602

Riedl, J., Esslinger, S., & Fauhl-Hassek, C. (2015). Review of validation and reporting of non-targeted fingerprinting approaches for food authentication. *Analytica Chimica Acta, 885*, 17–32. https://doi.org/10.1016/j.aca.2015.06.003

Rinnan, A., Amigo, J. M., Skov, T. (2014). Multiway methods in food science, in: D. Granato, G. Ares (Eds.), Mathematical and Statistical Methods in Food Science and Technology, IFT Press/Wiley Blackwell, Chichester 143-174.

Ríos-Reina, R., Aparicio-Ruiz, R., Morales, M. T., & García-Gonzalez, D. L. (2023). Contribution of specific volatile markers to green and ripe fruity attributes in extra virgin olive oils studied with three analytical methods. *Food Chemistry, 399*, Article 133942. https://doi.org/10.1016/j.foodchem.2022.133942

Sales, C., Portolés, T., Johnsen, L. G., Danielsen, M., & Beltran, J. (2019). Olive oil quality classification and measurement of its organoleptic attributes by untargeted GC–MS and multivariate statistical-based approach. *Food Chemistry, 271*, 488–496. https://doi.org/10.1016/j.foodchem.2018.07.200

Sammarco, G., Dall`Asta, C., & Suman, M. (2023). Near infrared spectroscopy and multivariate statistical analysis as rapid tools for the geographical origin assessment of Italian hazelnuts. *Vibrational Spectroscopy, 126*, Article 103531. https://doi.org/10.1016/j.vibspec.2023.103531

Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H. P., & Hollender, J. (2014). Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environmental Science & Technology, 48*, 2097–2098. https://doi.org/10.1021/es5002105

Tomasi, G., Van den Berg, F., & Andersson, C. (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics, 18*, 231–241. https://doi.org/10.1002/cem.859

Torres-Cobos, B., Quintanilla-Casas, B., Romero, A., Ninot, A., Alonso-Salces, R. M., Toschi, T. G., Bendini, A., Guardiola, F., Tres, A., & Vichi, S. (2021). Varietal authentication of virgin olive oil: Proving the efficiency of sesquiterpene fingerprinting for Mediterranean Arbequina oils. *Food Control, 128*, Article 108200. https://doi.org/10.1016/j.foodcont.2021.108200

Torres-Cobos, B., Quintanilla-Casas, B., Vicario, G., Guardiola, F., Tres, A., & Vichi, S. (2023). Revealing adulterated olive oils by triacylglycerol screening methods: Beyond the official method. *Food Chemistry, 409*, Article 135256. https://doi.org/10.1016/j.foodchem.2022.135256

Tüfekci, F., & Karataş, Ş. (2018). Determination of geographical origin Turkish hazelnuts according to fatty acid composition. *Food Science & Nutrition, 6*, 557–562. https://doi.org/10.1002/fsn3.595

Xu, B., Zhang, L., Ma, F., Zhang, W., Wang, X., Zhang, Q., Luo, D., Ma, H., & Li, P. (2018). Determination of free steroidal compounds in vegetable oils by comprehensive two-dimensional gas chromatography coupled to time-of-flight mass spectrometry. *Food Chemistry, 245*, 415–425. https://doi.org/10.1016/j.foodchem.2017.10.114

Zannella, C., Carucci, F., Aversano, R., Prohaska, T., Vingiani, S., Carputo, D., & Adamo, P. (2017). Genetic and geochemical signatures to prevent frauds and counterfeit of high-quality asparagus and pistachio. *Food Chemistry, 237*, 545–552. https://doi.org/10.1016/j.foodchem.2017.05.158

5.4 Enhancement, validation and comparison of metabolic methods for hazelnut geographical and varietal authentication

5.4.1 Triacylglycerol and Unsaponifiable fraction fingerprinting

Drawing on the results obtained from the PoC studies, fingerprinting emerged as the most suitable approach for developing authentication models based on the UF of hazelnuts. It outperformed the untargeted profiling method, achieving higher classification accuracy in authenticating both cultivar and provenance (Publication 7). Therefore, subsequent method enhancement and validation was conducted using this approach, with an expanded sample set that included a greater number of samples and encompassed a wider variability. (Publication 8)

On the other hand, the PoC study demonstrated that HT-GC-MS fingerprinting is an effective fast-screening tool for analysing TAGs, outperforming other tested methods (Publication 6). Consequently, it could be selected as a simpler and faster alternative to UF fingerprinting or as a complementary method, enhancing analytical efficiency by enabling the analysis of numerous samples.

To develop, evaluate and validate both methods under a higher natural variability, they were tested on the same extensive sample set (n = 309) that included hazelnuts from various harvest years, regions, and cultivars. This testing allowed for a reliable assessment of their performances.

This extensive sample set included hazelnuts collected from 2019 to 2022 of:

- the same TG cultivar (n = 207) from different countries of origin: CHL (n = 40,), ESP (n = 91,), Georgia (GEO, n = 40,) and ITA (ITA, n = 36); these samples had been also used to develop the isotope-based models (section 5.2.1).
- different cultivars from the same origin, Spain (n = 102).

Separate PLS-DA classification models were built and externally validated, using the UF and TAG fingerprinting data (**Figure 13**):

1) Cultivar model: a binary PLS-DA including only samples from Spain, to discriminate TG hazelnuts from other cultivars (non-TG).

2) Geographical origin model: A multi-class PLS-DA model, including only TG samples, to distinguish between the four countries of origin (CHL, ESP, GEO, ITA).

Finally, a combined strategy that integrates the rapidity of the TAG method and the efficiency UF method (**Figure 13**), was tested and evaluated with the purpose of achieving optimal discrimination while reducing the analytical workload (Publication 8).

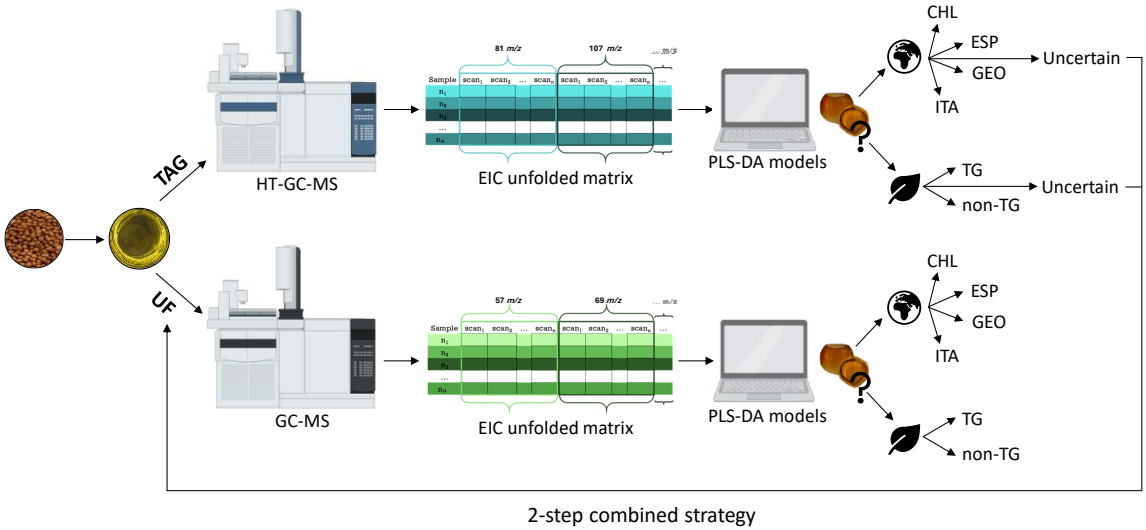


Figure 13. Graphical abstract of Publication 8. UF: unsaponifiable fraction, TAG: triacylglycerol, HT-GC-MS: high-temperature–gas chromatography–mass spectrometry, GC-MS: gas chromatography–mass spectrometry, EIC: extracted ion chromatogram, PLS-DA: partial least square-discriminant analysis, CHL: Chile, ESP: Spain, GEO: Georgia, ITA: Italy, TG: Tonda di Giffoni, non-TG: other cultivars.

5.4.2 Publication 8

Meeting the challenge of varietal and geographical authentication of hazelnuts through lipid metabolite fingerprinting

Berta Torres-Cobos, Soriana Beatrice Nicotra, Mercè Rovira, Agustí Romero, Francesc Guardiola, Alba Tres, Stefania Vichi

Food Chemistry, ***2025****, 463(2), 141203.*
https://doi.org/10.1016/j.foodchem.2024.141203

Supplementary material available in **Annex 7**

Food Chemistry 463 (2025) 141203

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem

Meeting the challenge of varietal and geographical authentication of hazelnuts through lipid metabolite fingerprinting

B. Torres-Cobos [a,b], S.B. Nicotra [a], M. Rovira [c], A. Romero [c], F. Guardiola [a,b], A. Tres [a,b,*], S. Vichi [a,b]

[a] Departament de Nutrició, Ciències de L'Alimentació i Gastronomia, Facultat de Farmàcia i Ciències de L'Alimentació, Universitat de Barcelona. Av Prat de La Riba, 171, 08921 Santa Coloma de Gramenet, Spain
[b] Institut de Recerca en Nutrició i Seguretat Alimentària (INSA-UB), Universitat de Barcelona. Av Prat de La Riba, 171, 08921 Santa Coloma de Gramenet, Spain
[c] Institute of Agrifood Research and Technology (IRTA), Ctra. de Reus – El Morell Km 3.8, Constantí 43120, Spain

ARTICLE INFO

Original content: Hazelnut triacylglycerol fingerprints obtained by concatenating GC-MS Extracted Ion Chromatograms (Original data)
Hazelnut unsaponifiable fraction fingerprints obtained by concatenating GC-MS Extracted Ion Chromatograms (Original data)

Keywords:
Hazelnut
Provenance and cultivar authentication
Unsaponifiable fraction
Fingerprinting
Triacylglycerols
PLSDA

ABSTRACT

Hazelnuts are high-quality products with significant economic importance in many European countries. Their market price depends on their qualitative characteristics, which are driven by cultivar and geographical origin, making hazelnuts susceptible to fraud. This study systematically compared two lipidomic fingerprinting strategies for the simultaneous authentication of hazelnut cultivar and provenance, based on the analysis of the unsaponifiable fraction (UF) and triacylglycerol (TAG) profiles by gas chromatography–mass spectrometry coupled with chemometrics. PLS-DA classification models were developed using a large sample set with high natural variability (n = 309) to discriminate hazelnuts by cultivar and origin. External validation results demonstrated the suitability of the UF fingerprint as a hazelnut authentication tool, both tested models showing a high efficiency (>94 %). The correct classification rate of the TAG fingerprinting method was lower (>80 %), but due to its faster analysis time, it is recommended as a complementary screening tool to UF fingerprinting.

1. Introduction

In recent years, the demand for high-quality food products has significantly increased, with typicity being a key quality trait. As the characteristics of plant-based foodstuffs are greatly influenced by environmental and genetic factors, consumers are showing a growing preference for typical food products that reliably indicate their geographical and varietal origin. These aspects are the cornerstone of denominations of origin such as the Protected Denomination of Origin and Protected Geographical Indication. These quality schemes stipulate that products included in these categories must be produced within defined geographic areas and use only specified cultivars (Regulation (EU) 2024/1143).

Hazelnut production is of significant economic importance in countries such as Turkey, Italy, the United States, Chile, Georgia, and Spain, among others (Food and Agriculture Organization of the United Nations, 2022). These nuts are highly valued for their sensory and compositional characteristics, which are closely associated with the cultivar and geographical origin (Amaral et al., 2006; Król & Gantner, 2020), leading to significant differences in the market price (Food and Agriculture Organization of the United Nations, 2022). Each hazelnut cultivar possesses distinctive technological traits including nut yield, shape, calibre, and pellicle removal, as well as unique kernel flavour and aroma. These inherent characteristics guide the industry's cultivar selection process to align with specific intended applications. The current market is dominated by round-shaped Italian cultivars such as 'Tonda Gentile', 'Tonda Romana' and 'Tonda di Giffoni' (TG), as well as Spanish cultivars like 'Negret' and 'Pauetet', along with Turkish cultivars such as 'Tombul', 'Palaz', and 'Fosa' (Silvestri et al., 2021). While most of these cultivars are primarily grown in their traditional countries of origin, certain

* Corresponding author at: Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Campus de l'Alimentació de Torribera, Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona, Av Prat de la Riba, 171. 08921 Santa Coloma de Gramenet, Spain.
E-mail address: atres@ub.edu (A. Tres).

https://doi.org/10.1016/j.foodchem.2024.141203
Received 26 May 2024; Received in revised form 3 September 2024; Accepted 6 September 2024
Available online 10 September 2024

cultivars, such as TG, have also gained popularity in new producer countries.

Differences in quality and market price among hazelnuts from distinct cultivars or regions, along with concerns about the safety of non-traceable foods, highlight the importance of authenticating hazelnuts based on their geographical and varietal origins. Developing reliable instrumental methods is crucial for assessing the authenticity of these characteristics, which are currently verified through documentation and phenotypic observations. To address this challenge, various strategies have been proposed. DNA-based methods provide accurate varietal authentication (Lang et al., 2021), while markers associated with the growing area, such as elemental composition and isotope ratios, contribute to reliable geographical authentication (Inaudi et al., 2020; Krauß et al., 2019; Oddone et al., 2009). However, these methods are often expensive and complex, and cannot simultaneously determine both varietal and geographical origin. Given their substantial influence on hazelnut traits and market value, being able to identify both origin and cultivar in a single analysis is crucial to counter the risk of economically motivated fraud. Metabolomics has great potential in this context, as plant metabolic profiles are strongly shaped by genetic traits and environmental conditions. These profiles can be effectively correlated with specific cultivars and geographical regions, allowing for robust authentication. Metabolomic studies have focused on distinguishing either the cultivar or the provenance of hazelnuts (Ciarmiello et al., 2014; Klockmann et al., 2016), but few attempts have been made to simultaneously authenticate their botanical and geographical origin (Ghisoni et al., 2020). In this regard, the unsaponifiable fraction (UF) of hazelnuts is a lipidic fraction that contains numerous families of secondary metabolites, including sterols, linear and terpene alcohols, and hydrocarbons (Benitez-Sanchez et al., 2003; Goriainov et al., 2021). The presence and profile of these metabolites is known to be driven by both genetic (Amaral et al., 2006; Matthäus & Özcan, 2012; Parcerisa et al., 1998) and environmental factors (Benitez-Sanchez et al., 2003; Ghisoni et al., 2020; Matthäus & Özcan, 2012), and is minimally influenced by aspects such as storage or processing. This positions the UF as a promising candidate for fingerprinting to verify the botanical and geographical origin of hazelnuts.

Among non-targeted metabolomic approaches, metabolic fingerprinting involves analysing high-dimensional data such as spectra or chromatograms to identify distinct patterns or fingerprints exclusive to a food sample with specific attributes (Ballin & Laursen, 2019; Bosque-Sendra et al., 2012). This approach has proven highly effective in authenticating both the varietal and geographical origin of food products (Quintanilla-Casas et al., 2022; Torres-Cobos, Quintanilla-Casas, Guardiola, et al., 2021; Torres-Cobos, Quintanilla-Casas, Romero, et al., 2021). Specifically, in a preliminary study on hazelnuts with a limited sample set, fingerprint analysis of UF data successfully classified hazelnut cultivar and provenance, achieving an accuracy rate of over 90 % in external validation (Torres-Cobos et al., 2024). Therefore, the analysis of secondary metabolites obtained through extraction and saponification of the lipid fraction, and subsequently determined by gas chromatography–mass spectrometry (GC–MS), has emerged as a promising and effective strategy for hazelnut authentication. However, to validate the results and ensure method reliability, it is crucial to further develop and evaluate authentication models using representative large-scale datasets.

Analytical efficiency can also be enhanced by exploring alternative screening methods based on the same fingerprinting approach but focusing on metabolic fractions that require simpler analytical procedures. For instance, variations in the profile of triacylglycerols (TAGs), the main constituents of the hazelnut lipid fraction, have been related to both genetic and environmental conditions (Amaral et al., 2006; Parcerisa et al., 1994). Moreover, TAG analysis is fast, simple, and does not require cumbersome sample preparation, making it suitable for routine analysis or as a screening method. Specifically, TAG fingerprinting by high temperature-gas chromatography–mass spectrometry (HT-GC–MS) has been recently described as more efficient than other analytical approaches in authenticating lipid-based foods (Torres-Cobos et al., 2023). However, despite its potential, only a few studies have explored the TAG fraction of hazelnuts to develop authentication tools for discriminating the cultivar or provenance. To the best of our knowledge, to date only Kiralan et al. (2015) have used the TAG profile coupled with chemometrics to develop hazelnut geographical classification models, and no studies have applied this method to distinguish between hazelnut cultivars. Therefore, our hypothesis is that TAG fingerprinting is a simpler method that can complement the UF fingerprinting approach for hazelnut botanical and geographical authentication.

For a reliable assessment of UF and TAG fingerprinting performance and their suitability for hazelnut authentication, both methods must be tested on a large sample set with high natural variability, including hazelnuts from different cultivars, regions, and harvest years.

The primary objectives of this study were to i) demonstrate the applicability of UF fingerprinting by GC–MS for hazelnut varietal and geographical authentication using a large-scale set of over 300 samples collected across different harvest seasons; ii) apply the same strategy and sample set to assess the viability of TAG fingerprinting by HT-GC–MS as a faster alternative method to UF fingerprinting, and iii) evaluate the effectiveness of an analytical strategy that combines the rapidity of the TAG method and the efficiency of the UF method, aiming to achieve optimal discrimination while minimizing analytical workload.

2. Material and methods

2.1. Sampling

A total of 309 traceable hazelnut samples were obtained from 2019 to 2022 as part of the TRACENUTS project (PID2020-117701RB-I00), directly from producers. From these samples, 207 were of the 'Tonda di Giffoni' (TG) cultivar, sourced from Chile (CHL, $n = 40$, 2019–2020), Spain (ESP, $n = 91$, 2019–2022), Georgia (GEO, n = 40, 2021–2022), and Italy (ITA, $n = 36$, 2019–2021). The remaining 102 samples were from seven different cultivars (non-TG) produced in Spain. Information about the sample set is summarized in Table S1 of Supplementary Information. Samples were stored vacuum-packed at 4 °C until analysis.

2.2. Unsaponifiable fraction fingerprinting

2.2.1. Material and reagents

Diethyl ether stabilized with 7 mg/L of BHT, anhydrous sodium sulphate, and anhydrous pyridine 99.5 % were purchased from Scharlau (Sentmenat, Spain). Potassium hydroxide 85 % for analysis in pellets was purchased from Thermo Scientific (Waltham, Massachusetts, USA). Methanol for GC-ECD and FID, and Horning's silylating mixture II (*N*,*O*-bis(trimethylsilyl)acetamide/chlorotrimethylsilane/1-(trimethylsilyl) imidazole, 3:2:3, v/v/v) were purchased from Merck (Darmstadt, Germany). Amberlite IRN78 OH hydroxide form was obtained from Supelco (Bellefonte, Pennsylvania, USA).

2.2.2. Sample preparation

Sample preparation was carried out as reported by Torres-Cobos et al. (2024). Briefly, about 30 g of hazelnuts were ground, and the lipid fraction was extracted with diethyl ether. The organic solvent was evaporated and a 1 g aliquot of the hazelnut oil was saponified by adding 2 M methanolic potassium hydroxide solution and heating for 30 min at 70 °C in a water bath. The reaction was quenched with ice and the UF was extracted with diethyl ether. The organic extract was purified by washing with distilled water and adding amberlite adsorbent to remove the excess dissolved free fatty acids. Anhydrous sodium sulphate was added to remove any remaining moisture. The solvent was evaporated, and the UF was reconstituted with 50 µL of pyridine. Finally, 100 µL of silylating reagent was added and allowed to react for 20 min prior to

injection.

2.2.3. Unsaponifiable fraction fingerprinting by selected ion monitoring (SIM) GC–MS

Samples were analysed under the conditions reported by Torres-Cobos et al. (2024). The UF fingerprint was acquired using an Agilent 6890 N Network GC system equipped with a Combi-PAL autosampler (CTC Analytics, Zwingen, Switzerland) and coupled to an Agilent 5975C Inert MSD quadrupolar mass selective analyser (Agilent Technologies, Santa Clara, California, USA). Helium was the carrier gas at a flow rate of 1.5 mL/min. Analytes were separated on a ZB-5 MS capillary column (60 m × 0.25 mm i.d., 0.25 μm film thickness) from Phenomenex (Torrance, California, USA). Column temperature was held at 150 °C for 2 min, increased to 260 °C at 10 °C/min, held for 2 min, and then increased to 320 °C at 2 °C/min, holding the last temperature for 13 min. The temperatures of the ion source and the transfer line were 230 and 300 °C, respectively. Mass spectra were recorded at 1.9 scan/s and the electron energy was set at 70 eV. Data acquisition was performed in the SIM mode by analysing the extracted ion chromatogram (EIC) of 15 characteristic ions for several compound families of the UF: *m/z* 57, 69, 73, 75, 81, 83, 93, 103, 117, 129, 189, 199, 204, 218 and 393 (Torres-Cobos et al., 2024). Ions of *m/z* 75, 81, 83, 93, 103 and 117 were acquired from 14.85 to 42.5 min (3171 scans for each EIC); *m/z* 57, 69, 73 and 129 from 14.85 to 58 min (4903 scans per EIC); and *m/z* 189, 199, 204, 218 and 393 from 42.5 to 58 min (1785 scans per EIC). These acquisition times were selected based on the retention time ranges of the main UF classes of compounds (Torres-Cobos et al., 2024). Then, a fingerprinting approach was followed using the intensities of the scans of the EICs. For each ion, a data matrix was built with all samples (rows, $n = 309$) and the scan intensities of each EIC as variables (columns) (15 different data matrices with 3171 scans × 6 ions, 4903 scans × 4 ions, or 1785 scans × 5 ions). Differences between injections were corrected by normalizing each EIC. The EICs of each ion matrix were then aligned using the correlation optimized warping (COW) algorithm in Matlab® (Nielsen et al., 1998) to correct retention time shifts among samples. Finally, the 15 aligned EIC matrices were concatenated into a two-way unfolded matrix (309 samples × 47,563 variables).

2.3. TAG fingerprinting

2.3.1. Material and reagents

Dichloromethane (SupraSolv® for GC-ECD/FID) was purchased from Merck (Darmstadt, Germany).

2.3.2. Sample preparation

Hazelnut oil was extracted as described in section *2.2.2 Sample preparation.* Then, a 30 mg aliquot of oil sample was dissolved in 1.5 mL of dichloromethane to a final concentration of 2 % (*w/v*).

2.3.3. TAG fingerprinting by SIM HT-GC–MS

The TAG fingerprint of hazelnut samples was analysed under the conditions reported by Torres-Cobos et al. (2023). HT-GC–MS analysis was performed on an Agilent Technologies 6890 N gas chromatograph coupled to an Agilent 5973 Network quadrupolar mass selective analyser (Agilent Technologies, Santa Clara, California, USA). The sample injection volume was 2 μL with a split ratio of 1:20. Helium was the carrier gas at a flow rate of 1.5 mL/min. Analytes were separated on a TRB-50HT column (Teknokroma, Sant Cugat del Vallès, Spain) (30 × 0.25 mm I.D., 0.10 μm). The initial column temperature was 290 °C, which was held for 2 min and then increased to 315 °C at 10 °C /min. Finally, it was increased to 350 °C at 2 °C /min and held for 20 min. The ion source and transfer line temperatures were set at 230 and 350 °C, respectively. Mass spectra were acquired in SIM mode at 1.3 scan/s and the electron energy was set at 70 eV. A fingerprinting approach was followed by acquiring the EICs of 16 TAG fragment ions: *m/z* 81, 107, 237, 239, 260, 262, 264, 265, 295, 311, 313, 323, 335, 337, 339, and 351 (Torres-Cobos et al., 2023) from 1.6 to 38.6 min (2920 scans × 16 ions = 46,720 variables). For each ion, a data matrix was built with the scan intensities of each EIC (columns) for all samples (rows) (309 samples × 46,720 variables). Baseline correction was performed using asymmetric least squares (Eilers, 2004) followed by normalization of each EIC. The EICs of each ion matrix were aligned using the COW algorithm in Matlab® (Nielsen et al., 1998). The 16 aligned matrices were then concatenated to obtain a two-way unfolded matrix.

2.4. Chemometrics

2.4.1. One analysis approach: UF and TAG partial least squares discriminant analysis (PLS-DA) models

Firstly, to explore the data and identify potential outliers according to Hotelling's T^2 range and model residuals, principal component analysis (PCA) was performed on both the UF and TAG fingerprinting datasets. Then, the data matrices of each method were separately used to calibrate and validate individual PLS-DA classification models using SIMCA v13.0© (Umetrics AB, Sweden). For each method (UF or TAG fingerprinting), two types of classification models were developed: i) a TG/non-TG binary cultivar model, which included only Spanish samples, to discriminate between TG samples and those of other cultivars (non-TG); and ii) a multi-class geographical origin model, which included only TG samples, to distinguish between four different countries of origin.

For each authentication model (TG/non-TG or origin model) and method (UF or TAG fingerprinting), the sample set was randomly split into training (80 % of samples from each category: TG/non-TG model, $n = 154$; origin model, $n = 166$) and validation sets (20 % of samples from each category: TG/non-TG model, $n = 39$; origin model, $n = 41$). This splitting process was run seven times (7 iterations) to evaluate the effect of the sample set composition and increase the robustness of the external validation. The sample set splitting information, including test and training sets, is summarized in Table S1 of the Supplementary Information.

In each iteration, a PLS-DA model (training model) was calibrated and internally validated through leave-10 %-out cross-validation using the samples in the training set of that iteration. In each case, the chosen data pre-processing and the number of latent variables (LV) for each PLS-DA model were those that achieved the lowest value of root mean squared error based on predicted values in cross validation (RMSEcv) (Quintanilla-Casas, Bustamante, et al., 2020; Torres-Cobos et al., 2024; Xu et al., 2018). To assess the model overfitting, permutation tests ($n = 20$ permutations) and ANOVA (p-value) on the cross-validated predictive residuals compared to those of a random model were performed as in previous studies (Eriksson et al., 2008; Ghisoni et al., 2020; Quintanilla-Casas et al., 2022). Then, each training model was externally validated by predicting the class of samples in the corresponding validation set, which had not been used to build the models. Therefore, for each type of model, 7 training PLS-DA models and the corresponding 7 external validations were obtained, to verify that results were not driven by specific influential samples and thus, to increase the robustness of the external validation (Torres-Cobos, Quintanilla-Casas, Romero, et al., 2021; Westerhuis et al., 2008). Results of each individual iteration can be found in in Tables S2 and S3 of the Supplementary Information. Unless otherwise indicated, results in the manuscript will be the mean of these 7 iterations.

In PLS-DA binary models, classes are expressed as PLS dummy variables (1 for non-TG class and 0 for TG). Whereas multi-class PLS-DA models work as multiple binary models of each class against the other samples, a dummy Y matrix holds as many classification vectors as classes and each vector has values of 1 for one class and 0 for all the other classes (1 for CHL, ESP, GEO, and ITA classes; and 0 for non-CHL, non-ESP, non-GEO, and non-ITA classes). The PLS predicted value (PV) of each sample was used for classification into one class or the other according to the highest PV criteria, provided it was above the

classification threshold. Samples with PVs below the classification threshold for all classes were not assigned to any class.

To maximize model performance, classification thresholds were optimized by analysing receiver operating characteristics (ROC) generated with the PVs obtained in the leave-10 %-out cross-validation. The ROC curve plots the sensitivity (true positives/ [true positives + false negatives]) and 1-specificity (1 - [true negatives/ (true negatives + false positives)]) values obtained when the PV threshold to assign samples to a diagnostic category (TG or non-TG for the cultivar model; and CHL, ESP, GEO, and ITA for the origin model) varies (Fawcett, 2006). In this case, the positive classes were non-TG, CHL, ESP, GEO, and ITA for the corresponding models. ROC analysis was applied to PV values from each individual PLS-DA model. Hence, a total of 35 ROC curves (7 random training sets per 5 diagnostic categories [1 for the cultivar and 4 for origin models]) were generated for each method (TAG and UF fingerprinting). Classification of the one-analysis approach models was based on thresholds optimized through the ROC curves; these optimal thresholds were those leading to the maximum value of sensitivity + specificity (Quintanilla-Casas, Marin, et al., 2020). Optimal thresholds are summarized in Table S4 of the Supplementary Information.

The performance of each PLS-DA model was evaluated using the cumulative Q^2 values and the percentage of correct classification. Sensitivity and specificity were assessed for the binary cultivar models. For the origin models, the number of unassigned samples (those with PVs below the thresholds for all diagnostic categories) was also evaluated.

The results of both approaches (UF and TAG fingerprinting) were compared and evaluated to determine their suitability for hazelnut authentication.

2.4.2. Combined TAG and UF fingerprinting strategy

To improve classification performance and enhance analytical efficiency, a combined strategy was applied following a 2-step classification approach as shown in Fig. 1. First, samples were assessed by the TAG fingerprinting method, and an uncertainty range for the classification was established instead of a single threshold. Uncertainty ranges were set according to the sensitivity and specificity obtained from ROC curves for each diagnostic category. Thus, the PVs that resulted in sensitivity and specificity equal to 1 were used as the lower and upper thresholds of the uncertainty range, respectively (Quintanilla-Casas, Marin, et al., 2020). If the same PV provided specificity and sensitivity values equal to 1, the upper and lower thresholds were set to ±0.15 of that PV (Quintanilla-Casas, Bustamante, et al., 2020) (Table S4 of Supplementary Information). Samples with PVs falling within the uncertainty range, or those not assigned to any class, were considered as boundary samples.

These boundary samples were further assessed by the UF fingerprinting method, applying the optimal threshold described in section 2.4.1 as the classification criterion. The combined strategy was also internally validated through leave-10 %-out cross-validation with each of the seven training sets and externally validated by predicting the class of the samples in the corresponding validation sets.

3. Results and discussion

The initial exploration of the UF and TAG fingerprinting data sets by PCA showed no outliers according to Hotelling's T^2 range and model residuals.

For all PLS-DA models, the optimal pre-processing, according to the lowest RMSEcv, was mean-centring and scaling to the unit of variance. Permutation tests and ANOVA *p*-values indicated that neither the origin nor the cultivar models were overfitted.

3.1. Performance of the PLS-DA models for the one-analysis approach

3.1.1. Unsaponifiable fraction fingerprinting

In the leave-10 %-out cross-validation (internal validation), all cultivar and origin models generated from the 7 training sets (7 iterations) achieved 100 % correct classification (Tables S2, S3, S5 and S6 of Supplementary Information). However, to confirm their suitability, these models were also externally validated using the training models to predict the samples of the corresponding validation sets. The results, expressed as mean values ± standard deviation from the 7 iterations, are summarized in Table 1 and Table 2.

For the binary cultivar model (Table 1), correct classification values for both TG and non-TG classes were equal to or higher than 91.2 %. The TG class showed slightly higher prediction accuracy (97.6 %), likely due to the greater variability of cultivars in the non-TG class, which challenged accurate prediction. This model displayed high mean sensitivity and specificity and achieved a mean global correct classification rate of 94.1 %.

These results indicate that the efficiency of UF fingerprinting in distinguishing hazelnut cultivars surpassed that of genetic methods (Giulia et al., 2022; Lang et al., 2021) and imaging-based methods for

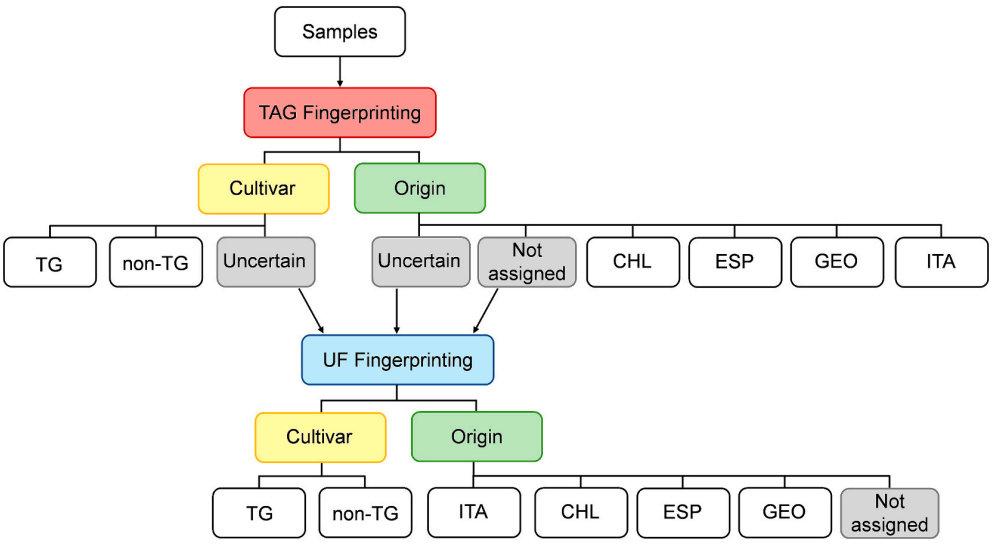


Fig. 1. Diagram of the 2-step combined strategy based on two consecutive analyses (TAG and UF fingerprinting) to predict hazelnut origin and cultivar.

Table 1
External validation of PLS-DA cultivar models developed independently on unsaponifiable fraction (UF) and triacylglycerol (TAG) fingerprinting data. Results are mean values (± standard deviation) obtained from seven iterations.

Cultivar model: TG/non-TG						
UF fingerprinting (LVs = 8, Q^2 = 0.82, RMSEcv = 0.21) [a]						
	n	non-TG (n)	TG (n)	Correct classification (%)	Sensitivity	Specificity
non-TG	21	19.1 ± 1.2	1.9 ± 1.2	91.2 ± 5.8	0.91 ± 0.06	
TG	18	0.6 ± 0.8	17.6 ± 0.8	97.6 ± 4.4		0.98 ± 0.04
Total	39			94.1 ± 4.1		
TAG fingerprinting (LVs = 9–10, Q^2 = 0.43, RMSEcv = 0.37) [a]						
	n	non-TG (n)	TG (n)	Correct classification (%)	Sensitivity	Specificity
non-TG	21	17.7 ± 1.6	3.3 ± 1.6	84.4 ± 7.6	0.84 ± 0.08	
TG	18	2.6 ± 1.6	15.4 ± 1.6	85.7 ± 9.0		0.86 ± 0.09
Total	39			85.0 ± 6.4		

For all models, ANOVA p-value < 0.05. [a] Model parameters: range of number of latent variables (LVs), and mean values (Q^2, RMSEcv) obtained with the training sets from 7 iterations. TG: 'Tonda di Giffoni'; non-TG: other cultivars.

whole shelled and unshelled hazelnuts (Keles & Taner, 2021; Menesatti et al., 2008). A direct comparison with other authentication methods based on hazelnut chemical composition is not always possible because many studies have used them to differentiate between groups of cultivars (Moscetti et al., 2015; Stella et al., 2013), rather than distinguishing a single hazelnut cultivar from others produced in the same region, as in the present study. Additionally, it is important to emphasize that UF fingerprinting successfully distinguished TG from other Italian cultivars within the non-TG class, such as 'San Giovanni' (non-TG5 samples, see Table S1 of Supplementary Information), 'Tonda Romana' (non-TG7 samples), and the hybrid 'Tonda Francescana' (non-TG6 samples), which is a cross between 'Tonda Romana' and TG (Farinelli et al., 2019).

External validation of the multi-class origin model (Table 2) also resulted in high correct classification percentages for all classes, reaching a global correct classification rate of 97.2 %. Notably, all hazelnuts from Georgia were correctly classified, while one Chilean and one Spanish sample were misclassified by one of the training models: the former was not assigned to any class and the latter was wrongly classified as Italian. Among the Italian samples, some were unassigned, and only two were misclassified as Chilean and Spanish, respectively. The greater challenge in assigning Italian samples likely stemmed from the similarity of environmental and climatic conditions to those of Spain. Again, the performance of this method cannot be directly compared to those of other methods because they have not been tested in the same challenging scenario. This study examined the same cultivar produced in different geographical areas, so the markers were influenced only by environmental and not genetic factors. Despite this, the method's efficiency was equal to or greater than in studies where hazelnut classes differed in both origin and cultivar (Bachmann et al., 2018; Biancolillo et al., 2018; Klockmann et al., 2016; Sammarco et al., 2023; Shakiba et al., 2022).

3.1.2. TAG fingerprinting

Leave-10 %-out cross-validation of TAG models resulted in 100 % and 99.5 % of correct classification for the cultivar (Table S2 and S5 of Supplementary Information) and origin models (Table S3 and S6 of Supplementary Information), respectively. These promising results were further evaluated through external validation (Table 1, Table 2 and Tables S3 and S4 of Supplementary Information).

Although external validation showed a slightly lower efficiency compared to the internal validation, an overall correct classification of 85 % was achieved for the cultivar model, with mean specificity and sensitivity values equal to or higher than 0.84 (Table 1). The model exhibited slightly more difficulty in classifying the non-TG class, as found with the UF cultivar model.

On the other hand, the origin TAG model (Table 2) successfully classified 82.2 % of the hazelnuts, with an accuracy exceeding 83 % for the Chilean, Georgian, and Spanish samples. Accuracy for the Italian samples was lower (61.2 %), with a significant number misclassified as Spanish or left unassigned. This behaviour, observed to a lesser extent in the UF data model, likely results from the similarity of environmental conditions in Spain and Italy, which appears to have a more pronounced effect on the TAG fingerprint. The classification accuracy of TAG fingerprinting is comparable to other high-resolution techniques that

Table 2
External validation of PLS-DA origin models developed independently on unsaponifiable fraction (UF) and triacylglycerol (TAG) fingerprinting data. Results are mean values (± standard deviation) obtained from seven iterations.

Geographical origin model: CHL/ESP/GEO/ITA							
UF fingerprinting (LVs = 11–13, Q^2 = 0.78, RMSEcv = 0.28) [a]							
	n	CHL (n)	ESP (n)	GEO (n)	ITA (n)	Not assigned (n)	Correct classification (%)
CHL	8	7.9 ± 0.4	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.1 ± 0.4	98.2 ± 4.7
ESP	18	0.0 ± 0.0	17.9 ± 0.4	0.0 ± 0.0	0.1 ± 0.4	0.0 ± 0.0	99.2 ± 2.1
GEO	8	0.0 ± 0.0	0.0 ± 0.0	8.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0
ITA	7	0.1 ± 0.4	0.1 ± 0.4	0.0 ± 0.0	6.1 ± 1.1	0.6 ± 1.0	87.8 ± 15.3
Total	41						97.2 ± 2.6
TAG fingerprinting (LVs = 11–12, Q^2 = 0.41, RMSEcv = 0.33) [a]							
	n	CHL (n)	ESP (n)	GEO (n)	ITA (n)	Not assigned (n)	Correct classification (%)
CHL	8	7.4 ± 0.8	0.3 ± 0.5	0.0 ± 0.0	0.0 ± 0.0	0.3 ± 0.8	92.9 ± 9.8
ESP	18	0.0 ± 0.0	15.0 ± 3.3	0.0 ± 0.0	1.0 ± 0.8	1.9 ± 3.2	83.3 ± 18.1
GEO	8	0.0 ± 0.0	0.3 ± 0.5	7.0 ± 0.8	0.1 ± 0.4	0.6 ± 0.8	87.5 ± 10.2
ITA	7	0.1 ± 0.4	1.7 ± 1.1	0.1 ± 0.4	4.3 ± 1.0	0.7 ± 0.8	61.2 ± 13.6
Total	41						82.2 ± 8.9

For all models, ANOVA *p*-value < 0.05. [a] Model parameters: range of number of latent variables (LVs), and mean values (Q^2, RMSEcv) obtained with the training sets from 7 iterations. CHL: Chile; ESP: Spain; GEO: Georgia; ITA: Italy.

involve more laborious and complex sample preparation (Klockmann et al., 2016) and outperforms other fast GC-based methods (Sammarco et al., 2023). Despite low efficiency in classifying Italian samples, the analytical simplicity and overall good performance of TAG fingerprinting support its suitability as a screening method, especially when used in combination with a confirmation method.

3.2. Comparison between UF and TAG fingerprinting methods

The results presented above demonstrate the efficacy of both UF and TAG fingerprinting for hazelnut authentication, due to the use of a large sample set of hazelnuts from different harvest years, cultivars, and geographical origins for the modelling and validation. Additionally, the reliability of the models was confirmed through septuplicate external validation. The low standard deviation observed in the external validation results for both UF and TAG models indicates minimal dependence on the composition of the validation sample set, affirming the robustness of both approaches.

These findings corroborate that the UF and TAG components are driven by both environmental and genetic factors and evidence their suitability as hazelnut authentication methods.

Comparing the results of the two methods revealed that UF fingerprinting outperformed TAG fingerprinting in both cultivar (global accuracy: UF: 94.1 %; TAG: 85.0 %) and origin (global accuracy: UF: 97.2 %; TAG: 82.2 %) models across all classes (Tables 1 and 2). These findings demonstrate that, despite requiring a more laborious analytical procedure, UF fingerprinting displays unmatched effectiveness in authenticating hazelnut cultivars (Giulia et al., 2022; Keles & Taner, 2021; Lang et al., 2021; Menesatti et al., 2008; Moscetti et al., 2015; Stella et al., 2013) and origin (Bachmann et al., 2018; Biancolillo et al., 2018; Klockmann et al., 2016; Sammarco et al., 2023), confirming our preliminary findings (Torres-Cobos et al., 2024).

On the other hand, TAG fingerprinting stands out as a straightforward method that requires minimal analytical expertise and is easily applicable to a large number of samples, yielding a global accuracy >80 % in external validation. It could therefore be considered as a suitable candidate for hazelnut authentication screening, especially when combined with a confirmatory tool such as UF fingerprinting method to resolve boundary or non-classified samples.

3.3. Combined strategy (TAG + UF fingerprinting)

To reduce the analytical workload while preserving high classification accuracy, we implemented a strategy combining TAG and UF methods in a 2-step classification procedure (Fig. 1). In the first step, samples underwent rapid screening by TAG fingerprinting, establishing an uncertainty range. Subsequently, samples falling within the uncertainty range or those not assigned to any class underwent further evaluation using the UF fingerprinting method.

In terms of cultivar authentication, the 2-step approach significantly improved the classification performance of the TAG fingerprinting method, achieving 90.8 % global correct classification in the external validation, with sensitivity and specificity values ≥0.89 (Table 3). Furthermore, the analytical efficiency of the combined method increased considerably compared to UF fingerprinting alone. In fact, most samples were directly classified by the TAG model, with only a quarter falling within the uncertainty range and requiring UF analysis.

Regarding geographical authentication, the classification performance of the combined strategy was significantly improved compared to the TAG fingerprinting method (Table 4). External validation provided 93.0 % overall correct classification, a notable increase from the 82.2 % achieved by the TAG model. The accuracy for individual classes was ≥95 % for Chilean, Georgian, and Spanish hazelnuts. Additionally, the classification of Italian samples experimented a substantial improvement, increasing from 61.2 % to 79.6 %. This approach also reduced the experimental workload and time, as less than a third of the samples were boundary (fell within the uncertainty range of the TAG method or were unassigned by that method) and required analysis by the UF fingerprinting method.

These outcomes confirm our hypothesis that the 2-step strategy integrating TAG and UF fingerprinting would enable the varietal and geographical authentication of hazelnuts, achieving overall accuracies of 90.8 % and 93.0 %, respectively. This method entailed an initial rapid screening by TAG fingerprinting followed by UF fingerprinting to confirm the 25 % and 28 % of uncertain samples for varietal and geographical authentication, respectively.

Table 3
External validation of PLS-DA cultivar models developed on triacylglycerol (TAG) fingerprinting and unsaponifiable fraction (UF) fingerprinting data following a 2-step combined strategy. Results are mean values (± standard deviation) obtained from seven iterations.

Cultivar model: TG/non-TG								
TAG fingerprinting					Combined strategy (TAG + UF)			
	n	Correct (%)	Incorrect (%)	Uncertain (%)	Correct (%)	Incorrect (%)	Sensitivity	Specificity
non-TG	21	69.4 ± 13.7	8.2 ± 5.3	22.4 ± 11.6	89.1 ± 6.6	10.9 ± 6.6	0.89 ± 0.07	
TG	18	69.0 ± 10.6	5.6 ± 4.5	25.4 ± 9.0	92.9 ± 6.2	7.1 ± 6.2		0.93 ± 0.06
Total	39	69.2 ± 6.5	5.6 ± 1.9	24.8 ± 6.4	90.8 ± 4.4	9.2 ± 4.4		

TG: 'Tonda di Giffoni'; non-TG: other cultivars.

Table 4
External validation of PLS-DA origin models developed on triacylglycerol (TAG) fingerprinting and unsaponifiable fraction (UF) fingerprinting data following a combined strategy. Results are mean values (± standard deviation) obtained from seven iterations.

Geographical origin model: CHL/ESP/GEO/ITA								
TAG fingerprinting						Combined strategy (TAG + UF)		
	n	Correct (%)	Incorrect (%)	Not assigned (%)	Uncertain (%)	Correct (%)	Incorrect (%)	Not assigned (%)
CHL	8	80.4 ± 12.2	1.8 ± 4.7	0.0 ± 0.0	17.9 ± 14.2	96.4 ± 6.1	1.8 ± 4.7	1.8 ± 4.7
ESP	18	65.1 ± 10.0	4.8 ± 3.8	1.6 ± 2.7	28.6 ± 9.8	95.2 ± 3.8	4.8 ± 3.8	0.0 ± 0.0
GEO	8	69.6 ± 17.5	3.6 ± 6.1	0.0 ± 0.0	26.8 ± 16.8	96.4 ± 6.1	3.6 ± 6.1	0.0 ± 0.0
ITA	7	49.0 ± 13.9	16.3 ± 19.2	0.0 ± 0.0	34.7 ± 16.2	79.6 ± 18.2	16.3 ± 19.2	4.1 ± 10.8
Total	41	66.2 ± 7.9	5.9 ± 3.9	0.7 ± 1.2	27.2 ± 7.0	93.0 ± 4.3	5.9 ± 3.9	1.0 ± 1.9

CHL: Chile; ESP: Spain; GEO: Georgia; ITA: Italy.

4. Conclusions

Although UF fingerprinting requires more extensive analytical procedures than TAG fingerprinting, it demonstrated great accuracy in hazelnut classification, achieving overall accuracies of 94.1 % and 97.2 % for the cultivar and origin models, respectively. This positions UF fingerprinting as a suitable tool for verifying challenging samples and as a confirmatory method to support fast-screening tools.

On the other hand, hazelnut TAG fingerprinting is a faster and simpler method, easily applicable to large sample sets. It exhibited promising classification results, achieving an overall accuracy higher than 80 % for both models. However, it may have limitations when discrimination is more challenging, such as when different geographical areas share similar environmental conditions (e.g., Italy and Spain).

By integrating both methods in a 2-step approach, in which initial screening by TAG fingerprinting was followed by UF fingerprinting to confirm the 25 % to 28 % of uncertain samples, we achieved overall accuracies >90 % while maintaining a reduced workload. This integrated metabolic fingerprinting strategy represents a powerful tool for both varietal and geographical authentication, as demonstrated in the present study. Consequently, it could be applied and assessed in commercial hazelnut samples for integration into real production scenarios.

Funding

This work was developed in the context of the project TRACENUTS, PID2020-117701RB-I00 supported by MICIU/AEI/https://doi.org/10.13039/501100011033. B. Torres-Cobos thanks the Spanish Ministry of Universities predoctoral fellowships FPU20/014540. A. Tres received a Ramon y Cajal grant (RYC-2017-23601) funded by MICIU/AEI/https://doi.org/10.13039/501100011033 and by "ESF Investing in your future".

CRediT authorship contribution statement

B. Torres-Cobos: Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation. **S.B. Nicotra:** Formal analysis. **M. Rovira:** Resources, Conceptualization. **A. Romero:** Writing – review & editing, Resources, Conceptualization. **F. Guardiola:** Writing – review & editing, Supervision. **A. Tres:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. **S. Vichi:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing 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

Data is published in a repository

Hazelnut triacylglycerol fingerprints obtained by concatenating GC-MS Extracted Ion Chromatograms (Original data) (Dataverse)

Hazelnut unsaponifiable fraction fingerprints obtained by concatenating GC-MS Extracted Ion Chromatograms (Original data) (Dataverse)

Acknowledgments

INSA-UB Maria de Maeztu Unit of Excellence (Grant CEX2021-001234-M) funded by MICIU/AEI/FEDER, UE. INSA-UB authors are part of the LiBiFOOD Consolidated Research Group (2021-SGR-00854) recognised by AGAUR (Catalan Government). The authors would like to express their gratitude to Ferrero Hazelnut Company and Tuscia University (Department of Agriculture and Forest Science) for providing the hazelnut samples from Chile and Georgia, and Italy, respectively.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2024.141203.

References

Amaral, J. S., Casal, S., Citova, I., Santos, A., Seabra, R. M., & Oliveira, B. P. P. (2006). Characterization of several hazelnut (Corylus avellana L.) cultivars based in chemical, fatty acid and sterol composition. *European Food Research and Technology, 222*, 274–280. https://doi.org/10.1007/s00217-005-0068-0

Bachmann, R., Klockmann, S., Haerdter, J., Fischer, M., & Hackl, T. (2018). 1H NMR spectroscopy for determination of the geographical origin of hazelnuts. *Journal of Agricultural and Food Chemistry, 66*, 11873–11879. https://doi.org/10.1021/acs.jafc.8b03724

Ballin, N. Z., & Laursen, K. H. (2019). To target or not to target? Definitions and nomenclature for targeted versus non-targeted analytical food authentication. *Trends in Food Science & Technology, 86*, 537–543. https://doi.org/10.1016/j.tifs.2018.09.025

Benitez-Sanchez, P. L., León-Camacho, M. L., & Aparicio, R. (2003). A comprehensive study of hazelnut oil composition with comparisons to other vegetable oils, particularly olive oil. *European Food Research and Technology, 218*, 13–19. https://doi.org/10.1007/s00217-003-0766-4

Biancolillo, A., de Luca, S., Bassi, S., Roudier, L., Bucci, R., Magrì, A. D., & Marini, F. (2018). Authentication of an Italian PDO hazelnut ("Nocciola Romana") by NIR spectroscopy. *Environmental Science and Pollution Research, 25*, 28780–28786. https://doi.org/10.1007/s11356-018-1755-2

Bosque-Sendra, J. M., Cuadros-Rodríguez, L., Ruiz-Samblás, C., & de la Mata, A. P. (2012). Combining chromatography and chemometrics for the characterization and authentication of fats and oils from triacylglycerol compositional data—a review. *Analytica Chimica Acta, 724*, 1–11. https://doi.org/10.1016/j.aca.2012.02.041

Ciarmiello, L. F., Mazzeo, M. F., Minasi, P., Peluso, A., de Luca, A., Piccirillo, P., … Carbone, V. (2014). Analysis of different European hazelnut (Corylus avellana L.) cultivars: Authentication, phenotypic features, and phenolic profiles. *Journal of Agricultural and Food Chemistry, 62*, 6236–6246. https://doi.org/10.1021/jf5018324

Eilers, P. H. C. (2004). Parametric time warping. *Analytical Chemistry, 76*, 404–411. https://doi.org/10.1021/ac034800e

Eriksson, L., Trygg, J., & Wold, S. (2008). CV-ANOVA for significance testing of PLS and OPLS® models. *Journal of Chemometrics, 22*, 594–600. https://doi.org/10.1002/cem.1187

Farinelli, D., Tombesi, A., & Boco, M. (2019). Varietà vegetale di *Corylus avellana* L. (Nocciolo) dal nome tonda francescana. https://hdl.handle.net/11391/1223164.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Food and Agriculture Organization of the United Nations. (2022). FAOSTAT producer prices. https://www.fao.org/faostat/es/#data/PP (accessed January 10, 2024).

Ghisoni, S., Lucini, L., Rocchetti, G., Chiodelli, G., Farinelli, D., Tombesi, S., & Trevisan, M. (2020). Untargeted metabolomics with multivariate analysis to discriminate hazelnut (Corylus avellana L.) cultivars and their geographical origin. *Journal of the Science of Food and Agriculture, 100*, 500–508. https://doi.org/10.1002/jsfa.9998

Giulia, T., Vallauri, G., Pavese, V., Valentini, N., Rufa, P., Botta, R., & Torello-Marinoni, D. (2022). Identification of the hazelnut cultivar in raw kernels and in semi-processed and processed products. *European Food Research and Technology, 248*, 2431–2440. https://doi.org/10.1007/s00217-022-04058-z

Goriainov, S. V., Esparza, C. A., Borisova, A. R., Orlova, S. V., Vandyshev, V. V., Hajjar, F., … Kalabin, G. A. (2021). Phytochemical study of the composition of the unsaponifiable fraction of various vegetable oils by gas chromatography-mass spectrometry. *Journal of Analytical Chemistry, 76*, 1635–1644. https://doi.org/10.1134/S1061934821140045

Inaudi, P., Giacomino, A., Malandrino, M., la Gioia, C., Conca, E., Karak, T., & Abollino, O. (2020). The inorganic component as a possible marker for quality and for authentication of the hazelnut's origin. *International Journal of Environmental Research and Public Health, 17*, 447. https://doi.org/10.3390/ijerph17020447

Keles, O., & Taner, A. (2021). Classification of hazelnut varieties by using artificial neural network and discriminant analysis. *Spanish Journal of Agricultural Research, 19*(4), e0211. https://doi.org/10.5424/sjar/2021194-18056

Kiralan, S., Yorulmaz, A., Şimşek, A., & Tekin, A. (2015). Classification of Turkish hazelnut oils based on their triacylglycerol structures by chemometric analysis. *European Food Research and Technology, 240*, 679–688. https://doi.org/10.1007/s00217-014-2371-0

Klockmann, S., Reiner, E., Bachmann, R., Hackl, T., & Fischer, M. (2016). Food fingerprinting: metabolomic approaches for geographical origin discrimination of hazelnuts (Corylus avellana) by UPLC-QTOF-MS. *Journal of Agricultural and Food Chemistry, 64*, 9253–9262. https://doi.org/10.1021/acs.jafc.6b04433

Krauß, S., Vieweg, A., & Vetter, W. (2019). Stable isotope signatures (δ^2H-, δ^{13}C-, δ^{15}N-values) of walnuts (Juglans regia L.) from different regions in Germany. *Journal of*

the Science of Food and Agriculture, 100, 1625–1634. https://doi.org/10.1002/jsfa.10174

Król, K., & Gantner, M. (2020). Morphological traits and chemical composition of hazelnut from different geographical origins: A review. *Agriculture, 10*, 375. https://doi.org/10.3390/agriculture10090375

Lang, C., Weber, N., Möller, M., Schramm, L., Schelm, S., Kohlbacher, O., & Fischer, M. (2021). Genetic authentication: Differentiation of hazelnut cultivars using polymorphic sites of the chloroplast genome. *Food Control, 130*, Article 108344. https://doi.org/10.1016/j.foodcont.2021.108344

Matthäus, B., & Özcan, M. M. (2012). The comparison of properties of the oil and kernels of various hazelnuts from Germany and Turkey. *European Journal of Lipid Science and Technology, 114*, 801–806. https://doi.org/10.1002/ejlt.201100299

Menesatti, P., Costa, C., Paglia, G., Pallottino, F., D'Andrea, S., Rimatori, V., & Aguzzi, J. (2008). Shape-based methodology for multivariate discrimination among Italian hazelnut cultivars. *Biosystems Engineering, 101*(4), 417–424. https://doi.org/10.1016/j.biosystemseng.2008.09.013

Moscetti, R., Radicetti, E., Monarca, D., Cecchini, M., & Massantini, R. (2015). Near infrared spectroscopy is suitable for the classification of hazelnuts according to protected designation of origin. *Journal of the Science of Food and Agriculture, 95*(13), 2619–2625. https://doi.org/10.1002/jsfa.6992

Nielsen, N. P. V., Carstensen, J. M., & Smedsgaard, J. (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A, 805*, 17–35. https://doi.org/10.1016/S0021-9673(98)00021-1

Oddone, M., Aceto, M., Baldizzone, M., Musso, D., & Osella, D. (2009). Authentication and traceability study of hazelnuts from Piedmont, Italy. *Journal of Agricultural and Food Chemistry, 57*, 3404–3408. https://doi.org/10.1021/jf900312p

Parcerisa, J., Rafecas, M., Castellote, A. I., Codony, R., Farran, A., Garcia, J., … Boatella, J. (1994). Influence of variety and geographical origin on the lipid fraction of hazelnuts (Coryllus avellana L.) from Spain:(II). Triglyceride composition. *Food chemistry, 50*(3), 245–249. https://doi.org/10.1016/0308-8146(94)90128-7.

Parcerisa, J., Richardson, D. G., Rafecas, M., Codony, R., & Boatella, J. (1998). Fatty acid, tocopherol and sterol content of some hazelnut varieties (Corylus avellana L.) harvested in Oregon (USA). *Journal of Chromatography A, 805*, 259–268. https://doi.org/10.1016/S0021-9673(98)00049-1

Quintanilla-Casas, B., Bustamante, J., Guardiola, F., García-González, D. L., Barbieri, S., Bendini, A., … Tres, A. (2020). Virgin olive oil volatile fingerprint and chemometrics: Towards an instrumental screening tool to grade the sensory quality. *LWT - Food Science and Technology, 121*, Article 108936. https://doi.org/10.1016/j.lwt.2019.108936

Quintanilla-Casas, B., Marin, M., Guardiola, F., García-González, D. L., Barbieri, S., Bendini, A., … Tres, A. (2020). Supporting the sensory panel to grade virgin olive oils: An in-house-validated screening tool by volatile fingerprinting and chemometrics. *Foods, 9*, 1509. https://doi.org/10.3390/foods9101509

Quintanilla-Casas, B., Torres-Cobos, B., Guardiola, F., Servili, M., Alonso-Salces, R. M., Valli, E., … Tres, A. (2022). Geographical authentication of virgin olive oil by GCMS sesquiterpene hydrocarbon fingerprint: Verifying EU and single country label-declaration. *Food Chemistry, 378*, Article 132104. https://doi.org/10.1016/j.foodchem.2022.132104

Regulation (EU) 2024/1143 of the European Parliament and of the Council of 11 April 2024 on geographical indications for wine. (2024). *Spirit drinks and agricultural products, as well as traditional specialities guaranteed and optional quality terms for agricultural products, amending Regulations (EU) No 1308/2013, (EU) 2019/787 and (EU) 2019/1753 and repealing Regulation (EU) No 1151/2012. Official Journal of the European Union, L series, 2024/1143, 23.4.*.

Sammarco, G., Bardin, D., Quaini, F., Dall'Asta, C., Christmann, J., Weller, P., & Suman, M. (2023). A geographical origin assessment of Italian hazelnuts: Gas chromatography-ion mobility spectrometry coupled with multivariate statistical analysis and data fusion approach. *Food Research International, 171*, Article 113085. https://doi.org/10.1016/j.foodres.2023.113085

Shakiba, N., Gerdes, A., Holz, N., Wenck, S., Bachmann, R., Schneider, T., Seifert, S., Fischer, M., & Hackl, T. (2022). Determination of the geographical origin of hazelnuts (Corylus avellana L.) by near-infrared spectroscopy (NIR) and a low-level fusion with nuclear magnetic resonance (NMR). *Microchemical Journal, 174*, Article 107066. https://doi.org/10.1016/j.microc.2021.107066

Silvestri, C., Bacchetta, L., Bellincontro, A., & Cristofori, V. (2021). Advances in cultivar choice, hazelnut orchard management, and nut storage to enhance product quality and safety: An overview. *Journal of the Science of Food and Agriculture, 101*, 27–43.

Stella, E., Moscetti, R., Carletti, L., Menghini, G., Fabrizi, F., Cristofori, V., Monarca, D., Cecchini, M., & Massantini, R. (2013). Near-infrared spectroscopy is feasible to discriminate hazelnut cultivars. *Journal of Agricultural Engineering, 44*(s2). https://doi.org/10.4081/jae.2013.297

Torres-Cobos, B., Quintanilla-Casas, B., Guardiola, F., Valli, E., Bendini, A., Gallina Toschi, T., … Vichi, S. (2021). Extracting different information from a sesquiterpene hydrocarbon fingerprint by PLS-DA: A case study for the geographical and varietal authentication of olive oil. Road to CAC, Zoom Conference 20–21 July 2021. https://cac2022.sciencesconf.org/data/CAC2022_BookOfAbstracts.pdf.

Torres-Cobos, B., Quintanilla-Casas, B., Romero, A., Ninot, A., Alonso-Salces, R. M., Gallina Toschi, T., … Vichi, S. (2021). Varietal authentication of virgin olive oil: Proving the efficiency of sesquiterpene fingerprinting for mediterranean arbequina oils. *Food Control, 128*, Article 108200. https://doi.org/10.1016/j.foodcont.2021.108200

Torres-Cobos, B., Quintanilla-Casas, B., Rovira, M., Romero, A., Guardiola, F., Vichi, S., & Tres, A. (2023). Revealing adulterated olive oils by triacylglycerol screening methods: Beyond the official method. *Food Chemistry, 409*, Article 135256. https://doi.org/10.1016/j.foodchem.2022.135256

Torres-Cobos, B., Quintanilla-Casas, B., Rovira, M., Romero, A., Guardiola, F., Vichi, S., & Tres, A. (2024). Prospective exploration of hazelnut's unsaponifiable fraction for geographical and varietal authentication: A comparative study of advanced fingerprinting and untargeted profiling techniques. *Food Chemistry, 441*, Article 138294. https://doi.org/10.1016/j.foodchem.2023.138294

Westerhuis, J. A., Hoefsloot, H. C., Smit, S., Vis, D. J., Smilde, A. K., van Velzen, E. J., … van Dorsten, F. A. (2008). Assessment of PLSDA cross validation. *Metabolomics, 4*, 81–89. https://doi.org/10.1007/s11306-007-0099-6

Xu, L., Hu, O., Guo, Y., Zhang, M., Lu, D., Cai, C. B., … She, Y. B. (2018). Representative splitting cross validation. *Chemometrics and Intelligent Laboratory Systems, 183*, 29–35. https://doi.org/10.1016/j.chemolab.2018.10.008

5.4.3 Spectroscopic fingerprinting

Spectroscopic techniques combined with chemometrics are powerful screening tools for food authentication. As NIR and MIR spectra are influenced by components of the lipid fraction, the successfulness of TAG and UF for hazelnut authentication suggested them as suitable rapid techniques.

To assess and compare various spectroscopic techniques, and to develop and validate a reliable method based on spectroscopic fingerprinting to simultaneously verify both hazelnut cultivar and provenance, the expanded hazelnut sample set that had already been used for the isotope-based models (Publication 5; section 5.2.1) and for the UF and TAG models (Publication 8; section 5.4.1) was analysed. This set included 302 hazelnut samples:

- 200 samples of TG cultivar hazelnuts from different countries (CHL, n=40), Spain (ESP, n=91), Georgia (GEO, n=34), and Italy (ITA, n=35).
- various cultivars of hazelnuts grown in Spain (n=91 TG; n=102 from other cultivars (non-TG).

The analyses were conducted using a NIR spectrometer, a handheld NIR spectrometer (hNIR), and a MIR spectrometer. Samples were analysed as whole raw kernels and in their ground form (**Figure 14**).

As in previous studies (Publication 8), independent PLS-DA models for cultivar and origin were developed and externally validated for each spectroscopic method. The regression coefficients of the best-performing models were examined to identify key variables that most effectively discriminated between cultivars and origins, and to tentatively link these variables to chemical compounds. The models' performances from each method and sample preparation were compared to determine the most suitable approach for hazelnut authentication (Publication 9).

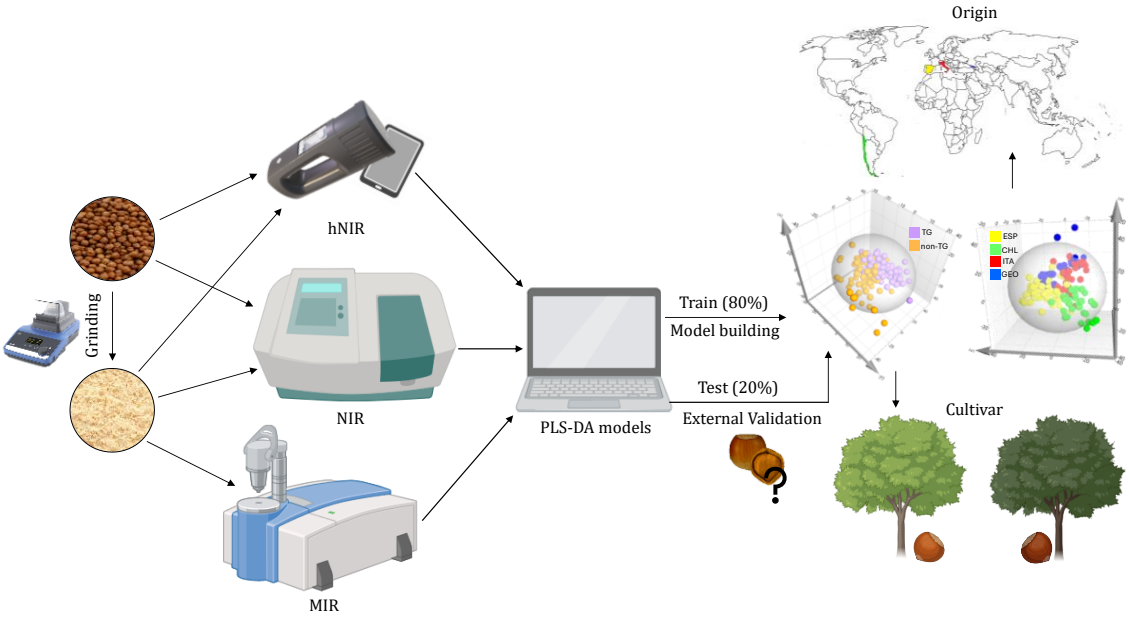


Figure 14. Graphical abstract of Publication 9. MIR: mid-infrared spectroscopy, NIR: near-infrared spectroscopy, hNIR: handheld near-infrared spectroscopy, PLS-DA: partial least square-discriminant analysis, CHL: Chile, ESP: Spain, GEO: Georgia, ITA: Italy, TG: Tonda di Giffoni, non-TG: other cultivars.

5.4.4 Publication 9

Comparative analysis of spectroscopic methods for rapid authentication of hazelnut cultivar and origin

Berta Torres-Cobos, Alba Tres, Stefania Vichi, Francesc Guardiola, Mercè Rovira, Agustí Romero, Vincent Baeten, Juan Antonio Fernández-Pierna

Spectrochimica Acta, Part A: Molecular and Biomolecular Spectroscopy, ***2025****, 326, 125367. https://doi.org/10.1016/j.saa.2024.125367*

Supplementary material available in **Annex 8**

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 326 (2025) 125367

Contents lists available at ScienceDirect

Spectrochimica Acta Part A:
Molecular and Biomolecular Spectroscopy

journal homepage: www.journals.elsevier.com/spectrochimica-acta-part-a-molecular-and-biomolecular-spectroscopy

Comparative analysis of spectroscopic methods for rapid authentication of hazelnut cultivar and origin

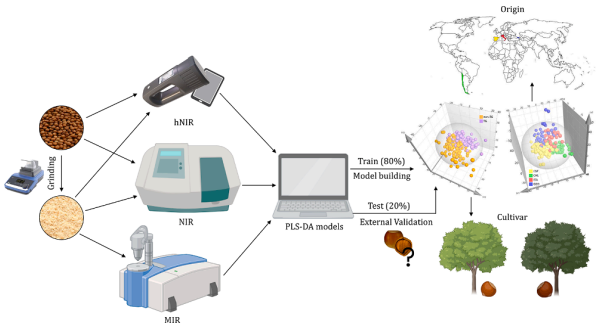
B. Torres-Cobos [a,b], A. Tres [a,b], S. Vichi [a,b,*], F. Guardiola [a,b], M. Rovira [c], A. Romero [c], V. Baeten [d], J.A. Fernández-Pierna [d]

[a] Departament de Nutrició, Ciències de L'Alimentació i Gastronomia, Facultat de Farmàcia i Ciències de L'Alimentació, Universitat de Barcelona, Av Prat de La Riba, 171, 08921 Santa Coloma de Gramenet, Spain
[b] Institut de Recerca en Nutrició i Seguretat Alimentària (INSA-UB), Universitat de Barcelona, Av Prat de La Riba, 171, 08921 Santa Coloma de Gramenet, Spain
[c] Institute of Agrifood Research and Technology (IRTA), Ctra. de Reus – El Morell Km 3.8, Constantí 43120, Spain
[d] Quality and Authentication of Products Unit, Knowledge and Valorization of Agricultural Products Department, Walloon Agricultural Research Centre, Chaussée de Namur 24, 5030 Gembloux, Belgium

HIGHLIGHTS

- Three spectroscopic methods were compared to verify hazelnut cultivars and origin.
- MIR and NIR spectroscopic methods achieved ≥93 % accuracy in classifying hazelnuts.
- Benchtop NIR spectroscopy showed superior performance for hazelnut authentication.
- Ground hazelnuts provide better results than whole kernels due to greater homogeneity.
- Models rely on protein and lipid composition for hazelnut discrimination.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Hazelnut
Geographical origin
Varietal
Authentication
Near infrared spectroscopy (NIR)
Handheld near infrared (hNIR)
Medium infrared spectroscopy (MIR)
PLS-DA

ABSTRACT

Hazelnut market prices fluctuate significantly based on cultivar and provenance, making them susceptible to counterfeiting. To develop an accurate authentication method, we compared the performances of three spectroscopic methods: near infrared (NIR), handheld near infrared (hNIR), and medium infrared (MIR), on over 300 samples from various origins, cultivars, and harvest years. Spectroscopic fingerprints were used to develop and externally validate PLS-DA classification models. Both cultivar and origin models showed high accuracy in external validation. The hNIR model effectively distinguished cultivars but struggled with geographic distinctions due to lower sensitivity. NIR and MIR models showed over 93 % accuracy, with NIR slightly outperforming MIR for geographic origin. NIR proved to be a fast and suitable tool for hazelnut authentication. This study is the first to systematically compare spectroscopic tools for authenticating hazelnut cultivar and origin using the same dataset, offering valuable insights for future food authentication applications.

* Corresponding author at: Departament de Nutrició, Ciències de L'Alimentació i Gastronomia, Facultat de Farmàcia i Ciències de L'Alimentació, Universitat de Barcelona, Av Prat de La Riba, 171, 08921 Santa Coloma de Gramenet, Spain.
E-mail address: stefaniavichi@ub.edu (S. Vichi).

https://doi.org/10.1016/j.saa.2024.125367
Received 12 July 2024; Received in revised form 1 October 2024; Accepted 30 October 2024
Available online 1 November 2024

1. Introduction

Hazelnuts are one of the most consumed nuts in Europe, as raw and roasted fruits or included in many preparations and traditional dishes. In addition, they are widely used in the chocolate and confectionary industries. The main hazelnut-producing countries are Turkey, Italy, the USA, Azerbaijan, Chile and Georgia [1]. Hazelnut nutritional values and sensory attributes are greatly influenced by geographical and varietal origin [2–4]. Consequently, their market price also fluctuates significantly depending on cultivar and provenance. Italian and Spanish hazelnuts are among the highest priced, with values of 3,416 USD/T and 2,434 USD/T, respectively, in contrast to Georgian kernels, which are priced at 1,287 USD/T [5]. Besides, hazelnuts with special geographical indications, such as Protected Designation of Origin or Protected Geographical Indication, are also highly appreciated by consumers, which is reflected in their price.

The high economic value of hazelnuts makes them susceptible to fraudulent practices. This susceptibility is accentuated by the absence of effective fraud detection methods. Appropriate tools to verify the cultivar and origin of hazelnuts are therefore necessary to ensure their authenticity and to safeguard consumers' rights.

Different analytical approaches have addressed this issue; morphological analysis [6,7], genetic methods [8,9], metabolomic analysis of proteins, phenolic and lipidic compounds by chromatographic techniques [6,10–12] and proton nuclear magnetic resonance (^{1}H NMR) [13] have been applied for hazelnut authentication. Nevertheless, these methods are usually laborious, time consuming, not suitable to be applied on-site (i.e. in the field, in the industry or in storage/retail facilities), require highly trained personnel and have a substantial cost.

Spectroscopic methods, such as mid-infrared (MIR) and near-infrared (NIR) spectroscopy, are fast, simple, environmentally friendly and non-destructive techniques, which can be applied directly to the solid sample without complex sample pretreatment steps. They allow simultaneous analysis of different chemical compounds and can be easily adapted to on-site or on-line applications, making them suitable for routine analysis.

These techniques, coupled with chemometric data analysis tools, have been successfully applied to the varietal and geographical authentication of different nuts [14–18]. In particular, previous works used NIR spectroscopy to authenticate an Italian PDO "Nocciola Romana" [19,20] to distinguish between Turkish hazelnut cultivars [21], or to discriminate hazelnuts from different origins [22,23]. However, none of them has been applied to simultaneously authenticate the origin and cultivar of hazelnuts. To the authors' knowledge, only Manfredi et al. [24] applied MIR spectroscopy to hazelnut halves, successfully differentiating between three cultivars. No further studies using MIR spectroscopy to authenticate hazelnuts have been found, underscoring the need for further evaluation of the technique's potential. Overall, studies using large sample sets, including higher variability in terms of harvest years, origins, cultivars and producers, are still needed to further corroborate the suitability of NIR and MIR spectroscopies as tools for geographical and varietal authentication of hazelnuts.

On the other hand, the performance of different spectroscopic methods for hazelnut authentication has not been systematically compared. Such a comparison would be particularly interesting in the case of MIR and NIR, as each provides different but complementary information. NIR spectroscopy includes bands issued from overtones and combination vibrations. It allows direct analysis of highly absorbing samples. NIR measurements often result in overlapping bands and complex spectra, so the bands are less specific in the NIR range than in MIR, which may hinder their interpretability. In contrast, MIR spectra include information of fundamental vibrations of specific functional groups, providing spectra with better resolved bands that can be used for structural identification [25]. In addition, NIR provides deeper light penetration into the sample matter, performing better with bulk or intact heterogeneous samples, minimizing the need for sample preparation and having a wider scope of applications. This makes it ideal for in situ analysis, as it requires less specificity requirements and less sample preparation [26,27]. On this basis, a large number of handheld NIR (hNIR) devices have been developed to authenticate a wide range of products [28–32]. These devices are small, compact, robust, more practical and affordable; they do not require a high level of expertise and can be applied for routine in-field analysis. However, handheld systems can reduce the accuracy of measurement, which can affect spectral quality compared to benchtop alternatives [33].

The aim of this research was to systematically apply, compare and evaluate the effectiveness of NIR and MIR spectroscopies for hazelnut authentication. The main goal was to develop an accurate method to simultaneously discriminate hazelnuts according to their geographical and varietal origin. For this purpose, in order to ensure a direct comparison between the different spectroscopic techniques and to properly evaluate their performances as hazelnut authentication tools, we analysed the same set of hazelnuts (both whole kernels and ground samples) from different origins, cultivars, and harvest years. This analysis was carried out by a NIR spectrometer, a hNIR spectrometer, and a MIR spectrometer (the latter only for ground samples). Subsequently, individual PLS-DA classification models for cultivar and origin discrimination were built for each technique. The geographical origin was discriminated between samples of the same cultivar ('Tonda di Giffoni', TG) produced in different countries, and the cultivar was discriminated between samples of the same origin (Spain). The cultivar models focused on discriminating TG hazelnuts, one of the most widespread cultivars worldwide, from other (non-TG) cultivars, while the geographical origin models were designed to classify samples according to their country of origin. The classification models were externally validated and evaluated for both fit and predictive ability.

2. Material and methods

2.1. Samples

A set of 302 traceable hazelnut samples was obtained directly from producers within the framework of the TRACENUTS project (PID2020-117701RB-I00). Samples were collected over four consecutive harvest years, from 2019 to 2022. From these samples, 200 were of the 'Tonda di Giffoni' cultivar (TG) from Chile (CHL, n = 40), Spain (ESP, n = 91), Georgia (GEO, n = 34) and Italy (ITA, n = 35), while 102 were from different cultivars (non-TG) produced in Spain (Table S1, Supplementary material). Samples were stored vacuum-packed at 4 °C until analysis.

2.2. Sample preparation

Raw hazelnut kernels with skin were directly analysed by NIR spectroscopy (benchtop and handheld device). Then, 15 g of the sample was ground with an IKA TUBE MILL control, (IKA, Staufen, Germany), and was analysed by NIR (benchtop and handheld device) and MIR spectroscopy.

2.3. NIR spectroscopy

2.3.1. Benchtop NIR spectrometer

The NIR measurements of the whole and ground hazelnut kernels were performed on a benchtop DS3 FOSS spectrometer (FOSS Analytics, Hilleroed, Denmark) acquiring spectra every 2 nm within the wavelength range of 400–2500 nm, (spectral resolution of 0.5 nm). Whole kernels were presented in a cell quarter cup, while ground samples were placed in ring cups. A background reference spectrum was collected before each measure. Two consecutive measures, being each the average of seven spectra at 4 different points of the cup (28 spectra), were collected for every sample. Both measures were averaged previously to chemometric analysis. ISIscan Nova ™ (2021, FOSS) was used for NIR

spectra recording.

2.3.2. Handheld NIR spectrometer (hNIR)

NIR spectra of the whole and ground samples, presented in the same way as before, were also acquired using a handheld device, NeoSpectra Scanner (Si-Ware Systems Inc., California, U.S.). Data was collected with the NeoSpectra Scan software v7.1. Spectra was acquired every 2 nm within the wavelength range of 1350–2500 nm, (spectral resolution of 16 nm) and a scanning time of 3 s. Two consecutive replicates were collected per sample and calibration was run every 10 measures (5 samples). Prior to data analysis the corresponding replicated spectra were averaged.

2.4. MIR spectroscopy

MIR measurements of hazelnut ground samples were performed on a Vertex 70 spectrometer (Bruker Optics, Ettlingen, Germany) equipped with Attenuated Total Reflectance (ATR) and an integrated press. Samples were directly deposited on the crystal and pressed against it to ensure optimum contact with the diamond. Each measure was the average spectrum of 64 scans. Spectra was acquired every 1.93 cm^{-1} within the wavelength range of 600–4000 cm^{-1}, (spectral resolution of 4 cm^{-1}). Background readings were collected prior to each sample spectrum collection. Three consecutive replicates were collected per sample and averaged prior to chemometric analysis. OPUS software (version 8.2.28) was used for MIR spectra recording.

2.5. Chemometrics

Spectral data matrices were extracted as csv files and processed using SIMCA v13.0© (Sartorius Lab Instruments GmbH & Co. KG, Gottingen, Germany).

2.5.1. Partial least squares discriminant analysis (PLS-DA)

For each method and sample preparation (NIR, whole or ground; hNIR, whole or ground; and MIR, ground), individual PLS-DA classification models were built to authenticate cultivar or geographical origin. Cultivar models were binary models to discriminate between two classes of Spanish samples (n = 193): TG cultivar (n = 91) and non-TG cultivar (n = 102). Origin models were multi-class models aimed at classifying TG samples (n = 200) according to their country of origin: CHL (n = 40), ESP (n = 91), GEO (n = 34) or ITA (n = 35).

Prior to model building, for each type of authentication model (origin or cultivar), each sample set was randomly split into training (80 % of the samples of each category: cultivar model, n = 154; origin model, n = 160) and validation set (20 % of the samples of each category: cultivar model, n = 39; origin model n = 40). The splitting was run seven times (7 iterations) for each authentication model to increase the robustness of the external validation, resulting in seven different training sets and their corresponding validation sets. Although randomly split, a stratified sampling strategy was followed by maintaining the variability and proportions of the sample set in the validation and training sets (Table S1, Supplementary material). The exact same splitting and resulting training and validation sets were used for all the methods to ensure the direct comparability between them.

In cultivar models (by binary PLS-DA), classes were expressed as PLS dummy variables, where 'non-TG' was represented as 1 and 'TG' as 0. The PLS predicted value (PV) of each sample was used for its classification into one of the two classes based on a classification threshold (PV = 0.5). Origin models (by multi-class PLS-DA) operated as multiple binary models, each comparing one class against the others. A dummy Y matrix held a set of classification vectors equal to the number of classes, where each vector had a value of 1 for one class (CHL, ESP, GEO or ITA) and 0 for all the other classes (non-CHL, non-ESP, non-GEO or non-ITA). Each sample was classified into the class corresponding to the vector leading to the highest PV, provided it was above the classification threshold (here, PV = 0.5). Samples with PV below the classification threshold (PV < 0.5) for all vectors were not assigned to any class.

Models were developed on the training sets and first internally validated through leave 10 %-out cross-validation using the samples of the training sets. The optimal number of latent variables (LV) and preprocessing were selected according to the lowest Root Mean Squared Error of Cross Validation (RMSEcv) criteria. Hotelling's T^2 and Q residuals were used to detect outliers [34]. Permutation test (n = 20 permutations) and ANOVA on the cross-validated predictive residuals (p-value) [35] were carried out to assess the risk of model overfitting. Finally, models were externally validated by predicting the class of the samples in the respective validation set, which had not been used to build the models. The suitability of each PLS-DA model was evaluated by the Q^2 values and efficiency, which was expressed as the rate of correct classification of each class. Additionally, for the binary cultivar models the sensitivity, Eq. (1) and specificity, Eq. (2) were also assessed. True positives were the non-TG samples correctly classified, and true positives + false negatives corresponded to the total non-TG samples. True negatives were the TG samples correctly classified, and true negatives + false positives corresponded to the total TG samples [36].

$$Sensitivity = \frac{true\ positives}{[true\ positives + false\ negatives]} \quad (1)$$

$$Specificity = \frac{true\ negatives}{[true\ negatives + false\ positives]} \quad (2)$$

The performance of models from each method and sample preparation was compared to determine the most suitable one for authentication.

2.5.2. Evaluation of PLS-DA regression coefficients

For the methods that gave the most promising results, models were rebuilt using all samples (cultivar n = 193, or origin n = 200), and their regression coefficients were explored to identify the most relevant variables for the classification (cultivar or origin) and tentatively link them to chemical species. The jack-knife standard error of cross-validation (SEcv) was used to evaluate the significance of the regression coefficients, with values exceeding their corresponding SEcv considered significant. Out of the significant variables, only those with the highest absolute values (25 % higher than the average of the coefficients) were considered.

3. Results and discussion

3.1. PLS-DA classification models

3.1.1. Whole kernels

For the PLS-DA models developed on the whole kernels analysed by the benchtop NIR and hNIR spectrometers data, the optimal preprocessing, according to the lowest RMSEcv criterion, was a first derivative, Savitzky–Golay smoothing (second order with a polynomial filter of 15-point window) and mean centring and scaling to unit of variance. For models developed with the benchtop NIR spectrometer, data was also pre-processed with the SNV. No outliers were detected according to the Hotelling's T^2 range and Q residuals.

The cross-validation results of the models built from training sets (7 iterations) were promising for TG cultivar discrimination for both NIR and hNIR models, with correct classification rates of 96.0 % and 87.2 %, respectively (Table S2, Supplementary material). However, the performance of the model for discrimination according to origin appears to be lower, in particular when using hNIR data, achieving only a 66.3 % of correct classification (Table S3, Supplementary material).

These findings were corroborated through external validation by predicting the class of the samples of the corresponding validation sets. Tables 1 and 2 present the mean values obtained from seven iterations of the external validation of each type of authentication model (Cultivar:

TG/non-TG; Geographical origin: CHL/ESP/GEO/ITA) developed with NIR and hNIR data from whole kernels. In all cases involving whole kernel analysis, the NIR models outperformed the hNIR models with higher sensitivity, specificity, and total correct classification rate (91.2 % vs 82.4 % in cultivar model; 81.1 % vs 53.2 % in origin model). The NIR and hNIR cultivar models showed high sensitivity (0.90 and 0.83) and specificity (0.93 and 0.82), proving their efficiency in discriminating TG samples from other cultivars (Table 2). Nonetheless, the origin models showed lower rates of correct classification for both NIR and hNIR, especially for GEO and ITA (Table 3). As observed, neither the NIR nor the hNIR were able to accurately classify the ITA samples, which were mostly not assigned to any class or were misclassified as ESP samples. This could be attributed to the proximity of these two regions and the similarity of their pedoclimatic conditions, which may result in similar effects on hazelnut composition. The poor performance of the hNIR origin models indicates that it is not suitable for accurate classification of whole hazelnuts according to their geographical origin.

3.1.2. Ground samples

For NIR models based on ground samples, a first derivative and smoothing by Savitzky–Golay (second-order polynomial filter with a 15-point window) was applied along with SNV and mean centring and scaling to unit of variance. For hNIR models, the optimal pre-processing was SNV followed by mean centring and scaling to unit of variance, whereas for MIR models a first derivative and smoothing (Savitzky–Golay second-order polynomial filter with a 5-point window) was applied. No outliers were detected according to the Hotelling's T^2 range and Q residuals. Both types of authentication models (origin and cultivar) developed on ground samples by NIR and MIR presented successful cross-validation results with a high overall rate of correct classification (98.7–99.8 % cultivar model; 95.4–98.0 % origin model), while the accuracy was lower for hNIR models (89 % overall correct classification for both models) (Tables S4 and S5 of Supplementary information).

According to the external validation results, hNIR model of ground samples could be useful to discriminate TG cultivar from other cultivars (0.80 sensitivity, 0.89 specificity and 83.9 % correct classification) (Table 3) but it was still unable to accurately distinguish Georgian and Italian samples from those produced in other regions (Table 4). This could be attributed to the lower sensitivity of the hNIR spectrometer compared to a benchtop spectrometer.

On the other hand, the NIR and MIR models based on ground sample data performed satisfactorily in external validation. For cultivar models, both techniques achieved a sensitivity equal or higher than 0.92 and a specificity of 0.98, resulting in an overall classification rate of 95 % for both techniques (Table 3). Concerning the geographical origin models, NIR outperformed MIR in classifying hazelnuts from GEO (91.8 % vs 85.7 %) and ITA (91.8 % vs 83.7 %), providing slightly better overall classification results (96.4 % vs 93.9 %, respectively) (Table 4). Consequently, NIR proved to provide the most successful spectroscopic model for hazelnut varietal and geographical authentication.

In all cases, the results for the ground samples outperformed those obtained with the whole kernels, in line with studies performed on other nuts [37]. This could be attributed to the higher homogeneity and representativeness of the ground samples compared to analysing only a

Table 1
External validation of PLS-DA cultivar models developed on whole kernels analysed by benchtop NIR and handheld NIR (hNIR) spectrometers. Results are mean values (±standard deviation) obtained from seven iterations.

Whole kernel – Cultivar model: TG/non-TG						
NIR spectrometer (LVs = 9, Q^2 = 0.47, RMSEcv = 0.38)[a]						
	n	Non-TG (n)	TG (n)	Correct classification (%)	Sensitivity	Specificity
Non-TG	21	18.9 ± 1.3	2.1 ± 1.3	89.8 ± 6.4	0.90 ± 0.06	
TG	18	1.3 ± 1.1	16.7 ± 1.1	92.9 ± 6.2		0.93 ± 0.06
Total	39			91.2 ± 4.4		
hNIR spectrometer (LVs = 6, Q^2 = 0.35, RMSEcv = 0.41)[a]						
	n	Non-TG (n)	TG (n)	Correct classification (%)	Sensitivity	Specificity
Non-TG	21	17.4 ± 1.7	3.6 ± 1.7	83.0 ± 8.2	0.83 ± 0.08	
TG	18	3.3 ± 1.3	14.7 ± 1.3	81.7 ± 7.0		0.82 ± 0.07
Total	39			82.4 ± 6.4		

For all models, ANOVA p-value < 0.05.
[a] Model parameters: mean values obtained with the training sets from 7 iterations. LVs: latent variables of the training model; Q^2: Cumulative fraction of Y variation predicted by the X training model up to the specified latent variable, according to cross-validation; RMSEcv: root mean square error of the cross validation of the training model. TG: 'Tonda di Giffoni'; non-TG: other cultivars.

Table 2
External validation of PLS-DA origin models developed on whole kernels analysed by benchtop NIR (NIR) and handheld NIR (hNIR) spectrometers. Results are mean values (±standard deviation) obtained from seven iterations.

Whole kernel – Geographical origin model: CHL/ESP/GEO/ITA							
NIR spectrometer (LVs = 14, Q^2 = 0.27, RMSEcv = 0.36)[a]							
	n	CHL (n)	ESP (n)	GEO (n)	ITA (n)	Not assigned (n)	Correct classification (%)
CHL	8	6.6 ± 1.1	0.6 ± 0.8	0.0 ± 0.0	0.0 ± 0.0	0.9 ± 1.2	82.1 ± 14.2
ESP	18	0.0 ± 0.0	16.7 ± 1.0	0.1 ± 0.4	0.3 ± 0.5	0.9 ± 0.9	92.9 ± 5.3
GEO	7	0.3 ± 0.5	0.4 ± 0.8	5.0 ± 1.5	0.3 ± 0.5	1.0 ± 1.0	71.4 ± 21.8
ITA	7	0.0 ± 0.0	1.6 ± 1.0	0.0 ± 0.0	4.1 ± 1.3	1.3 ± 1.7	59.2 ± 19.2
Total	40					4.0 ± 2.2	81.1 ± 7.2
hNIR spectrometer (LVs = 9, Q^2 = 0.23, RMSEcv = 0.36)[a]							
	n	CHL (n)	ESP (n)	GEO (n)	ITA (n)	Not assigned (n)	Correct classification (%)
CHL	8	3.7 ± 1.6	0.9 ± 0.7	0.0 ± 0.0	0.4 ± 0.5	3.0 ± 1.2	46.4 ± 20.0
ESP	18	0.9 ± 0.9	12.9 ± 2.0	0.6 ± 0.5	0.0 ± 0.0	3.7 ± 2.3	71.4 ± 10.8
GEO	7	0.0 ± 0.0	1.9 ± 0.9	2.1 ± 1.1	0.1 ± 0.4	2.9 ± 1.6	30.6 ± 15.3
ITA	7	0.7 ± 0.5	2.1 ± 1.5	0.0 ± 0.0	2.6 ± 0.8	1.6 ± 1.3	36.7 ± 11.2
Total	40					11.1 ± 3.2	53.2 ± 5.7

For all models, ANOVA p-value < 0.05.
[a] Model parameters: mean values obtained with the training sets from 7 iterations. LVs: latent variables of the training model; Q^2: Cumulative fraction of Y variation predicted by the X training model up to the specified latent variable, according to cross-validation; RMSEcv: root mean square error of the cross validation of the training model. CHL: Chile; ESP: Spain; GEO: Georgia; ITA: Italy.

Table 3
External validation of PLS-DA cultivar models developed on ground samples analysed by benchtop NIR (NIR), handheld NIR (hNIR) and MIR spectrometers. Results are mean values (±standard deviation) obtained from seven iterations.

Ground samples – Cultivar model: TG/non-TG						
NIR spectrometer (LVs = 10, Q^2 = 0.53, RMSEcv = 0.34)[a]						
	n	Non-TG (n)	TG (n)	Correct classification (%)	Sensitivity	Specificity
Non-TG	21	19.3 ± 0.8	1.7 ± 0.8	91.8 ± 3.6	0.92 ± 0.04	
TG	18	0.4 ± 0.5	17.7 ± 0.5	98.4 ± 2.7		0.98 ± 0.03
Total	39			94.9 ± 2.1		
hNIR spectrometer (LVs = 9, Q^2 = 0.39, RMSEcv = 0.39)[a]						
	n	Non-TG (n)	TG (n)	Correct classification (%)	Sensitivity	Specificity
Non-TG	21	16.1 ± 1.0	4.3 ± 1.0	79.6 ± 4.5	0.80 ± 0.05	
TG	18	2.0 ± 0.8	16.0 ± 0.8	88.9 ± 4.5		0.89 ± 0.05
Total	39			83.9 ± 2.4		
MIR spectrometer (LVs = 10, Q^2 = 0.63, RMSEcv = 0.30)[a]						
	n	Non-TG (n)	TG (n)	Correct classification (%)	Sensitivity	Specificity
Non-TG	21	19.6 ± 1.4	1.4 ± 1.4	93.2 ± 6.7	0.93 ± 0.07	
TG	18	0.4 ± 0.5	17.6 ± 0.5	97.6 ± 3.0		0.98 ± 0.03
Total	39			95.2 ± 2.3		

For all models, ANOVA p-value < 0.05.
[a] Model parameters: mean values obtained with the training sets from 7 iterations. LVs: latent variables of the training model; Q^2: Cumulative fraction of Y variation predicted by the X training model up to the specified latent variable, according to cross-validation; RMSEcv: root mean square error of the cross validation of the training model. TG: 'Tonda di Giffoni'; non-TG: other cultivars.

single or a few kernels per sample. Additionally, grinding not only improves sample homogeneity but also causes oil release from the cells, which affects the primary spectral signals. In whole kernels with skin, the dominant signals arise primarily from the lignocellulosic structure, whereas in ground samples, the signal from the released oil becomes more prominent. Consequently, the difference in model performance can be partly attributed to this shift in dominant signals between the two sample types.

3.2. Regression coefficients

The regression coefficients of the two best performing methods (NIR and MIR spectroscopic analysis of ground hazelnuts) were explored to identify the most informative variables in PLS-DA models for cultivar or origin discrimination and to relate them to the chemical composition of the samples.

3.2.1. NIR spectroscopy

The interpretation of NIR spectra might be hindered by the fact that some of the bands in the analysed range are overtones containing generic information corresponding to different molecular vibrations of different functional groups. Even so, prior knowledge of the hazelnut's chemical composition enables the tentative identification of variables relevant in classifying the samples within each category. These variables can then be correlated with the main compositional constituents of the samples (lipids, proteins, carbohydrates).

Table 4
External validation of PLS-DA origin models developed on ground samples analysed by benchtop NIR (NIR), handheld NIR (hNIR) and MIR spectrometers. Results are mean values (±standard deviation) obtained from seven iterations.

Ground samples – Geographical origin model: CHL/ESP/GEO/ITA								
NIR spectrometer (LVs = 14, Q^2 = 0.66, RMSEcv = 0.25)[a]								
	n	CHL (n)	ESP (n)	GEO (n)	ITA (n)	Not assigned (n)	Correct classification (%)	
CHL	8	8.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	
ESP	18	0.0 ± 0.0	17.7 ± 0.5	0.0 ± 0.0	0.0 ± 0.0	0.3 ± 0.5	98.4 ± 2.7	
GEO	7	0.0 ± 0.0	0.3 ± 0.5	6.4 ± 0.4	0.0 ± 0.0	0.4 ± 0.5	91.8 ± 7.6	
ITA	7	0.1 ± 0.4	0.0 ± 0.0	0.0 ± 0.0	6.4 ± 0.8	0.4 ± 0.5	91.8 ± 11.2	
Total	40					1.1 ± 0.9	96.4 ± 2.4	
hNIR spectrometer (LVs = 9, Q^2 = 0.50, RMSEcv = 0.29)[a]								
	n	CHL (n)	ESP (n)	GEO (n)	ITA (n)	Not assigned (n)	Correct classification (%)	
CHL	8	7.3 ± 0.5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.7 ± 0.5	91.1 ± 6.1	
ESP	18	0.1 ± 0.4	17.1 ± 0.9	0.0 ± 0.0	0.0 ± 0.0	0.7 ± 1.0	95.2 ± 5.0	
GEO	7	0.1 ± 0.4	0.7 ± 0.5	4.1 ± 1.1	0.3 ± 0.8	1.7 ± 1.3	59.2 ± 15.3	
ITA	7	0.0 ± 0.0	0.6 ± 0.5	0.3 ± 0.5	4.6 ± 0.5	1.6 ± 0.8	65.3 ± 7.6	
Total	40					4.7 ± 1.8	82.9 ± 1.7	
MIR spectrometer (LVs = 12, Q^2 = 0.61, RMSEcv = 0.26)[a]								
	n	CHL (n)	ESP (n)	GEO (n)	ITA (n)	Not assigned (n)	Correct classification (%)	
CHL	8	8.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	
ESP	18	0.0 ± 0.0	17.7 ± 0.5	0.0 ± 0.0	0.0 ± 0.0	0.3 ± 0.5	98.4 ± 2.7	
GEO	7	0.0 ± 0.0	0.6 ± 0.5	6.0 ± 0.6	0.0 ± 0.0	0.4 ± 0.8	85.7 ± 8.2	
ITA	7	0.0 ± 0.0	0.1 ± 0.4	0.0 ± 0.0	5.9 ± 0.7	1.0 ± 0.6	83.7 ± 9.9	
Total	40					1.7 ± 1.1	93.9 ± 2.4	

For all models, ANOVA p-value < 0.05.
[a] Model parameters: mean values obtained with the training sets from 7 iterations. LVs: latent variables of the training model; Q^2: Cumulative fraction of Y variation predicted by the X training model up to the specified latent variable, according to cross-validation; RMSEcv: root mean square error of the cross validation of the training model. Chile; ESP: Spain; GEO: Georgia; ITA: Italy.

In Fig. 1, the regression coefficients of the NIR PLS-DA models for cultivar (a) and origin (b) are displayed against the pre-processed and the raw spectra. For the cultivar model, the highest regression coefficients of the TG class corresponded to bands around 1482–1490 nm, 1588–1606 nm, and 2180–2220 nm. All these bands were related to protein functional groups. The first two could be assigned to the first

overtone of the N—H stretching of the peptide bond and amino acids side chains, while the last ones belonged to the combination of amide I and amide III bonds [21,38]. The most relevant coefficients for the non-TG class were assigned to the bands around 1952–1990 nm and 2220–2232 nm, which were related to primary amide groups and to the methylene group combination bands of fatty acids, respectively [20,38]. For the origin model, the most relevant coefficients for the ESP class corresponded to the following bands: 722–774 nm, belonging to an O—H stretching overtone, which can originate from alkyl, and primary alcohols, phenols or water [38]; 1640–1658 nm, assigned to the first C—H stretching overtone associated with a secondary alkyl group, which could be related to unsaturated fatty acids [19]; 1816–1844 nm, corresponding to a combination of second O—H/C—O stretching overtones of cellulose [38]; 2110–2150 nm related to combination bands of C—H/C═O stretching associated to lipids [38]; 2266–2270 nm, corresponding to a combination of O—H/C—O stretching of cellulose or to the peptide bond $CONH_2$ in β-sheet structures; and 2476–2482 nm, assigned to the combination bands of the C—H stretching in methyl group of lipid and aliphatic compounds [20,38]. For the non-ESP class, the most significant bands were: 582–615 nm, assigned to the fourth OH stretching overtone in alkyl alcohols [38]; 1516–1532 nm, related to the first NH stretching overtone of the amide group [38]; 1724–1728 nm, corresponding to first overtone of the C—H vibration of triolein [39,40]; 2192–2208 nm, associated to the combination of amide I and amide III bonds [21,38]; 2226–2238 nm and 2440–2444, both assigned to combination bands of the C—H stretching in methyl group of lipidic compounds [20].

Although some of the main discriminant bands coincided in both cultivar and origin models, such as the band around 2180–2220 nm (a combination band associated to amide bonds) and that around 2220–2240 nm (related to methyl groups in lipid compounds), in general, the most relevant variables for each discriminant model were different. The protein-related bands, together with those associated to lipids, were the most relevant variables to distinguish the TG cultivar. These findings are consistent with previous studies [20,21], which reported that the most influential spectral bands in hazelnut cultivar classification corresponded to protein and lipid compounds.

On the other hand, the most significant bands for discriminating the geographical origin were related to a wider variety of compounds: mainly lipids (1640–1658, 1724–1728, 2110–2150, 2226–2238 and 2440–2482 nm), proteins (1516–1532, 2192–2208 and 2266–2270 nm), complex carbohydrates (1816–1844 and 2266–2270 nm) and a few bands associated to compounds containing hydroxy groups (722–774 nm). These results agreed with previous research [22,23] reporting similar findings in Italian and Georgian hazelnuts and in samples from other origins (France, Germany and Turkey).

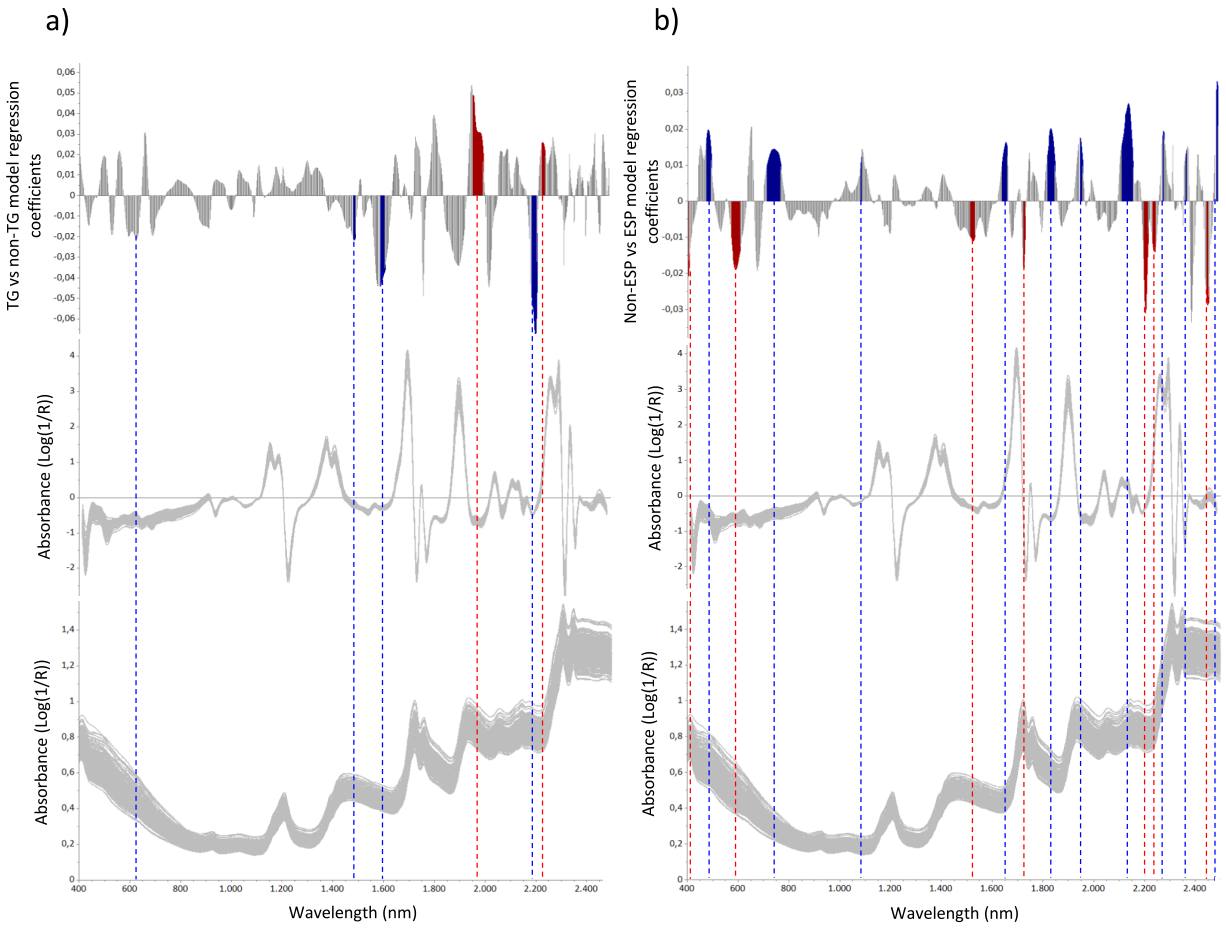


Fig. 1. Regression coefficients of the PLS-DA models developed on ground samples analysed by NIR benchtop spectrometer; a) cultivar model ('Tonda di Giffoni' TG vs non-TG) b) origin model ('Spain' ESP vs non-ESP). Regression coefficients (above) are plotted against the pre-processed (middle) and raw spectra (below). For each model, the most relevant coefficients for the prediction of the TG and ESP classes are highlighted in blue and those relevant for non-TG, non-ESP are highlighted in red.

3.2.2. MIR spectroscopy

Fig. 2 shows the regression coefficients of the MIR PLS-DA cultivar (a) and origin (b) models against the corresponding pre-processed and raw spectra. In the cultivar model, some of the most significant regression coefficients, for both TG and non-TG classes, corresponded to bands in the low-frequency region (600–900 cm^{-1}). This region is associated to N—H wagging of primary and secondary amines and amides [41]. The highest regression coefficients of both classes corresponded to the band around 900–1200 cm^{-1}, which belonged to the C—O stretching of the ester groups in triacylglycerols [41,42]. For the non-TG class, several relevant regression coefficients corresponded to a broad interval with numerous bands around 1230–1470 cm^{-1} that were assigned to C—H bending of CH_2 and CH_3 of lipids, and to the bands in the region 1438–1480 cm^{-1}, which are characteristic of the unsaponifiable fraction compounds [42–44]. Finally, the band around 1500–1570 cm^{-1}, associated with amide II vibrations, arises from mixed N—H bending and C—N stretching vibration in protein structures [24,45].

For the origin model, the most significant coefficients are in the 1000–1750 cm^{-1} range. For the ESP class, the most discriminant bands were: the broad interval with numerous bands associated to CH_2 bending of lipids and unsaponifiable fraction compounds (1300–1470 cm^{-1}) [43]; the amide II vibration band around 1500–1570 cm^{-1} related to protein structures; the amide I band associated C═O stretching of protein amides (1600–1670 cm^{-1}) and the narrow band at 1710–1780 cm^{-1} corresponding to C═O stretching of ester groups of triglycerides [24,45]. Two of these bands, the broad CH_2 band around 1300–1470 cm^{-1} and the amide I band (1600–1670 cm^{-1}), were also relevant for the discrimination of the non-ESP class. Additionally, the band around 900–1200 cm^{-1} assigned to the C-O stretching of the ester groups in triacylglycerols was particularly significant for the non-ESP class.

In general, for both types of models, the most discriminating variables were associated to protein, lipid and unsaponifiable fraction compounds and were found in the 800–1800 cm^{-1} region of the spectral fingerprint. The absorption pattern in this region is complex but contains valuable information, as it is highly specific for each molecular species. In the cultivar models, the low frequency region of the spectra also appeared to be relevant for discrimination. In contrast, in the origin model, the final section of the fingerprint region, between 1300 and 1700 cm^{-1}, was the most significant for discrimination.

Unlike the NIR method, which seemed to find proteins as the most discriminant compounds between TG and non-TG classes, the MIR method relied mainly on lipids, including both triacylglycerols and unsaponifiable fraction compounds, to achieve this discrimination. Regarding the geographical discrimination, the NIR model identified bands related to unsaturated fatty acids, other lipid compounds and complex carbohydrates such as cellulose as very distinctive for the ESP category, whereas in the MIR models, triacylglycerols and unsaponifiable fraction compounds were more relevant in distinguishing between origins.

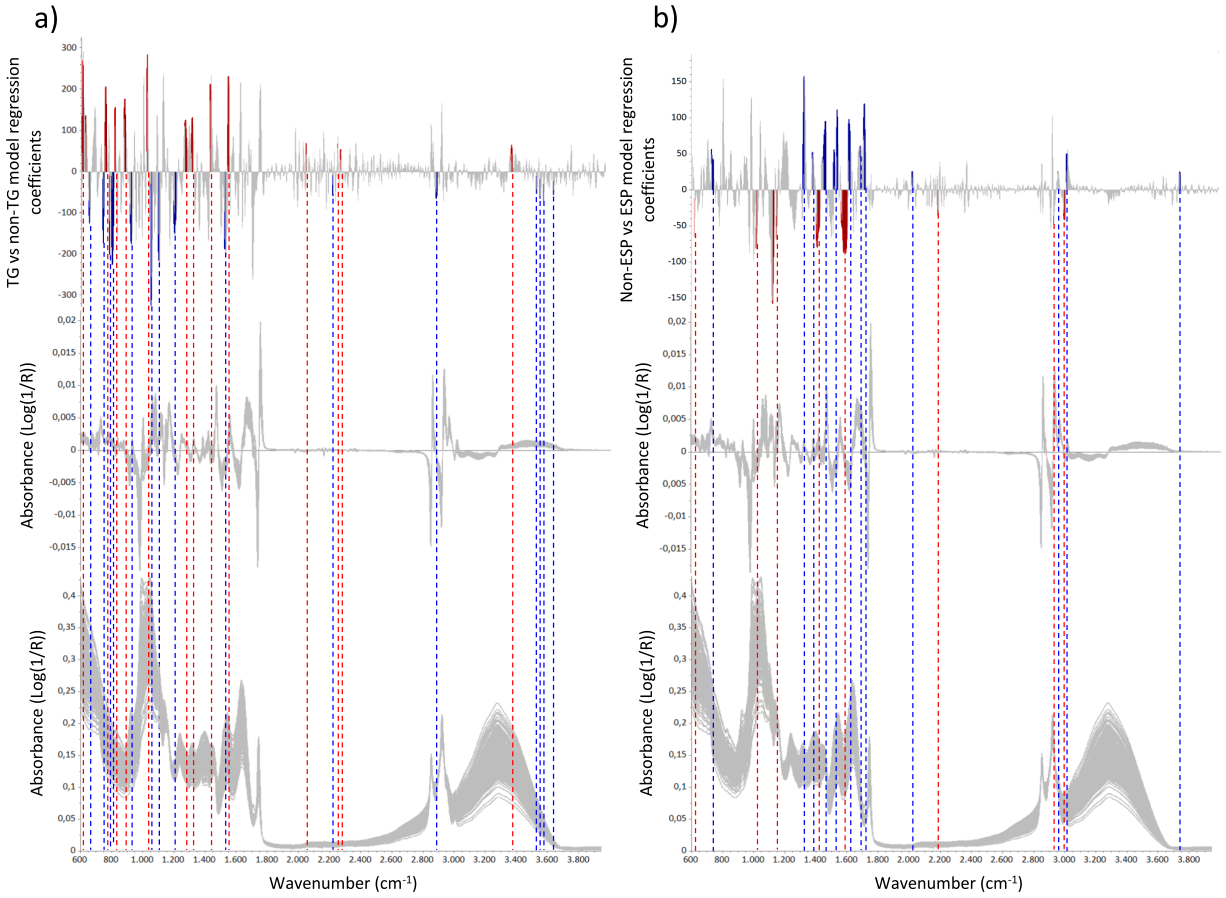


Fig. 2. Regression coefficients of the PLS-DA models developed on ground samples analysed by MIR spectrometer; (a) cultivar model ('Tonda di Giffoni' TG vs non-TG) (b) origin model ('Spain' ESP vs non-ESP). Regression coefficients (above) are plotted against the pre-processed (middle) and raw spectra (below). For each model, the most relevant coefficients for the prediction of the TG and ESP classes are highlighted in blue and those relevant for non-TG, non-ESP are highlighted in red.

4. Conclusions

Three different spectroscopic methods were tested to authenticate hazelnut cultivar and geographical origin: NIR spectroscopy using a benchtop instrument, NIR spectroscopy using a handheld device, and MIR spectroscopy. The analysis of ground hazelnuts yielded significantly better results than whole kernels, owing to the greater homogeneity, improved sample representativeness, and the prominence of oil content signals. Among the analysis conducted on the ground samples, the benchtop NIR spectrometer demonstrated superior performance, with a sensitivity of 0.92 and a specificity of 0.98 for cultivar models, as well as high correct classification rates for all origins (≥91 %). This resulted in overall correct classification rates of 95 % and 96 %, for cultivar and origin models respectively, closely followed by the MIR method.

Exploring the regression coefficients of the most promising models, based on MIR and NIR applied to ground samples, revealed their reliance on distinct sets of information for discrimination. It was observed that the discrimination of hazelnut cultivar and origin was mainly driven by proteins and lipid composition.

In conclusion, this study allowed for a straightforward comparison of three spectroscopic techniques that offer valuable insights into their performance when applied to exactly the same dataset of hazelnuts from different origins and cultivars. The present work showed that the NIR method could be a fit-for-purpose tool for hazelnut geographical and varietal authentication. However, optimal models need to be further developed and evaluated through extensive datasets, including higher natural heterogeneity of samples, producing regions, main cultivars and multiple harvest years.

5. Research data

Torres-Cobos, B., Tres, A., Vichi, S., Guardiola, F., Rovira, M., Romero, A., Baeten, V., & Fernández-Pierna, J.A. (2024). Near Infrared (NIR) and Mid Infrared (MIR) spectra of whole and ground hazelnuts [dataset]. CORA. Repositori de Dades de Recerca, https://doi.org/10.34810/data1725.

CRediT authorship contribution statement

B. Torres-Cobos: Formal analysis, Investigation, Methodology, Validation, Data curation, Visualization, Writing – original draft. **A. Tres:** Conceptualization, Writing – review & editing. **S. Vichi:** Conceptualization, Resources, Project administration, Funding acquisition, Writing – review & editing. **F. Guardiola:** Writing – review & editing. **M. Rovira:** Resources, Writing – review & editing. **A. Romero:** Resources, Writing – review & editing. **V. Baeten:** Conceptualization, Methodology, Supervision, Writing – review & editing. **J.A. Fernández-Pierna:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Funding

This work was developed in the context of the project TRACENUTS, PID2020-117701RB-I00 supported by MICIU/AEI/10.13039/501100011033. B. Torres-Cobos thanks the Spanish Ministry of Universities predoctoral fellowships FPU20/01454 and EST23/00155.

Declaration of competing 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.

Acknowledgements

The authors would like to thank all who participated in the project: François Stevens for data analysis, Quentin Arnould, Delphine Delhotte, Louis Paternostre, Stéphane Brichard and Sandrine Mauro for formal analysis, Antoine Deryck for data analysis (all from the Walloon Agricultural research Centre (CRA-W) in Gembloux, Belgium). INSA-UB Maria de Maeztu Unit of Excellence (Grant CEX2021- 001234-M) funded by MICIU/AEI/ERDF, EU. The authors would like to express their gratitude to Ferrero Hazelnut Company and Tuscia University (Department of Agriculture and Forest Science) for providing the hazelnut samples from Chile and Georgia, and Italy, respectively.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.saa.2024.125367.

Data availability

Data is published in a repository: Torres-Cobos, B., Tres, A., Vichi, S., Guardiola, F., Rovira, M.,Romero, A., Baeten, V., & Fernández-Pierna, J. A. (2024). Near Infrared(NIR) and Mid Infrared (MIR) spectra of whole and ground hazelnuts[dataset]. CORA. Repositori de Dades de Recerca, https://doi.org/10.34810/data1725.

References

[1] International Nut and Dried Fruits (INC), Nuts & Dried Fruits Statistical Yearbook 2022/2023, https://inc.nutfruit.org/publications/ (accessed Apr 10, 2024).
[2] V. Cristofori, S. Ferramondo, G. Bertazza, C. Bignami, Nut and kernel traits and chemical composition of hazelnut (*Corylus avellana* L.) cultivars, Sci. Food Agric. 88 (2008) 1091–1098, https://doi.org/10.1002/jsfa.3203.
[3] K. Król, M. Ganter, A. Piotrowska, Morphological traits, kernel composition and sensory evaluation of hazelnut (*Corylus avellana* L.) cultivars grown in Poland, Agronomy 9 (2019) 703, https://doi.org/10.3390/agronomy9110703.
[4] E. Cazzaniga, N. Cavallini, A. Giraudo, G. Gavoci, F. Geobaldo, M. Pariani, D. Ghirardello, G. Zeppa, F. Savorani, Lipids in a nutshell: quick determination of lipid content in hazelnuts with NIR spectroscopy, Foods 12 (2022) 34, https://doi.org/10.3390/foods12010034.
[5] Food and Agriculture Organization of the United Nations (FAO), agricultural data 2022, https://www.fao.org/faostat/es/#data/PP (accessed Mar 23, 2024).
[6] L.F. Ciarmiello, M.F. Mazzeo, P. Minasi, A. Peluso, A. De Luca, P. Piccirillo, R. A. Siciliano, V. Carbone, Analysis of different European hazelnut (*Corylus avellana* L.) cultivars: authentication, phenotypic features, and phenolic profiles, J. Agric. Food Chem. 62 (2014) 6236–6246, https://doi.org/10.1021/jf5018324.
[7] K. Król, M. Gantner, Morphological traits and chemical composition of hazelnut from different geographical origins: a review, Agriculture 10 (2020) 375, https://doi.org/10.3390/agriculture10090375.
[8] C. Lang, N. Weber, M. Möller, L. Schramm, S. Schelm, O. Kohlbacher, M. Fischer, Genetic authentication: differentiation of hazelnut cultivars using polymorphic sites of the chloroplast genome, Food Control 130 (2021) 108344, https://doi.org/10.1016/j.foodcont.2021.108344.
[9] T. Giulia, G. Vallauri, V. Pavese, N. Valentini, P. Rufa, R. Botta, D. Torello-Marinoni, Identification of the hazelnut cultivar in raw kernels and in semi-processed and processed products, Eur. Food Res. Technol. 248 (2022) 2431–2440, https://doi.org/10.1007/s00217-022-04058-z.
[10] S. Klockmann, E. Reiner, R. Bachmann, T. Hackl, M. Fischer, Food fingerprinting: metabolomic approaches for geographical origin discrimination of hazelnuts (*Corylus avellana*) by UPLC-QTOF-MS, J. Agric. Food Chem. 64 (2016) 9253–9262, https://doi.org/10.1021/acs.jafc.6b04433.
[11] S. Ghisoni, L. Lucini, G. Rocchetti, G. Chiodelli, D. Farinelli, S. Tombesi, M. Trevisan, Untargeted metabolomics with multivariate analysis to discriminate hazelnut (*Corylus avellana* L.) cultivars and their geographical origin, J. Sci. Food Agric. 100 (2020) 500–508, https://doi.org/10.1002/jsfa.9998.
[12] B. Torres-Cobos, B. Quintanilla-Casas, M. Rovira, A. Romero, F. Guardiola, S. Vichi, A. Tres, Prospective exploration of hazelnut's unsaponifiable fraction for geographical and varietal authentication: a comparative study of advanced fingerprinting and untargeted profiling techniques, Food Chem. 441 (2024) 138294, https://doi.org/10.1016/j.foodchem.2023.138294.
[13] R. Bachmann, S. Klockmann, J. Haerdter, M. Fischer, T. Hackl, 1H NMR spectroscopy for determination of the geographical origin of hazelnuts, J. Agric. Food Chem. 66 (2018) 11873–11879, https://doi.org/10.1021/acs.jafc.8b03724.
[14] R. Vitale, M. Bevilacqua, R. Bucci, A.D. Magrì, A.L. Magrì, F. Marini, A rapid and non-invasive method for authenticating the origin of pistachio samples by NIR spectroscopy and chemometrics, Chemom. Intell. Lab. Syst. 121 (2013) 90–99, https://doi.org/10.1016/j.chemolab.2012.11.019.

[15] P. Firmani, R. Bucci, F. Marini, A. Biancolillo, Authentication of "Avola almonds" by near infrared (NIR) spectroscopy and chemometrics, J. Food Compos. Anal. 82 (2019) 103235, https://doi.org/10.1016/j.jfca.2019.103235.

[16] L. Amendola, P. Firmani, R. Bucci, F. Marini, A. Biancolillo, Authentication of sorrento walnuts by NIR spectroscopy coupled with different chemometric classification strategies, Appl. Sci. 10 (11) (2020) 4003, https://doi.org/10.3390/app10114003.

[17] H. Zhu, J. Xu, Authentication and provenance of walnut combining Fourier transform mid-infrared spectroscopy with machine learning algorithms, Molecules 25 (21) (2020) 4987, https://doi.org/10.3390/molecules25214987.

[18] M. Arndt, M. Rurik, A. Drees, C. Ahlers, S. Feldmann, O. Kohlbacher, M. Fischer, Food authentication: determination of the geographical origin of almonds (*Prunus dulcis* Mill.) via near-infrared spectroscopy, Microchem. J. 160 (B) (2021) 105702, https://doi.org/10.1016/j.microc.2020.105702.

[19] R. Moscetti, E. Radicetti, D. Monarca, M. Cecchini, R. Massantini, Near infrared spectroscopy is suitable for the classification of hazelnuts according to Protected Designation of Origin, J. Sci. Food Agric. 95 (13) (2015) 2619–2625, https://doi.org/10.1002/jsfa.6992.

[20] A. Biancolillo, S. De Luca, S. Bassi, L. Roudier, R. Bucci, A.D. Magrì, F. Marini, Authentication of an Italian PDO hazelnut ("Nocciola Romana") by NIR spectroscopy, Environ. Sci. Pollut. Res. 25 (2018) 28780–28786, https://doi.org/10.1007/s11356-018-1755-2.

[21] H. Ayvaz, R. Temizkan, H.E. Genis, M. Mortas, D.O. Genis, M.A. Dogan, B. A. Nazlim, Rapid discrimination of Turkish commercial hazelnut (*Corylus avellana* L.) varieties using Near-Infrared Spectroscopy and chemometrics, Vib. Spectrosc. 119 (2022) 103353, https://doi.org/10.1016/j.vibspec.2022.103353.

[22] N. Shakiba, A. Gerdes, N. Holz, S. Wenck, R. Bachmann, T. Schneider, S. Seifert, M. Fischer, T. Hackl, Determination of the geographical origin of hazelnuts (*Corylus avellana* L.) by Near-Infrared spectroscopy (NIR) and a Low-Level Fusion with nuclear magnetic resonance (NMR), Microchem. J. 174 (2022) 107066, https://doi.org/10.1016/j.microc.2021.107066.

[23] G. Sammarco, C. Dall'Asta, M. Suman, Near infrared spectroscopy and multivariate statistical analysis as rapid tools for the geographical origin assessment of Italian hazelnuts, Vib. Spectrosc. 126 (2023) 103531, https://doi.org/10.1016/j.vibspec.2023.103531.

[24] M. Manfredi, E. Robotti, F. Quasso, E. Mazzucco, G. Calabrese, E. Marengo, Fast classification of hazelnut cultivars through portable infrared spectroscopy and chemometrics, Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 189 (2018) 427–435, https://doi.org/10.1016/j.saa.2017.08.050.

[25] L.E. Rodriguez-Saona, M.E. Allendorf, Use of FTIR for rapid authentication and detection of adulteration of food, Annu. Rev. Food Sci. Technol. 2 (2011) 467–483, https://doi.org/10.1146/annurev-food-022510-133750.

[26] V. Bellon-Maurel, A. McBratney, Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils – critical review and research perspectives, Soil Biol. Biochem. 43 (7) (2011) 1398–1410, https://doi.org/10.1016/j.soilbio.2011.02.019.

[27] G.A. de Oliveira, F. de Castilhos, C.M.G.C. Renard, S. Bureau, Comparison of NIR and MIR spectroscopic methods for determination of individual sugars, organic acids and carotenoids in passion fruit, Food Res. Int. 60 (2014) 154–162, https://doi.org/10.1016/j.foodres.2013.10.051.

[28] J. Nogales-Bueno, L. Feliz, B. Baca-Bocanegra, J.M. Hernández-Hierro, F. J. Heredia, J.M. Barroso, A.E. Rato, Comparative study on the use of three different near infrared spectroscopy recording methodologies for varietal discrimination of walnuts, Talanta 206 (2020) 120189, https://doi.org/10.1016/j.talanta.2019.120189.

[29] C. McVey, T.F. McGrath, S.A. Haughey, C.T. Elliott, A rapid food chain approach for authenticity screening: the development, validation and transferability of a chemometric model using two handheld near infrared spectroscopy (NIRS) devices, Talanta 222 (2021) 121533, https://doi.org/10.1016/j.talanta.2020.121533.

[30] Y. Pu, D. Pérez-Marín, N. O'Shea, A. Garrido-Varo, Recent advances in portable and handheld NIR spectrometers and applications in milk, cheese and dairy powders, Foods 10 (2021) 2377, https://doi.org/10.3390/foods10102377.

[31] J. Müller-Maatsch, S.M. van Ruth, Handheld devices for food authentication and their applications: a review, Foods 10 (2021) 2901, https://doi.org/10.3390/foods10122901.

[32] G.S. Folli, L.P. Santos, F.D. Santos, P.H.P. Cunha, I.F. Schaffel, F.T. Borghi, I.H.A. S. Barros, A.A. Pires, A.V.F.N. Ribeiro, W. Romão, P.R. Filgueiras, Food analysis by portable NIR spectrometer, Food Chem. Adv. 1 (2022) 100074, https://doi.org/10.1016/j.focha.2022.100074.

[33] D. Sorak, L. Herberholz, S. Iwascek, S. Altinpinar, F. Pfeifer, H.W. Siesler, New developments and applications of handheld Raman, mid-Infrared, and near-Infrared spectrometers, Appl. Spectrosc. Rev. 47 (2) (2012) 83–115, https://doi.org/10.1080/05704928.2011.625748.

[34] S.N. Thennadil, M. Dewar, C. Herdsman, A. Nordon, E. Becker, Automated weighted outlier detection technique for multivariate data, Control Eng. Pract. 70 (2018) 40–49, https://doi.org/10.1016/j.conengprac.2017.09.018.

[35] L. Eriksson, J. Trygg, S. Wold, CV-ANOVA for significance testing of PLS and OPLS® models, J. Chemom. 22 (2008) 594–600, https://doi.org/10.1002/cem.1187.

[36] B. Magnusson, U. Örnemark, Eurachem guide: The Fitness for Purpose of Analytical Methods - A Laboratory Guide to Method Validation and Related Topics, second edition, Eurachem, Belgium, 2014.

[37] M. Arndt, M. Rurik, A. Drees, K. Bigdowski, O. Kohlbacher, M. Fischer, Comparison of different sample preparation techniques for NIR screening and their influence on the geographical origin determination of almonds (*Prunus dulcis* MILL.), Food Control 115 (2020) 107302, https://doi.org/10.1016/j.foodcont.2020.107302.

[38] J. Workman, L. Weyer, Practical Guide to Interpretive Near-infrared Spectroscopy, first edition, CRC Press, Boca Raton, Florida, 2007, doi: 10.1201/9781420018318.

[39] T. Sato, S. Kawano, M. Iwamoto, Near infrared spectral patterns of fatty acid analysis from fats and oils, J. Am. Oil Chem. Soc. 68 (1991) 827–833, https://doi.org/10.1007/BF02660596.

[40] V. Baeten, R. Aparicio, N. Marigheto, R. Wilson, Olive oil analysis by infrared and Raman spectroscopy: methodologies and applications, in: J. Harwood, R. Aparicio (Eds.), Handbook of Olive Oil: Analysis and Properties, Springer, Boston, MA, 2000, pp. 209–248, https://doi.org/10.1007/978-1-4757-5371-4_8.

[41] R.M. Silverstein, F.X. Webster, Spectrometric Identification of Organic Compounds, sixth edition, John Wiley & Sons, New Jersey, 1998, ISBN978-0471134572.

[42] S. Ng, O. Lasekan, K. Muhammad, R. Sulaiman, N. Hussain, Effect of roasting conditions on color development and Fourier transform infrared spectroscopy (FTIR-ATR) analysis of Malaysian-grown tropical almond nuts (*Terminalia catappa* L.), Chem. Cent. J. 8 (2014) 55, https://doi.org/10.1186/s13065-014-0055-2.

[43] V. Baeten, J.A. Fernández Pierna, P. Dardenne, M. Meurens, D.L. García-González, R. Aparicio-Ruiz, Detection of the presence of hazelnut oil in olive oil by FT-Raman and FT-MIR spectroscopy, J. Agric. Food Chem. 53 (16) (2005) 6201–6206, https://doi.org/10.1021/jf050595n.

[44] A. Valdés García, A. Beltrán-Sanahuja, M.C. Garrigós-Selva, Characterization and classification of almond cultivars by using spectroscopic and thermal techniques, J. Food Sci. 78 (2) (2013) 138–144, https://doi.org/10.1111/1750-3841.12031.

[45] A. Dogan, G. Siyakus, F. Severcan, FTIR spectroscopic characterization of irradiated hazelnut (*Corylus avellana* L.), Food Chem. 100 (2007) 1106–1114, https://doi.org/10.1016/j.foodchem.2005.11.017.

CHAPTER 6.

PINE NUT GEOGRAPHICAL AND VARIETAL AUTHENTICATION

This **chapter 6** deals with the **authentication of pine nuts species and origin through isotopic and metabolic methods**, as defined in **objectives 5** and **6**.

Although isotopic [23,93,88-91,247] and terpene markers [113,136,137,140,141,199, 232] have shown promising results in the authentication of plant-based foods, they have not yet been explored for verifying the provenance and species of pine nuts. These markers could hold valuable information and serve as fit-for-purpose tools for pine nuts botanical and geographical authentication.

6.1 Development of isotopic methods for pine nut geographical authentication: Proof-of-concept study

6.1.1 Sr isotopic ratios and Sr and Rb elemental analysis

A preliminary study was conducted on a small but representative sample set (**Figure 6**), encompassing multiple origins, producers, harvest years, and species to assess the feasibility of using isotopic ratios as authenticity markers. The Sr isotope ratio was chosen as a potential marker because it strongly correlates with the geological and geographical characteristics of the pine nut growing area and is minimally affected by fertilizers in natural forest crops such as pine nuts. The Sr and Rb content were also determined to optimise the isotopic analysis, ensuring that the Sr concentration was sufficient for detection, and that Rb was effectively removed during the purification process, as ^{87}Rb interferes with the measurement of ^{87}Sr.

The Sr and Rb content and the ^{87}Sr/^{86}Sr ratio were analysed in 96 samples collected over three consecutive harvest years (2020-2022) from four countries of origin. Among these 96 samples:

- 60 samples were from two regions in Spain (ESP): Catalonia (CAT, n = 30) and Central Spain: Castile and Leon and Madrid (CS, n = 30).
- 36 samples were commercial samples from China (CHN, n= 15), Russia (RUS, n = 15) and Turkey (TUR, n= 6) (**Figure 15**).

Differences in the Sr and Rb content and Sr isotopic ratios across various origins (individual countries, or Spanish regions) and years (2020, 2021 or 2022) were studied using statistical tests. For data following a normal distribution (determined by the Shapiro-Wilk test), independent-samples t-tests and one-way ANOVA were employed. For data that did not follow a normal distribution, the Mann-Whitney U and Kruskal-

Wallis tests were used. Dispersion plots combining Sr and Rb concentration and Sr isotopic ratio were also explored to evaluate sample clustering and discrimination according to the different regions of origin.

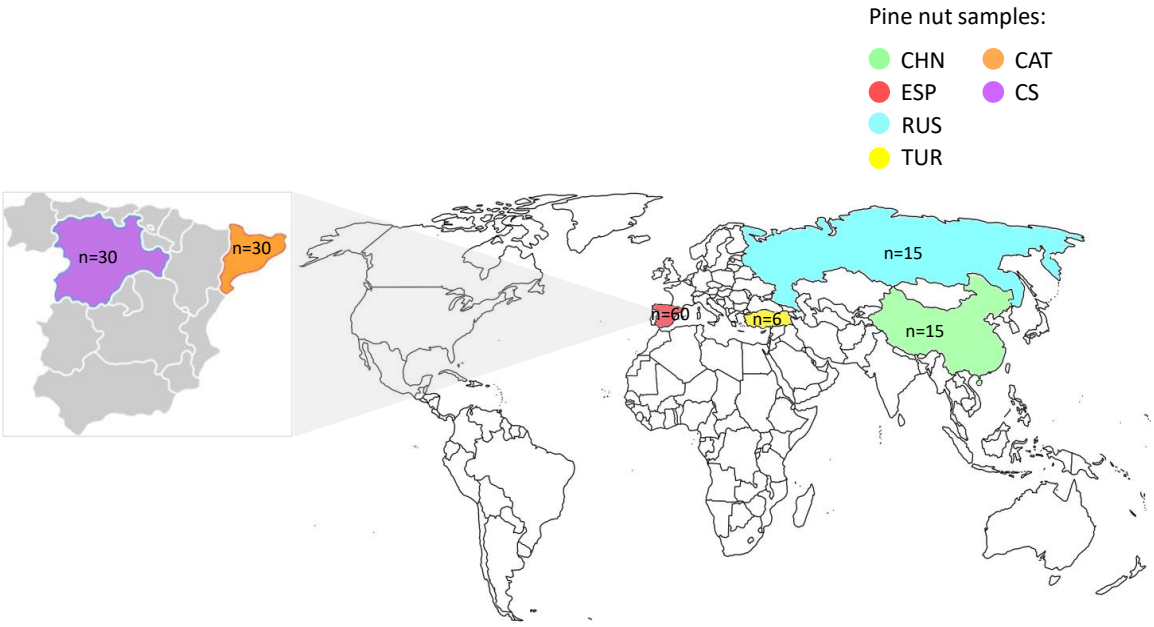


Figure 15. Pine nut samples of the PoC study on strontium isotope ratio for pine nut geographical authentication. ESP: Spain, CHN: China, RUS: Russia, TUR: Turkey, CAT: Catalonia, CS: Central Spain (Publication 10).

6.1.2 Publication 10

Evaluating Sr isotope ratios, and Sr and Rb elemental analysis for pine nuts geographical authentication

Berta Torres-Cobos, Mònica Rosell, Albert Soler, Soriana Beatrice Nicotra, Neus Aletà, Anna Teixidó, Mercè Rovira, Agustí Romero, Francesc Guardiola, Alba Tres, Stefania Vichi

Manuscript under preparation 2024

Evaluating Sr isotope ratios, and Sr and Rb elemental analysis for pine nuts geographical authentication

B. Torres-Cobos[1,2], M. Rosell[3], A. Soler[3], S. Nicotra[1,2], N. Aletà[4,5], A. Teixidó[5], M. Rovira[6], A. Romero[6], F. Guardiola[1,2], A. Tres[1,2], S. Vichi[1,2,*]

[1] Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Universitat de Barcelona. Av Prat de La Riba, 171, 08921, Santa Coloma de Gramenet, Spain.

[2] Institut de Recerca en Nutrició i Seguretat Alimentària (INSA-UB), Universitat de Barcelona. Av Prat de La Riba, 171, 08921, Santa Coloma de Gramenet, Spain.

[3] Grup MAiMA, Mineralogia Aplicada, Geoquímica i Hidrogeologia, Departament de Mineralogia, Petrologia i Geologia Aplicada, Institut de Recerca de l'Aigua (IdRA), Universitat de Barcelona, Martí i Franqués s/n, 08028, Barcelona, Spain.

[4]Institute of Agrifood Research and Technology (IRTA). Fruit Tree Program. Torre Marimon. 08140 Caldes de Montbui, Spain

[5] Forest Science and Technology Centre (CTFC). Multifunctional Forest Management Program. 25280 Solsona, Lleida, Spain.

[6] Institute of Agrifood Research and Technology (IRTA), Ctra. de Reus – El Morell Km 3.8, Constantí, 43120, Spain.

*Corresponding author: Stefania Vichi, PhD, Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Campus de l'Alimentació Torribera, Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona. Av Prat de la Riba, 171. Edifici Gaudí, 08921 Santa Coloma de Gramenet, Spain.

ABSTRACT

In this study, Sr isotope ratios and Sr and Rb concentrations of pine nuts are evaluated as authentication markers to verify the geographical origin of pine nuts. Recently, the global pine nut market has grown significantly, driven by their nutritional benefits and unique flavour. Main commercial species include *Pinus koraiensis* from China, *Pinus sibirica* from Russia, and *Pinus pinea* from the Mediterranean, which commands the highest prices. This high value and price variation, make pine nuts susceptible to fraud, underscoring the need for effective geographical authentication.

Stable isotope analysis is a valuable tool for determining the geographical origin of agrifood products, as isotope ratios are influenced by geological and pedoclimatic conditions. Sr isotope ratios, which vary with the age and composition of the bedrock, provide key insights into food provenance. Alternatively, elemental analysis, widely applied to verify plant food origins, examines the relationship between soil mineral profiles and plant composition, with Sr and Rb concentrations being particularly useful for geographical origin differentiation.

This preliminary study explores the potential of Sr isotope ratios and Sr and Rb elemental analysis to distinguish between Mediterranean and Asian pine nuts, and between two Spanish regions. The results of this study reveal that while Sr isotopic ratios are insufficient to verify pine nuts origin, the combination of Sr and Rb exhibits significant potential for geographical authentication, providing a foundation for developing reliable authentication methods.

Keywords: Pine nuts; Authenticity; Strontium isotope; Elemental strontium and rubidium; Geographical origin.

1. Introduction

In recent years, the global market for pine nuts has experienced significant growth (Awan et al., 2017; Evaristo et al., 2013). These nuts are highly valued for their nutritional benefits, being rich in unsaturated fatty acids, vitamins, minerals, and for their distinctive flavour (Gonçalves et al., 2022; Takala et al., 2022). As a premium gourmet product that can command prices exceeding 100 EUR/kg based on their origin (Mutke, 2022), pine nuts are particularly vulnerable to fraudulent practices.

The primary commercial species of pine nuts include the Chinese *Pinus koraiensis* Siebold & Zucc. (*P. koraiensis*); the *Pinus sibirica* (*P. sibirica*), which is primarily grown in Russia, and the *Pinus pinea* L. (*P. pinea*), sourced from the Mediterranean basin, with Turkey, Spain, and Portugal being the main producers (Awan et al., 2017; Moscetti et al., 2021). Each of these species and geographical provenances possesses distinct characteristics that influence both quality and market value (Awan et al., 2017; Evaristo et al., 2013; Mutke, 2022). The Mediterranean *P. pinea* pine nuts reach the highest prices, often doubling or even tripling the price of other species and provenances (Evaristo et al., 2013; Moscetti et al., 2021; Mutke, 2022; Ríos-Reina et al., 2021).

According to EU Regulation No 2023/2429, declaring the geographical origin of pine nuts is mandatory to safeguard consumer interests and promote trust and transparency within the food supply chain. In this context, verifying the authenticity of the declared geographical origin is crucial for consumer protection, quality assurance, and the prevention of fraudulent practices in the global market.

Over the past few years, stable isotope analysis has emerged as a powerful approach for determining the provenance of agrifood products (Camin et al, 2017; Kelly et al., 2005; Pustjens et al., 2016). The success of isotopic methods for geographical authentication relies on the direct influence of geological and pedoclimatic characteristics of the growing region on the isotope ratios present in the food commodity (Kelly et al., 2005; Laursen et al., 2016). The light bioelements, C, H, N, O and S, along with the heavy geo element Sr, are among the most commonly analysed elements for food authentication purposes (Drivelos et al., 2012).

The isotopic ratios of these elements are intricately linked to the environmental conditions of each geographical area, but in distinct ways. Light bioelements undergo

fractionation—shifts in the ratio of 'heavy' to 'light' isotopes—due to natural processes such as water evaporation, condensation, and plant photosynthesis. These processes are influenced by geographical factors such as altitude, latitude, and proximity to the sea, making light bioelements valuable indicators of the geographical origin of agrifood products. However, they may be less effective in distinguishing between samples from the same climate zone.

In contrast, the fractionation of Sr isotopic ratios is minimal due to the high mass of its nuclides relative to the differences in isotope masses (Drivelos et al., 2012). The radiogenic isotope ^{87}Sr is produced from the decay of ^{87}Rb, leading to variations in the $^{87}Sr/^{86}Sr$ ratio based on the age and Rb/Sr ratio of the underlying rock. The Sr isotopic composition in plants and animals primarily reflects the bioavailable Sr from the bedrock beneath them. The variation in $^{87}Sr/^{86}Sr$ ratios across different geological formations—shaped by their unique Rb/Sr ratios and geological ages—makes Sr isotopic signatures particularly useful for verifying the geographical authenticity of natural food products. This geological insight is especially valuable when the stable isotopes of lighter elements are insufficient for differentiating between samples from similar climate zones (Kelly et al., 2005).

On the other hand, elemental analysis has been widely reported as a powerful tool for the geographical authentication of foods of plant origin (D'Archivio et al., 2014; Di Paola-Naranjo et al., 2011; Epova et al., 2019; Griboff et al., 2019; Horacek, et al., 2023; García-Ruiz et al., 2007). The elemental composition of plants reflects the bioavailable portion of mineral elements present in the soil where they grow. The availability of trace elements is influenced by factors, such as soil pH, moisture, porosity, and the presence of clay and humic substances (Kim et al., 1993). As a result, the diversity of soils and their unique bioavailability profiles can lead to distinctive elemental compositions in food, which serve as reliable markers for identifying geographical origin. Among the inorganic elements studied for this purpose, trace elements like Sr and Rb have been identified as some of the most effective geographical discriminants (D'Archivio et al., 2014; Di Paola-Naranjo et al., 2011; Epova et al., 2019; Griboff et al., 2019; Horacek, et al., 2023; García-Ruiz et al., 2007).

Besides, unlike agricultural crops, where the Sr isotope ratios and modify Sr and Rb concentrations are modified through anthropogenic influences such as fertilitzation

(Ariyama et al., 2006; Bora et al., 2018; Danezis et al., 2022; Sen et al., 2014; Teacher et al., 2017; Vitòria et al., 2004), pine nuts are harvested from natural forests where fertilizers are not applied. This ensures that elemental and isotopic profiles of pine nuts accurately reflect the soil's geochemical signature. Moreover, Previous research has demonstrated the efficacy of Sr isotope ratios, and Sr and Rb elemental analysis in authenticating the geographical origin of various foodstuffs and beverages (Baffi et al., 2016; Drivelos et al., 2012; Epova et al., 2019; Fortunato et al., 2003; García-Ruiz et al., 2007; Geană et al., 2017; Asfaha, et al., 2011; Horacek, et al., 2023; Kelly et al, 2005). However, the application of these methods for pine nuts authentication remains underexplored.

The hypothesis of this preliminary study is that Sr isotope ratio, as well as Sr and Rb elemental analysis are suitable for the geographical authentication of pine nuts, a high-value commodity with significant economic and cultural importance in many regions.

Therefore, the primary objective of this study is to evaluate the effectiveness of Sr isotope ratios, along with Sr and Rb concentrations in discriminating between Mediterranean and Asian pine nuts, as well as between pine nuts of the same species from two distinct regions in Spain. By analysing samples from various regions and species, we aim to develop a robust analytical framework for verifying the authenticity of pine nuts. Additionally, this study seeks to compare the performance of the isotopic analysis with the elemental analysis to determine the most reliable approach for geographical authentication.

2. Material and methods

2.1 Sampling

A set of 96 pine nut samples was obtained between 2020 and 2022 in the framework of the Tracenuts project (PID2020-117701RB-I00). Among these, 60 samples were *P. pinea* kernels from Spain (ESP), specifically from two regions: Central Spain [Castile and Leon and Madrid (CS, n = 30)], and Catalonia (CAT, n = 30). These were traceable samples provided by the Institute of Forest Science (ICIFOR-INIA, Madrid, Spain), the “Centro de Servicios y Promoción Forestal y de su Industria de Castilla y León” (CESEFOR Foundation, Soria, Spain), and the Institute of Agrifood Research and

Technology (IRTA-Torre Marimon, Caldes de Montbui, Spain). The remaining 36 samples were commercial pine nuts sourced from China (CHN, n = 15), Russia (RUS, n = 15), and Turkey (TUR, n = 6) (**Table 1**). Based on the natural distribution of pine nut species, it was presumed that the samples from China and Russia were mainly from *P. koraiensis* and *Pinus sibirica*, which are the most commercially significant species in those countries (Awan et al., 2017). Commercial samples from Turkey likely included both *P. pinea* and other local species (Bonari et al., 2020). All samples were stored at 4°C until analysis.

Table 1. Geographical origin and harvest year of the 96 pine nut samples analysed.

	n	2020	2021	2022
Spain	60	20	16	24
Catalonia (CAT)	30	10	10	10
Central Spain (CS)	30	10	6	14
China (CHN)	15	0	15	0
Russia (RUS)	15	0	5	10
Turkey (TUR)	6	0	0	6

2.2 Elemental analysis of Sr and Rb, and isotopic analysis of $^{87}Sr/^{86}Sr$

The concentrations of Sr and Rb and the $^{87}Sr/^{86}Sr$ isotopic ratio were determined as detailed by Torres-Cobos et al., (2024). The procedure involved grinding a 10 g aliquot of the sample, followed by the digestion of 0.5 g of the ground sample with HNO_3. The quantification of Sr and Rb was carried out by Inductively Coupled Plasma-Mass Spectrometry (ICP-MS). The Sr/Rb ratio in the sample must be ≥ 500 to ensure that the ^{87}Rb does not interfere with the ^{87}Sr analysis. Therefore, for the Sr isotopic ratio, digested samples were purified with a Sr-SPEC resin. Given the high Rb concentrations in pine nuts, a second purification step was carried out to ensure effective removal of the ^{87}Rb interferent. However, since chemical purification cannot completely eliminate all Rb from the sample, a correction was applied by measuring ^{85}Rb to estimate the residual ^{87}Rb. This correction was only valid when the Rb concentration in the sample was low. Finally, the Sr isotopic composition was determined using a Plasma 3 Multi

Collector Inductively Coupled Plasma Mass Spectrometer (MC-ICP-MS) (Torres-Cobos et al., 2024).

2.3 Statistical analysis

The statistical analysis was performed in IBM SPSS Statistics v29.0© (IBM Corp., Armonk, New York, USA). Statistical tests were applied to study differences in isotopic ratios among multiple origins (individual countries or Spanish regions) or among harvest years (2020, 2021 and 2022).

First, to examine the population distribution of the samples, Shapiro-Wilk test was applied. Since the isotopic ratios did not follow a normal distribution or had fewer than 30 samples in the compared populations, the medians and quartiles were calculated. The independent samples median test (non-parametric) was used to determine if the Sr and Rb concentrations or the ^{87}Sr /^{86}Sr ratio varied according to the geographical origin (countries of origin or Spanish regions (CAT/CS) or between harvest years. When more than two groups were compared in the median test, differences between groups were established by pair-wise comparisons. In all cases, $p<0.05$ was considered significant. Finally, to explore the samples distribution: 1) Samples were plotted versus $^{87}Sr/^{86}Sr$ values and 2) log ([Sr]) was plotted versus log([Rb]). In both plots, samples were coloured according to their country of origin.

3. Results and discussion

First, the interannual variability of Sr ratio medians across different harvest years was evaluated for each studied region of origin. **Table 2** presents the median and quartile results for the ESP and RUS, as well as for ESP regions. CHN and TUR were not analysed since samples of only one harvest year were included within these classes. Significant differences were observed in the median $^{87}Sr/^{86}Sr$ ratio across different years for the ESP samples (**Table 2**). Specifically, within the two regions conforming the ESP class, the Sr ratio for CS samples from 2022 showed a significant deviation from those in 2020 and 2021. However, these differences are more likely attributed to the specific locations of the CS samples across the harvest years rather than to a dependence of Sr isotopic ratios on the harvest year. In fact, CS samples collected in 2020-2021 and 2022 were from a different location of CS.

To explore and evaluate the differences in the Sr ratio across various geographical origins, **Table 3** displays the median and quartile results for the four countries of origin (ESP, CHN, RUS, TUR), and the two Spanish regions (CAT, CS).

The results of the statistical tests indicate that the median $^{87}Sr/^{86}Sr$ of CNH and RUS were significantly different from ESP and TUR, while no statistical differences were found between ESP and TUR and between the two Spanish regions, CAT and CS.

Table 2. $^{87}Sr/^{86}Sr$ ratio of pine nuts harvested in different years from various geographical origins.

		2020		**2021**	
	n	Median	Q_1 - Q_3	Median	Q_1 - Q_3
ESP $^{87}Sr/^{86}Sr$	60	0.7108 [a]	0.7105 – 0.7114	0.7107 [a]	0.7105 – 0.7112
CAT $^{87}Sr/^{86}Sr$	30	0.7106 [a]	0.7104 – 0.7108	0.7107 [a]	0.7104 – 0.7107
CS $^{87}Sr/^{86}Sr$	30	0.7114 [a]	0.7110 – 0.7115	0.7113 [a]	0.7106 – 0.7119
RUS $^{87}Sr/^{86}Sr$	15	-	-	0.7098 [a]	0.7089 – 0.7099
		2022			
	n	Median	Q_1 - Q_3	p-value	
ESP $^{87}Sr/^{86}Sr$	60	0.7099 [b]	0.7099 – 0.7106	0.006 *[1]	
CAT $^{87}Sr/^{86}Sr$	30	0.7107 [a]	0.7105 – 0.7108	1.000 [1]	
CS $^{87}Sr/^{86}Sr$	30	0.7099 [b]	0.7098 – 0.7099	<0.001 *[1]	
RUS $^{87}Sr/^{86}Sr$	15	0.7095 [a]	0.7082 – 0.7102	0.608 [1]	

*p-value < 0.05 means are significantly different. [1] p-value obtained from independent samples median test. Differences between origins were noted as a > b > c. ESP: Spain; CAT: Catalonia; CS: Central Spain; RUS: Russia.

Table 3. $^{87}Sr/^{86}Sr$ ratio of pine nuts from different geographical origins: a) Countries of origin; and b) Spanish regions of origin.

a)

	ESP (n= 60)		**CHN** (n= 15)		
	Median	Q_1 - Q_3	Median	Q_1 - Q_3	
$^{87}Sr/^{86}Sr$	0.7107 [a]	0.7102 – 0.7108	0.7095 [b]	0.7087 – 0.7111	
	RUS (n=15)		**TUR** (n=6)		
	Median	Q_1 - Q_3	Median	Q_1 - Q_3	p-value
$^{87}Sr/^{86}Sr$	0.7097 [b]	0.7088 – 0.7099	0.7138 [a]	0.7098 – 0.7157	<0.001 *[1]

*p-value < 0.05 means are significantly different. [1] p-value obtained from independent samples median test. Differences between origins were noted as a > b. ESP: Spain; CHN: China; RUS: Russia; TUR: Turkey.

b)

	CAT (n= 30)		**CS** (n= 30)		
	Median	Q_1 - Q_3	Median	Q_1 - Q_3	p-value
$^{87}Sr/^{86}Sr$	0.7107 [a]	0.7104 – 0.7107	0.7105 [a]	0.7099 – 0.7114	0.796 [1]

*p-value < 0.05 means are significantly different. [1] p-value obtained from independent samples median test. Differences between regions were noted as a > b. CAT: Catalonia; CS: Central Spain.

However, despite the significant differences observed among most classes, the corresponding quartile values of Sr ratios revealed substantial overlaps between origins, hindering a clear identification of pine nut provenance. This overlapping is clearly visible when plotting the samples vs their Sr ratios (**Figure 1a**). In fact, the Sr ratios of all analysed pine nut samples ranged from 0.707 to 0.717, regardless of geographical origin.

This lack of clear differentiation may be attributed to the similarities certain regional lithology across different areas of the world. These results are consistent with other research highlighting the limitations of using $^{87}Sr/^{86}Sr$ ratio to verify the geographical origin of various agrifood products (Baffi et al., 2016; Bong et al., 2012; Durante et al., 2015; Epova et al., 2019). Studies have reported that bedrock type has a much greater impact on the Sr isotope ratios than soil age (Horacek, 2022a). Variations in bedrock geology within regions can hinder geographical distinctions due to substantial overlaps of $^{87}Sr/^{86}Sr$ values both within and between regions (Asfaha, et al., 2011; Horacek, et al., 2022a; Horacek, et al., 2022b). Consequently, the heterogeneous geology of large countries such as CHN, RUS, and ESP (FAO/UNESCO 1971-1981; Gutiérrez et al., 2013; Horacek, et al., 2022b) introduces substantial variability, affecting the ability to differentiate samples based on Sr isotope ratios. In the case of CHN and RUS, the commercial samples were sourced from different producers, potentially representing various regions within the country, each with its own distinct geological characteristics.

On the other hand, the concentrations of the mineral elements Sr and Rb yielded more promising results. Non-parametric independent samples median tests revealed significant differences in the median concentrations of these elements between most countries of origin. No significant differences were found in the median concentrations of Sr and Rb between samples from CHN and RUS (**Table 4**). Regarding the two Spanish

regions, median Sr concentrations were similar, significant in Rb concentrations, suggest that Rb could effectively differentiate between these two regions (**Table 4**).

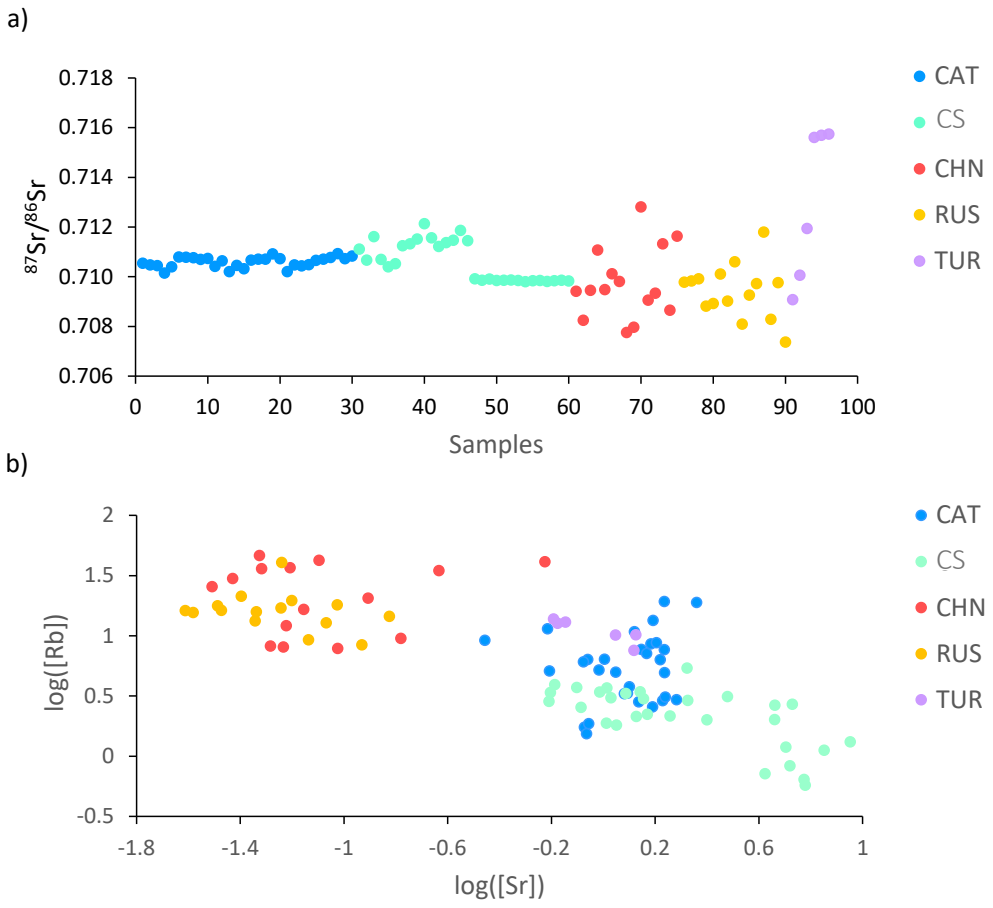


Figure 1. Dispersion plots of samples, color-coded by region of origin: Catalonia CAT (dark blue), Central Spain CS (light blue), China CHN (red), Russia RUS (yellow) and Turkey TUR (purple). (a) Plot of samples versus $^{87}Sr/^{86}Sr$ values in pine nuts. b) Plot of log([Sr]) versus log([Rb]) values in pine nuts.

These findings are further corroborated by the log([Sr]) vs. log([Rb]) plot (**Figure 1b**), which shows distinct clustering among samples from different origins. Despite the variability in the Sr and Rb values within CHN, RUS and ESP, the plot reveals a clear separation between the Asian pine nuts (RUS and CHN) and the Mediterranean ones (ESP and TUR) based on [Sr] values. Additionally, although the CAT and TUR sample groups partially overlapped, a differentiation between them and CS samples was observed, with TUR and CAT pine nuts displaying higher median [Rb] values and less dispersion compared to CS. However, no clear distinction was observed between samples from CHN and RUS. As previously mentioned, these commercial samples may be sourced from multiple regions within each country. Hence, such dispersion is

expected given the heterogeneous bedrock in these areas. In light of the findings from this study, elemental analysis appears to be more effective than geoelement isotopic approaches for the geographical authentication of pine nuts. This aligns with some previous studies (Asfaha et al., 2011; D'Archivio et al., 2014; Epova et al., 2019; García-Ruiz et al., 2007; Geană et al., 2017; Horacek et al., 2023; Li et al., 2016) which demonstrates that mineral composition—specifically Rb and Sr—either alone or in combination with Sr isotopic ratio or with other elements such as Al, B, Ca, Cu, Ga, K or Na, provides better geographical discrimination for agrifood products. This approach consistently outperformed the use of the Sr isotope ratio to verify pine nut provenance.

Table 4. Sr and Rb concentrations of pine punts from different origins a) Countries of origin; b) Spanish regions of origin.

a)

	ESP (n= 60)		**CHN** (n= 15)		
	Median	Q_1 - Q_3	Median	Q_1 - Q_3	
[Sr] (µg/g)	1.41 [a]	0.98 – 2.05	0.06 [c]	0.05 – 0.12	
[Rb] (µg/g)	3.22 [c]	2.14 – 5.90	25.61 [a]	9.53 – 36.67	
	RUS (n=15)		**TUR** (n=6)		
	Median	Q_1 - Q_3	Median	Q_1 - Q_3	p-value
[Sr] (µg/g)	0.06 [c]	0.03 – 0.09	0.91 [b]	0.66 – 1.32	<0.001 *[1]
[Rb] (µg/g)	16.18 [a]	13.29 – 18.10	11.46 [b]	9.49 – 13.18	<0.001 *[1]

*p-value < 0.05 means are significantly different. [1] p-value obtained from independent samples median test. Differences between origins were noted as a > b > c. ESP: Spain; CHN: China; RUS: Russia; TUR: Turkey.

b)

	CAT (n= 30)		**CS** (n= 30)		
	Median	Q_1 - Q_3	Median	Q_1 - Q_3	p-value
[Sr] (µg/g)	1.34 [a]	0.88 – 1.67	1.64 [a]	1.03 – 4.70	0.439 [1]
[Rb] (µg/g)	5.63 [a]	3.06 – 8.60	2.60 [b]	1.69 – 3.34	<0.001 *[1]

*p-value < 0.05 means are significantly different. [1] p-value obtained from independent samples median test. Differences between regions were noted as a > b. CAT: Catalonia; CS: Central Spain.

4. Conclusions

Although significant differences were observed between the median Sr isotopic ratios of Spanish *P. pinea* and Asian pine nuts from Russia and China, these differences were not sufficient to clearly distinguish the origins, as the groups still overlapped. This outcome could be attributed to the heterogeneous bedrock geology of the studied areas. Therefore, these findings demonstrate that that relying solely on the analysis of Sr isotopic ratios may be insufficient to effectively authenticate the provenance of pine nuts.

On the other hand, the combination of Sr and Rb concentrations has shown promising results for the geographical authentication of pine nuts. Significant differences were detected in the median concentrations of Sr and Rb between the Mediterranean and Asian pine nuts, as well as among some individual countries. Within ESP, notable differences in Rb concentrations were observed between CAT and CS, suggesting that Rb could be useful for distinguishing between Spanish regions.

Additionally, the Sr vs. Rb concentration plot effectively differentiated between Asian (RUS and CHN) and Mediterranean (ESP and TUR) pine nuts, with partial distinction between CAT and CS regions.

This study establishes the bases for the development of efficient and robust tools based on Sr and Rb elemental composition for the geographical authentication of pine nuts. However, further evaluation with large-scale datasets that encompass natural variability, diverse production regions, different harvest years, and various species is needed. Additionally, exploring the use of these markers in conjunction with other mineral elements or light bioelements isotope ratios could further improve the accuracy and efficiency of geographical discrimination.

References

Ariyama, K., Nishida, T., Noda, T., Kadokura, M., Yasui, A. (2006). Effects of fertilization, crop year, variety, and provenance factors on mineral concentrations in onions. *Journal of Agricultural and Food Chemistry*, *54*(9), 3341-3350. https://doi.org/10.1021/jf0525481

Asfaha, D. G., Quétel, C. R., Thomas, F., Horacek, M., Wimmer, B., Heiss, G., Dekant, C., Deters-Itzelsberger, P., Hoelzl, S. Rummel, S., Brach-Papa, C., van Bocxstaele, M., Jamin, E., Baxter, M. Heinrich, K., Kelly, S., Bertoldi, D., Bontempo, L., Camin, F.,Larcher, R., Perini, M., Rossmann, A., Schellenberg, A., Schlichti, C., Froeschl, H., Hoogewerff, J., Ueckermann, H. (2011). Combining isotopic signatures of n (^{87}Sr)/n (^{86}Sr) and light stable elements (C, N, O, S) with multi-elemental profiling for the authentication of provenance of European cereal samples. *Journal of Cereal Science*, *53*(2), 170-177. https://doi.org/10.1016/j.jcs.2010.11.004

Awan, H. U. M., Pettenella, D. (2017). Pine nuts: a review of recent sanitary conditions and market development. *Forests*, *8*, Article 367. https://doi.org/10.3390/f8100367

Azmy, K., Veizer, J., Wenzel, B., Bassett, M. G., Copper, P. (1999). Silurian strontium isotope stratigraphy. *Geological Society of America Bulletin*, *111*, 475–483. https://doi.org/10.1130/0016-7606(1999)111%3C0475:SSIS%3E2.3.CO;2

Baffi, C., Trincherini, P. R. (2016). Food traceability using the ^{87}Sr/^{86}Sr isotopic ratio mass spectrometry. *European Food Research and Technology*, *242*, 1411-1439. https://doi.org/10.1007/s00217-016-2712-2

Bonari, G., Chytrý, K., Çoban, S., Chytrý, M. (2020). Natural forests of *Pinus pinea* in western Turkey: a priority for conservation. *Biodiversity and Conservation*, *29*, 3877–3898. https://doi.org /10.1007/s10531-020-02052-z

Bong, Y. S., Shin, W. J., Gautam, M. K., Jeong, Y. J., Lee, A. R., Jang, C. S., Lim, Y. P., Chung, G. S., Lee, K. S. (2012). Determining the geographical origin of Chinese cabbages using multielement composition and strontium isotope ratio analyses. *Food Chemistry*, *135*(4), 2666-2674. https://doi.org/10.1016/j.foodchem.2012.07.045

Bontempo, L., Paolini, M., Franceschi, P., Ziller, L., García-González, D. L., Camin, F. (2019). Characterisation and attempted differentiation of european and extraeuropean olive oils using stable isotope ratio analysis. *Food Chemistry*, *276*, 782–789. https://doi.org/10.1016/j.foodchem.2018.10.077

Bora, F. D., Donici, A., Rusu, T., Bunea, A., Popescu, D., Bunea, C. I. (2018). Elemental Profile and 207Pb/206Pb, 208Pb/206Pb, 204Pb/206Pb, $^{87}Sr/^{86}Sr$ Isotope Ratio as Fingerprints for Geographical Traceability of Romanian Wines. *Notulae Botanicae Horti Agrobotanici Cluj-Napoca*, *46*(1), 223–239. https://doi.org/10.15835/nbha46110853

Camin, F., Boner, M., Bontempo, L., Fauhl-Hassek, C., Kelly, S. D., Riedl, J., Rossmann, A. (2017). Stable isotope techniques for verifying the declared geographical origin of food in legal cases. *Trends in Food Science & Technology*, *61*, 176–187. https://doi.org/10.1016/j.tifs.2016.12.007

Danezis, G. P., Georgiou, C. A. (2022). Elemental metabolomics: Food elemental assessment could reveal geographical origin. *Current Opinion in Food Science*, *44*, 100812. https://doi.org/10.1016/j.cofs.2022.100812

D' Archivio, A. A., Giannitto, A., Incani, A., Nisi, S. (2014). Analysis of the mineral composition of Italian saffron by ICP-MS and classification of geographical origin. *Food Chemistry*, *157*, 485-489. https://doi.org/10.1016/j.foodchem.2014.02.068

Di Paola-Naranjo, R. D., Baroni, M. V., Podio, N. S., Rubinstein, H. R., Fabani, M. P., Badini, R. G., Inga, M., Ostera, H. A., Cagnoni, M., Gallegos, E., Gautier, E., Peral-García, P., Hoogewerff, J., Wunderlin, D. A. (2011). Fingerprints for main varieties of Argentinean wines: terroir differentiation by inorganic, organic, and stable isotopic analyses coupled to chemometrics. *Journal of Agricultural and Food Chemistry*, *59*(14), 7854-7865. https://doi.org/10.1021/jf2007419

Drivelos, S. A., Georgiou, C. A. (2012). Multi-element and multi-isotope-ratio analysis to determine the geographical origin of foods in the European Union. *Trends in Analytical Chemistry*, *40*, 38–51. https://doi.org/10.1016/j.trac.2012.08.003

Durante, C., Baschieri, C., Bertacchini, L., Bertelli, D., Cocchi, M., Marchetti, A., Manzini, D., Papotti, G., Sighinolfi, S. (2015). An analytical approach to Sr isotope ratio determination in Lambrusco wines for geographical traceability purposes. *Food Chemistry*, *173*, 557-563. https://doi.org/10.1016/j.foodchem.2014.10.086

Epova, E. N., Bérail, S., Séby, F., Vacchina, V., Bareille, G., Médina, B., Sathou, L., Donard, O. F. (2019). Strontium elemental and isotopic signatures of Bordeaux wines for authenticity and geographical origin assessment. *Food Chemistry*, *294*, 35-45. https://doi.org/10.1016/j.foodchem.2019.04.068

Evaristo I., Batista D., Correia I., Correia P., Costa R. (2013). Chemical profiling of Portuguese *Pinus pinea* L. nuts and comparative analysis with Pinus koraiensis Sieb. & Zucc. commercial kernels. *Options Méditerranéennes. Séries A: Mediterranean Seminars*, *105*, 99–104. http://om.ciheam.org/om/pdf/a105/00006787.pdf

FAO/UNESCO. (1971-1981). Soil Map of the World (1:5,000,000). Food and Agriculture Organization of the United Nations. https://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/en/ (accessed Aug 20, 2024)

Fortunato, G., Mumic, K., Wunderli, S., Pillonel, L., Bosset, J. O., Gremaud, G. (2004). Application of strontium isotope abundance ratios measured by MC-ICP-MS for food authentication. *Journal of Analytical Atomic Spectrometry*, *19*(2), 227-234. https://doi.org/10.1039/B307068A

García-Ruiz, S., Moldovan, M., Fortunato, G., Wunderli, S., Alonso, J. I. G. (2007). Evaluation of strontium isotope abundance ratios in combination with multi-elemental analysis as a possible tool to study the geographical origin of ciders. *Analytica Chimica Acta*, *590*(1), 55-66. https://doi.org/10.1016/j.aca.2007.03.016.

Geană, E. I., Sandru, C., Stanciu, V., Ionete, R. E. (2017). Elemental profile and $^{87}Sr/^{86}Sr$ isotope ratio as fingerprints for geographical traceability of wines: an approach on Romanian Wines. *Food Analytical Methods*, *10*, 63-73. https://doi.org/10.1007/s12161-016-0550-2

Gonçalves, B., Pinto, T., Aires, A., Morais, M. C., Bacelar, E., Anjos, R., Ferreira-Cardoso, J., Oliveira, I., Vilela, A., Cosme, F. (2023). Composition of nuts and their potential health benefits—An overview. *Foods*, *12*(5), 942. https://doi.org/10.3390/foods12050942

Griboff, J., Baroni, M. V., Horacek, M., Wunderlin, D. A., Monferran, M. V. (2019). Multielemental+ isotopic fingerprint enables linking soil, water, forage and milk composition, assessing the geographical origin of Argentinean milk. *Food Chemistry*, *283*, 549-558. https://doi.org/10.1016/j.foodchem.2019.01.067

Gutiérrez, F., Harvey, A., García-Ruiz, J. M., Silva, P., Cendrero, A. (2013). Geomorphological research in Spain. *Geomorphology*, *196*, 1-12. https://doi-org.sire.ub.edu/10.1016/j.geomorph.2012.12.014

Horacek, M. (2022a). The Need to Consider Geochemistry When Interpreting Sr-Isotopes. Comment on Gregorčič et al. The Provenance of Slovenian Milk Using $^{87}Sr/^{86}Sr$ Isotope Ratios. *Foods*, *11*(4), 564. https://doi.org/10.3390/foods11040564

Horacek, M., Klcova, L., Hudcovicova, M., Ondreickova, K., Gubis, J., Hölzl, S. (2022b). Differentiation of Apricots of Different Geographic Origin in Central and Southern Europe by Applying $^{87}Sr/^{86}Sr$ Analysis: Potential and Limitations. *Foods*, *11*(15), 2239. https://doi.org/10.3390/foods11152239

Horacek, M., Lage, M., Vakhlu, J. (2023). Exploring the potential of $^{87}Sr/^{86}Sr$ isotope ratio with strontium and rubidium levels to assess the geographic origin of saffron. *Foods*, *12*(15), 2830. https://doi.org/10.3390/foods12152830

Kelly, S., Heaton, K., Hoogewerff, J. (2005). Tracing the geographical origin of food: The application of multi-element and multi-isotope analysis. *Trends in Food Science & Technology*, *16*, 555–567. https://doi.org/10.1016/j.tifs.2005.08.008

Kim, K. W., Thornton, I. (1993). Influence of Ordovician uraniferous black shales on the trace element composition of soils and food crops, Korea. *Applied Geochemistry*, *8*, 249-255. https://doi.org/10.1016/S0883-2927(09)80045-0

Laursen, K. H., Bontempo, L., Camin, F., Roßmann, A. (2016). Advances in isotopic analysis for food authenticity testing. In G. In, & D. G. Woodhead (Eds.), *Publishing Series in Food Science, Technology and Nutrition, Advances in Food Authenticity Testing* (pp. 227–252). Woodhead Publishing. https://doi.org/10.1016/B978-0-08-100220- 9.00009-6

Li, C., Dong, H., Luo, D., Xian, Y., Fu, X. (2016). Recent developments in application of stable isotope and multi-element analysis on geographical origin traceability of cereal grains. *Food Analytical Methods*, *9*, 1512-1519. https://doi.org/10.1007/s12161-015-0328-y

Moscetti, R., Berhe, D. H., Agrimi, M., Haff, R. P., Liang, P., Ferri, S., Monarca, D., Massantini, R. (2021). Pine nut species recognition using NIR spectroscopy and image analysis. *Journal of Food Engineering*, *292*, Article 110357. https://doi.org/10.1016/j.jfoodeng.2020.110357

Mutke, S. (2022). Producción y comercio de piñones comestibles en el mundo. *Montes*, *147*, 40–43. https://www.revistamontes.net/Buscador.aspx?id=15180 (accessed Aug 20, 2024)

Nier, A. O. (1938). The isotopic constitution of strontium, Barium, bismuth, thallium and mercury. *Physical Review Journals*, *54*, 275–278. https://doi.org/10.1103/ PhysRev.54.275

Pustjens, A. M., Muilwijk, M., Weesepoel, Y., van Ruth, S. M. (2016). Advances in authenticity testing of geographical origin of food products. In G. In, & D. G. Woodhead (Eds.), *Publishing Series in Food Science, Technology and Nutrition, Advances in Food Authenticity Testing* (pp. 339-367). Woodhead Publishing. https://doi.org/10.1016/B978-0-08-100220-9.00013-8

Quintanilla-Casas, B., Torres-Cobos, B., Guardiola, F., Servili, M., Alonso-Salces, R. M., Valli, E., Bendini, A., Toschi, T. G., Vichi, S., Tres, A. (2022). Geographical authentication of virgin olive oil by GCMS sesquiterpene hydrocarbon fingerprint: Verifying EU and single country label-declaration. *Food Chemistry*, *378*, 132104. https://doi.org/10.1016/j.foodchem.2022.132104

Regulation (EU) No 2023/2429 of the European Parliament and of the Council of 17 August 2023 as regards marketing standards for the fruit and vegetables sector, certain processed fruit and vegetable products and the bananas sector, pp. 1–77. http://data.europa.eu/eli/reg_del/2023/2429/oj

Ríos-Reina, R., Callejón, R. M., Amigo, J. M. (2021). Feasibility of a rapid and non-destructive methodology for the study and discrimination of pine nuts using near-infrared hyperspectral analysis and chemometrics, *Food Control*, *130*, 108365. https://doi.org/10.1016/j.foodcont.2021.108365

Sen, I., Tokatli, F. (2014). Characterization and classification of Turkish wines based on elemental composition. *American Journal of Enology and Viticulture*, *65*(1), 134-142. https://doi.org/ 10.5344/ajev.2013.13081

Takala, R., Ramji, D. P., Choy, E. (2023). The beneficial effects of pine nuts and its major fatty acid, pinolenic acid, on inflammation and metabolic perturbations in inflammatory disorders. *International Journal of Molecular Sciences*, *24*(2), 1171. https://doi.org/10.3390/ijms24021171

Techer, I., Medini, S., Janin, M., Arregui, M. (2017). Impact of agricultural practice on the sr isotopic composition of food products: Application to discriminate the geographic origin of olives and olive oil. *Applied Geochemistry*, *82*, 1–14. https://doi. org/10.1016/j.apgeochem.2017.05.010

Vitòria, L., Otero, N., Soler, A., Canals, A. (2004). Fertilizer Characterization: Isotopic data (N, S, O, C, and Sr). *Environmental Science & Technology*, *38*, 3254–3262. https:// doi.org/10.1021/es0348187

6.2 Enhancement and validation of metabolic methods for pine nut geographical and botanical authentication

6.2.1 Sesquiterpene fingerprinting

To assess the suitability of mono- and sesquiterpene compounds as tools for **verifying pine nut species and provenance**, the volatile and semi-volatile terpene fractions of 253 pine nut samples was analysed using HS-SPME-GC–MS (**Figure 6**). These samples were collected between 2020 and 2023. Among them:

- 170 were traceable *P. pinea* samples from two regions in Spain (ESP): Catalonia (CAT, n = 74) and Central Spain, Madrid and Castile and Leon (CS, n = 96).
- The remaining 83 samples were commercial products from China (CHN, n = 53), Russia (RUS, n = 22), and Turkey (TUR, n = 8). Based on the natural distribution of pine nut species, it was assumed that the CHN and RUS samples primarily belonged to *P. koraiensis* and *P. sibirica*, respectively. In contrast, the commercial TUR samples could include both *P. pinea* and other local species.

Two PLS-DA classification models were developed and externally validated (**Figure 16**):

1) A multispecies-geographical origin model (binary PLS-DA) to differentiate between Spanish *P. pinea* pine nuts vs. Asian pine nuts of other species.

2) A binary *P. pinea*-geographical origin model to distinguish pine nuts of the same species (*P. pinea*) from two distinct Spanish regions (CAT and CS).

The PLS-DA regression coefficients of both models were examined to tentatively identify the variables that significantly contributed to the discrimination between Asian and Spanish kernels, as well as between the two Spanish regions (CAT vs. CS) (Publication 11).

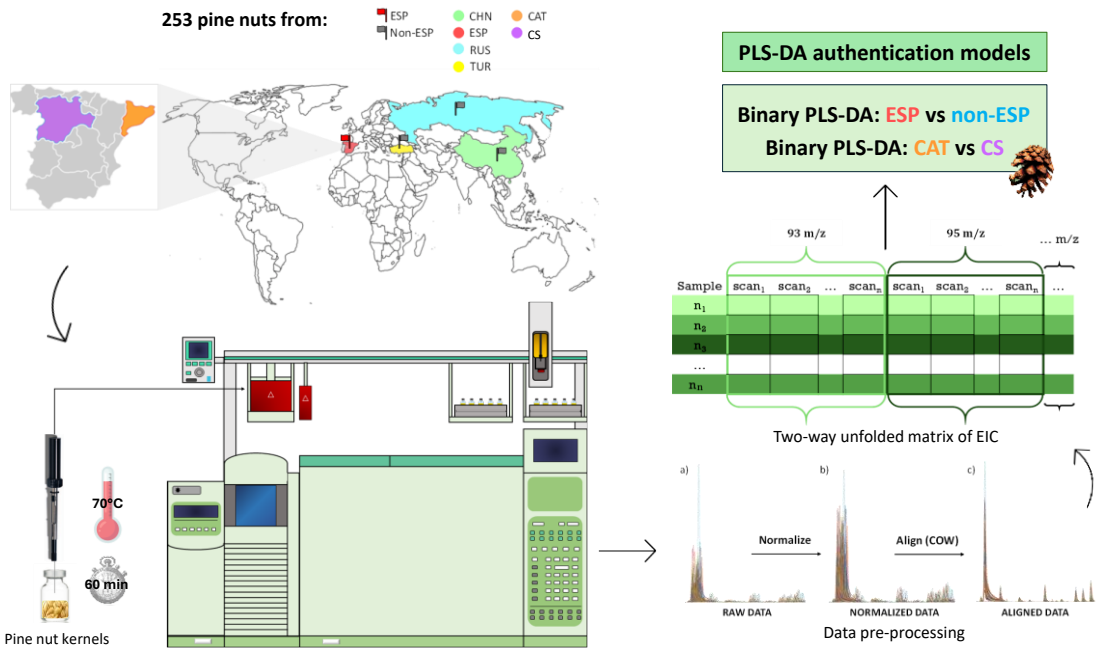


Figure 16. Graphical abstract of Publication 11. ESP: Spain, non-ESP: the other countries (CHN: China, RUS: Russia, TUR: Turkey), CAT: Catalonia, CS: Central Spain, EIC: extracted ion chromatogram, PLS-DA: partial least square-discriminant analysis.

6.2.2 Publication 11

Mono- and sesquiterpenoid fingerprinting: a powerful and streamlined solution for pine nut authentication

Berta Torres-Cobos, Soriana Beatrice Nicotra, Celia Asensio-Manzano, Neus Aletà, Anna Teixidó, Mercè Rovira, Agustí Romero, Francesc Guardiola, Stefania Vichi, Alba Tres

Under review in Food Chemistry 2024

Supplementary material available in **Annex 9**

Mono- and sesquiterpenoid fingerprinting: a powerful and streamlined solution for pine nut authentication

B. Torres-Cobos[1,2], S. Nicotra[1,2], C. Asensio-Manzano[1], N. Aletà[3,4], A. Teixidó[4], M. Rovira[5], A. Romero[5], F. Guardiola[1,2], S. Vichi[1,2*], A. Tres[1,2]

[1] University of Barcelona, Department of Nutrition, Food Sciences and Gastronomy, Prat de la Riba 171, Santa Coloma de Gramenet, 08921, Spain.

[2] University of Barcelona, Institute of Research on Food Nutrition and Safety (INSA-UB), Prat de la Riba 171, Santa Coloma de Gramenet, 08921, Spain.

[3] Institute of Agrifood Research and Technology (IRTA). Fruit Tree Program. Torre Marimon. 08140 Caldes de Montbui, Spain

[4] Forest Science and Technology Centre (CTFC). Multifunctional Forest Management Program. 25280 Solsona, Lleida, Spain.

[5] Institute of Agrifood Research and Technology (IRTA), Ctra. de Reus – El Morell Km 3.8, Constantí, 43120, Spain.

*Corresponding author: S. Vichi, stefaniavichi@ub.edu. Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Campus de l'Alimentació Torribera, Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona. Av Prat de la Riba, 171. Edifici Gaudí 08921 Santa Coloma de Gramenet, Spain.

ABSTRACT

Pine nuts are highly valued worldwide for their quality and commercial importance, with China and Russia as major producers of Pinus koraiensis and Pinus sibirica, and the Iberian Peninsula and Turkey supplying Pinus pinea. Their species and origin influence their sensory and nutritional characteristics and affect their price, making them vulnerable to counterfeiting. This study proposes a novel authentication method using mono- and sesquiterpene fingerprints (extracted ion chromatograms from specific ions) analysed via solid-phase microextraction combined with gas chromatography-mass spectrometry, combined with chemometrics (partial least squares – discriminant analysis). Tested on 253 samples from China, Russia, Spain, and Turkey across different harvest years, it achieved 100% accuracy in external validation when distinguishing Spanish from non-Spanish pine nuts, and 99% accuracy in differentiating Pinus pinea samples from two distinct Spanish regions. This simple, affordable, and automatable approach proves effective as a screening tool that could be applied to support official controls.

Keywords: Pine nuts; Fraud; Terpenoids; Fingerprint; Authenticity; Geographical origin; PLS-DA.

1. Introduction

Pine nuts, popularly known as the “white gold”, are the most expensive nuts on the market. They account for only 1% of global tree nut production but have a supply value of more than 1.3 billion USD (INC, 2023). Asia stands as the primary global producer of pine nuts, with China, Russia, and North Korea leading the output, followed by the Mediterranean basin, with Turkey and the Iberian Peninsula as the main producers (INC, 2023).

While many agri-food products have cultivars farmed worldwide, the species of pine nuts are strictly tied to their geographical origins. The most common species of pine nuts among the 20 commercially available are *Pinus pinea* L. (*P. pinea*), predominantly growing in the Mediterranean region, and *Pinus koraiensis* Siebold & Zucc. (*P. koraiensis*) and *Pinus sibirica* (*P. sibirica*), primarily sourced from China and Russia, respectively (Awan and Pettenella, 2017; Moscetti et al., 2021). The sensory attributes, nutritional values and market price of pine nuts are highly dependent on the specie and region of origin (Awan et al., 2017; Evaristo et al., 2013; Mutke, 2022). Mediterranean pine nuts are highly valued and appreciated by consumers, reaching prices higher than 100 EUR/kg (Mutke, 2022); however, they account for only 5% of the world average pine nut production (INC, 2023). In contrast, Chinese or Russian pine nuts are usually sold at much lower prices, often less than a third of the value of Mediterranean ones (Evaristo et al., 2013; Moscetti et al., 2021; Mutke, 2022).

Despite the differences among pine nuts from various origins and species, non-expert consumers often find them difficult to distinguish. Consequently, EU regulations mandate the declaration of the country of origin on pine nut packaging (Regulation (EU) No 2023/2429). Additionally, international commercial labelling standards recommend including the botanical species (UNECE, 2013). These label claims need to be verified by regulatory bodies to prevent fraud and protect consumers. Indeed, due to the significant price difference between pine nuts of different origins, they are highly vulnerable to economically motivated fraudulent practices such as counterfeiting or adulteration. These practices can have serious consequences not only for the economy, impacting both the market and producers, but also for consumers' health, as they compromise the traceability chain of food products (Moscetti et al., 2021). In the

particular case of pine nuts, misrepresenting their origin carries an added risk because Chinese *P. koraiensis* is sometimes marketed mixed with other pine seed species like *Pinus armandii* Franch., which has been linked to the dysgeusia called 'Pine Mouth Syndrome' (Mutke et al., 2013; Destillats et al., 2011).

For all these reasons, disposing of reliable methods for pine nut authentication is crucial to safeguard the interests of both producers and consumers. Traditional methods to authenticate pine nuts have been based on phenotypic observations of physical traits such as pine nut kernel morphology (Fardin-Kia et al., 2012; Loewe-Muñoz et al., 2018; Mikkelsen et al., 2014) but their susceptibility to external agents, and the fact that the evaluation is limited to whole kernels, hinder their effectiveness. Consequently, some studies have focused on genetic analysis to distinguish pine nuts species (Handy et al., 2011). Despite their reliability and accuracy, these methods are laborious, complex, destructive, and expensive (Fardin-Kia et al., 2012; Ríos-Reina et al., 2021), and thus, hardly applicable for routine analysis of large sample sets.

Alternatively, pine nut composition has been investigated using different analytical approaches, including targeted fatty acids analysis (Destaillats et al., 2010, 2011; Evaristo et al., 2013; Fardin-Kia et al., 2012) and comprehensive spectroscopic techniques like near infrared spectroscopy (Loewe et al., 2017; Moscetti et al., 2021). Image analysis was also proposed for pine nut authentication (Ríos-Reina et al., 2021), although its application is restricted to entire kernels. While these methods showed promising results, there is still a need to fully evaluate their efficiency on sample sets that sufficiently represent the natural diversity of pine nut production, covering a wider range of origins, producers, harvest years, and species. Therefore, it is essential to develop a fit-for-purpose analytical method to verify pine nut authenticity.

In this context, previous research has demonstrated that mono- and sesquiterpenes could be reliable markers of varietal and geographical origin of different plant species and vegetable-derived products such as spices, alcoholic beverages and oils (Avula et al., 2015; Matsushita et al., 2018; Vichi et al., 2005; Marti et al., 2014; Quintanilla-Casas et al., 2022a; Torres-Cobos et al., 2021). These terpenes are secondary plant metabolites whose presence and composition are highly dependent on the plant botanical and geographical origin. In fact, they are shaped by environmental and genetic factors, with minimal impact from other factors such as processing or storage

conditions (Quintanilla-Casas et al., 2020; Vichi et al., 2010, 2018). When applied to virgin olive oils, sesquiterpene chromatographic fingerprint coupled with pattern recognition techniques, such as Partial Least Squares-Discriminant Analysis (PLS-DA) has been shown to be fast, robust and efficient for varietal and geographical authentication (Quintanilla-Casas et al., 2022a; Torres-Cobos et al., 2021). While other nut species typically lack appreciable amounts of terpenoids in their kernels, conifers produce an abundant amount of volatile and semi-volatile terpene metabolites and some of them have also been identified in pine nut kernels (Adelina et al., 2021; Rogachev et al., 2015). Recent reports documenting variations in the mono- and sesquiterpene composition among different pine species and origins (Arrabal et al., 2012; Faria et al., 2021; Kim et al., 2024) position the volatile terpene fingerprint as a promising marker for pine nut authentication. Moreover, monoterpenoids were reported as the main compounds in the essential oils of the pine bark, wood, needles, and cones, as well as in the volatile fraction of raw pine nut kernels (Adelina et al., 2021; Nikolic et al., 2022; Rogachev et al., 2015). The hypothesis of our study is that the mono- and sesquiterpene fingerprint of pine nuts can serve as a reliable and efficient marker to discriminate among pine nut kernel species and provenances.

The objective of the present research is to develop a fast, efficient, and reliable method to enhance the discrimination of pine nuts based on the volatile and semi-volatile terpene fingerprint obtained by Headspace Solid Phase Microextraction-Gas Chromatography-Mass Spectrometry (HS-SPME-GC–MS) from a wide sample set reflecting their natural variability. This involved the development and external validation of two PLS-DA classification models: (i) a multispecies geographical model to distinguish between Spanish *P. pinea* kernels vs. Asian kernels of other species, and (ii) a *P. pinea* geographical model to differentiate pine nuts of the same species from two distinct Spanish regions.

2. Material and methods

2.1 Sampling

A set of 253 pine nut samples from different geographical regions and species was obtained from 2020 to 2023 in the frame of the TRACENUTS project (PID2020-

117701RB-I00) (**Table S1** of **Supplementary Information**). Among these samples, 83 were commercial samples from: China (CHN, n = 53), Russia (RUS, n = 22) and Turkey (TUR, n = 8). According to the natural distribution of pine nut species, it was assumed that CHN and RUS samples did not originate from *P. pinea* but primarily belonged to *P. koraiensis* and *P. sibirica*, the most commercially significant species from these countries (Awan and Pettenella, 2017). Commercial TUR samples may have belonged to both *P. pinea* and other local species (Bonari et al., 2020). The remaining 170 samples were *P. pinea* kernels from Spain (ESP), sourced from two regions: Central Spain, Madrid and Castile and Leon (CS, n= 96) and Catalonia (CAT, n= 74). All the Spanish samples were traceable and were supplied by the Institute of Forest Science (ICIFOR-INIA, Madrid, Spain), the "*Centro de Servicios y Promocion Forestal y de su Industria de Castilla y León*" (CESEFOR foundation, Soria, Spain), and the Institute of Agrifood Research and Technology (IRTA-Torre Marimon, Caldes de Montbui, Spain). Samples were directly harvested from forests and were stored at 4 °C until analysis.

2.2 Headspace-solid phase microextraction (HS-SPME)

Pine nut samples were analysed under conditions adapted from Vichi et al. (2006) using a Combi-pal autosampler (CTC Analytics, Zwingen, Switzerland). An aliquot of approximately 1 g of whole pine nuts was weighed into a 10 mL vial fitted with a PTFE/silicone septum. The sample was conditioned at 70 °C for 10 min, followed by exposing a divinylbenzene/carboxen/polydimethylsiloxane (DVB/CAR/PDMS) fiber (2 cm length, 50/30 µm film thickness) from Supelco (Bellefonte, Pennsylvania, USA) to the sample headspace for 60 min, at the same temperature. Then, the fiber was desorbed at 260 ˚C for 10 min in the gas chromatograph injection port, the injector was maintained in split-less mode for the first 5 min. To monitor carryover, blank samples were alternated between injections.

2.3 Gas Chromatography-Mass Spectrometry (GC-MS)

The mono- and sesquiterpene fingerprint was acquired by an Agilent 6890N Network GC system coupled to a quadrupolar mass selective analyser Agilent 5975C Inert MSD (Agilent Technologies, Santa Clara, California, USA). Helium was the carrier gas, at a flow of 1.5 mL/min. Analytes were separated on a Supelcowax-10 capillary column (60

m × 0.25 mm i.d., 0.25 µm film thickness) (Supelco, Bellefonte, Pennsylvania, USA). Column temperature was held at 40 °C for 3 min, increased to 100 °C at 4 °C/min, then, to 200 °C at 5 °C/min and to 260 °C at 15 °C/min, holding the last temperature for 5 min. The temperatures of the ion source and the transfer line were 230 and 280°C, respectively. Mass spectra were recorded at 2.338 scan/s and the electron energy was 70 eV. Data acquisition was performed in the selected ion monitoring (SIM) mode between 0 and 47.2 min, by registering the Extracted Ion Chromatogram (EIC) of 7 ions which have been reported to be characteristic of the mono and sesquiterpene compounds and their oxygenated derivates: *m/z* 93, 95, 119, 159, 161, 189 and 204 (Maleknia et al., 2007; Reed, 1963; Tani et al., 2003; Torres-Cobos et al., 2021; Vichi et al., 2010). Therefore, for each ion, the intensities of a total of 6621 scans were acquired and used as fingerprinting data (section 2.3) to build the authentication models (section 2.4).

After fingerprinting models were developed, we tentatively identified the compounds corresponding to the scans leading to the most relevant regression coefficients (section 2.4.3). To do so, acquisition was carried out in the full scan mode in the range m/z 35-350 and the MS spectra at the retention times corresponding to the relevant scans were obtained. This tentative molecular structure identification was a level 3 identification (tentative candidate, evidence exists for possible structure, but insufficient information for one exact structure only) according to Schymanski et al. (2014).

2.4 Fingerprinting approach

A fingerprinting approach was followed using the EICs of the 7 selected ions. Scan intensities were considered from 0 to 47.2 min for each ion (6621 scans × 7 ions = 46347 variables per sample). The acquisition interval has been extended from previous studies (Torres-Cobos et al., 2021) to include the monoterpenoids that appear at the initial times of the chromatogram (from 0 to 30 minutes) due to their relevance and abundance in pine nuts. A data matrix was built for each ion, with the scan intensities of each EIC (columns) for all the samples (rows) (7 matrices of 6621 columns x 253 rows). Differences between injections were corrected by normalizing each EIC, which consisted in dividing each scan intensity by the total sum of intensities (Nam et al.,

2020). Then, the EICs in each matrix were aligned by Correlation Optimized Warping (COW) algorithm in Matlab® (Nielsen et al., 1998). Finally, the 7 aligned EIC matrices were concatenated conforming a two-way unfolded matrix (253 samples × 46347 variables).

2.5 Chemometrics

2.5.1. Data exploration and preliminary multi-class geographical model

The data treatment and model building were performed with SIMCA software v13.0© (Sartorius, Göttingen, Germany). A Principal Component Analysis (PCA) was performed for the exploration of data and to identify potential outliers, according to Hotelling's T^2 range and Q-residuals model parameters. The exploratory analysis of the dataset showed no outliers.

A preliminary multi-class PLS-DA classification model was built to discriminate among the four countries of origin (CHN, TUR, RUS, ESP), to assess the potential of the terpene fingerprinting to distinguish pine nuts from different origins. Multi-class PLS-DA models operate as multiple binary models; each class being compared to the rest of the samples. A dummy Y matrix is used, containing as many classification vectors as classes. Each vector has values of 1 for a specific class and 0 for all other classes (i.e., 1 for CHN, TUR, RUS, and ESP classes, and 0 for non-CHN, non-TUR classes, non-RUS, and non-ESP). The multi-class PLS-DA model was internally validated through leave 10%-out cross validation. The model's optimal preprocessing and number of latent variables (LV) were selected according to the lowest Root Mean Squared Error of Cross Validation (RMSEcv) criteria. The pre-processing was mean centring and scaling to the unit of variance. To evaluate model overfitting, permutation test (n = 20 permutations) and ANOVA on the cross-validated predictive residuals (p-value) were carried out (Eriksson et al., 2008; Quintanilla-Casas et al., 2020). The suitability PLS-DA model was evaluated by the Q^2 values and the percentage of correct classification of each class, and the resulting score plot was examined to identify any clustering among samples.

2.5.2. Partial Least Squares Discriminant Analysis (PLS-DA) binary classification models

After the initial data exploration and the exclusion of origins represented by fewer than 20 samples (TUR), two PLS-DA binary classification models were built: (i) a multi-species geographical model to discriminate between pine nuts from ESP (*P. pinea*) and

non-ESP (other species), and (ii) a *P. pinea* geographical model to classify the ESP *P. pinea* samples by their region of production: CAT and CS.

For each authentication model, the sample set was split following a stratified random sampling strategy into training (80 % of the samples of each class: ESP/non-ESP model, n = 196; CAT/CS model, n = 136) and validation set (20 % of the samples of each class: ESP/non-ESP model, n = 49; CAT/CS model, n = 34). This splitting was run seven times (7 iterations) to evaluate the effect of the sample set composition and to increase the robustness of the external validation. The sample set splitting information, including validation and training sets, is summarized in **Table S1** of **Supplementary Information**.

In each iteration, a PLS-DA binary model (training model) was fitted and internally validated through leave 10%-out cross validation. The model's optimal pre-processing and LV were selected according to the RMSEcv criteria. For all training models, the optimal pre-processing was mean centring and scaling to the unit of variance. To evaluate model overfitting, permutation test (n = 20 permutations) and ANOVA on the cross-validated predictive residuals (p-value) were carried out (Eriksson et al., 2008; Quintanilla-Casas et al., 2020). None of the training models was overfitted according to the permutation test and ANOVA p-value results. Subsequently, each training model was externally validated by predicting the class of the samples in the corresponding validation set, which had not been used to build the model. Therefore, for each type of model, seven training PLS-DA models and the corresponding seven external validations were obtained, to verify that results were not driven by specific influential samples and thus, to increase the robustness of the external validation.

In PLS-DA binary models, classes are expressed as PLS dummy variables (here, 1 for non-ESP, and CS classes, and 0 for ESP and CAT classes). The PLS predicted value (PV) of each sample was used for its classification into one class or the other according to a classification threshold (here, PV = 0.5).

The performance of each PLS-DA model was evaluated by the Q^2 values and efficiency, which was expressed as the percentage of correct classification of each class, the sensitivity (the number of true positive results/ [the number of true positive results + the number of false negative results]) and specificity (the number of true negative results/ [the number of true negative results + the number of false positive results]) values. Wilson score intervals were calculated to establish confidence intervals for

models' sensitivity and specificity (Wilson, 1927). Non-ESP and CS samples were arbitrarily defined as the positive samples for the corresponding models.

2.5.3. Evaluation of PLS-DA regression coefficients

The regression coefficients of the PLS-DA models were studied to tentatively identify the key variables that mainly contributed to the discrimination between classes. The jack-knife standard error of cross-validation (SEcv) was used to evaluate the significance of the regression coefficients, considering as significant those with values higher than their corresponding SEcv (Torres-Cobos et al., 2024). Among the significant variables, only the ones with the highest absolute values (the 3 % higher regression coefficients) were considered, and the corresponding compounds were tentatively identified based on their mass spectra and elution order from full scan injections as explained in section 2.4.

3. Results and discussion

3.1 Exploratory analysis and preliminary multi-class geographical model

The preliminary multi-class PLS-DA model to classify the samples according to their country of origin showed promising results. The inspection of the score plot (**Figure 1**) evidenced a clear clustering of samples by country of origin. LV1 was useful in discriminating TUR and ESP samples from CHN and RUS ones, whereas LV3 distinguished TUR from ESP and RUS from CHN pine nuts, achieving four groups quite differentiated from one another.

The leave 10%-out cross validation (**Table 1**) yielded a 100% correct classification for the pine nuts from ESP and TUR, and high correct classification rates for the CHN (92%) and RUS (95%) classes, with only 2 CHN samples misclassified as RUS, and one RUS misclassified as CHN. The misclassification may be attributed to the greater similarity between these two classes, as both CHN and RUS samples belong to species other than P. pinea and originate from regions geographically distant from the Mediterranean. Additionally, this similarity is evident in the scores plot (Figure 1), where the clusters for RUS and CHN samples show significant dispersion and partial overlap. This overlap suggests that the chemical fingerprinting approach may have difficulty distinguishing between these two classes due to their more closely related terpene composition.

Permutation test and ANOVA p-value allowed excluding model overfitting (**Table S2** of **Supplementary Information**).

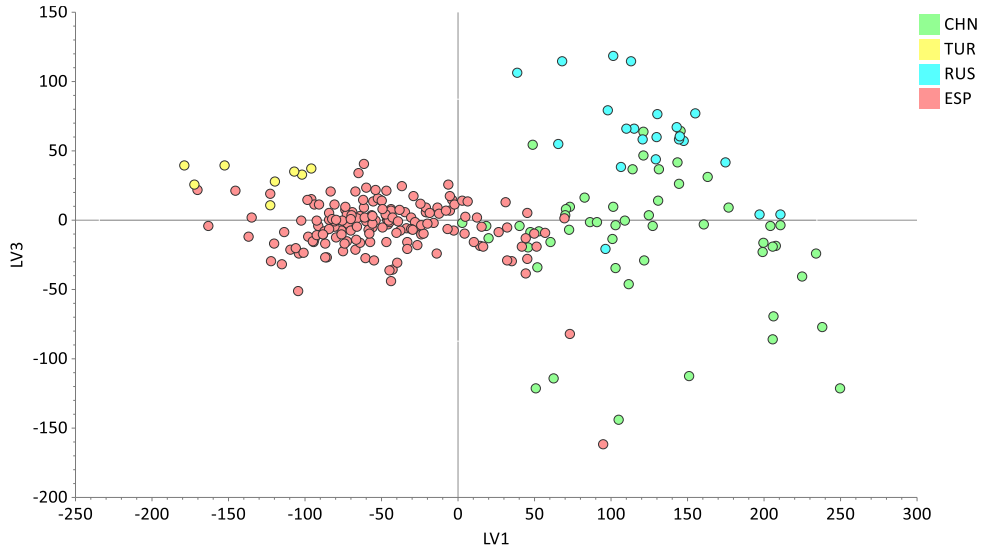


Figure 1. Score scatter plot (LV1 vs LV3) of the PLS-DA classification model developed by country of origin (n = 253, 9 LVs, Q2 = 0.659, RMSEcv = 0.216, ANOVA p-value < 0.05), based on pine nut sesquiterpene fingerprinting data. CHN: China, TUR: Turkey, RUS: Russia, ESP: Spain.

Table 1. Leave 10%-out cross validation of the four-class PLS-DA model developed by country of origin, based on pine nut mono- and sesquiterpene fingerprinting data.

Multi species geographical model: CHN/TUR/RUS/ESP							
	n	CHN (n)	TUR (n)	RUS (n)	ESP (n)	Not assigned (n)	Correct classification (%)
CHN	53	49	0	2	0	2	92.5
TUR	8	0	8	0	0	0	100.0
RUS	22	1	0	21	0	0	95.5
ESP	170	0	0	0	170	0	100.0
Total	253						98.0

N = 253, 9 LVs, Q^2 = 0.659, RMSEcv = 0.216, ANOVA p-value < 0.05. CHN: China; TUR: Turkey; RUS: Russia; ESP: Spain.

Categories with n < 20, such as TUR, are not suitable for proper external validation. Therefore, constructing further binary models for broader and better-represented categories was the chosen option to yield reliable results. However, these preliminary findings indicate the potential for developing future models to authenticate pine nuts by country of origin based on mono- and sesquiterpene fingerprint, with the appropriate sampling.

3.2 PLS-DA binary classification models

3.2.1 Multi-species geographical PLS-DA model

Leave 10%-out cross-validation of the 7 binary ESP/non-ESP PLS-DA training models (80% of the samples) provided a 100% of correct classification of both classes for all training models. To verify the reliability of these promising outcomes, models' performances were assessed through external validation. This involved predicting the class of samples from the respective validation sets. The external validation results were expressed as mean values ± standard deviation obtained from the 7 iterations (**Table 2**).

Table 2. External validation of the binary PLS-DA model to discriminate samples into "ESP" and "non-ESP" based on pine nut mono- and sesquiterpene fingerprinting data. Results are mean values (± standard deviation) obtained from seven iterations.

Multi species geographical model: ESP/non-ESP						
	n	ESP (n)	non-ESP (n)	Correct classification (%)	Sensitivity	Specificity
ESP	34	34.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0		1.0 ± 0.0
non-ESP	15	0.0 ± 0.0	15.0 ± 0.0	100.0 ± 0.0	1.0 ± 0.0	
Total	49			100.0 ± 0.0		

N = 196, 5 LVs, Q^2 = 0.969, RMSEcv = 0.086, ANOVA p-value < 0.05.

The external validation outcomes corroborated the leave 10-out cross validation, correctly classifying all samples of the validation sets into either ESP or non-ESP categories, with maximum sensitivity and specificity, and without deviation. These results evidenced the exceptional effectiveness of terpene fingerprinting in distinguishing ESP *P. pinea* kernels from those of other geographical and botanical origins potentially used for counterfeiting, regardless of the specific region, harvest year, or commercial brand.

3.2.2 P. pinea geographical PLS-DA model

To assess the capability and potential limitations of using volatile terpene fingerprint for authenticating the origin of pine nuts, a classification model was built in a more challenging scenario. The goal was to discriminate between samples from the same species, *P. Pinea,* produced in the same country, ESP, but in distinct regions, CAT and CS.

In this case as well, internal validation of the 7 binary PLS-DA models built using the training sets obtained from the 7 iterations of the sample-set splitting achieved 100% accuracy, correctly classifying all the training samples into their respective region of origin. These results were further corroborated by the external validation (**Table 3**), where all CAT pine nuts were correctly classified, providing a specificity of 1. Only one CS sample was misclassified as CAT, resulting in a sensitivity of 0.98 and an overall correct classification rate of 99%.

Table 3. External validation of the Spanish PLS-DA model to discriminate samples into Catalonia and Central Spain based on pine nut mono- and sesquiterpene fingerprinting data. Results are mean values (± standard deviation) obtained from seven iterations.

***P. pinea* geographical model: CAT/CL**						
	n	CS (n)	CAT (n)	Correct classification (%)	Sensitivity	Specificity
CS	19	18.7 ± 0.5	0.3 ± 0.5	98.0 ± 3.0	0.98 ± 0.03	
CAT	15	0.0 ± 0.0	15.0 ± 0.0	100.0 ± 0.0		1.00 ± 0.00
Total	34			99.0 ± 1.0		

N = 136, 5 LVs, Q^2 = 0.907, RMSEcv = 0.158, ANOVA p-value < 0.05. CAT: Catalonia; CS: Central Spain.

These findings confirm the ability of the volatile terpene fingerprint to distinguish pine nuts from different regions, even when they originate from the same species and relatively close geographical areas. These results align with previous studies demonstrating the influence of pedoclimatic factors on the sesquiterpene composition of olive oil, enabling the differentiation of olive oils from various Catalan Protected Designation of Origin (PDO) regions, even when derived from the same cultivar and nearby geographical areas (Quintanilla-Casas et al., 2022b).

The developed model can be applied to verify the identity of pine nuts from these regions/species, but this opens future research for further expanding the model for other regions and species, by developing and validating the models with new samples from these regions and species. Similarly, further research might address open questions, as for instance, which would be the ability of the model in revealing mixed samples from various origins or species, or if the model would work for identifying pine nut identity in complex foods.

3.3 Exploration of PLS-DA regression coefficients

Unlike other food matrices where sesquiterpene fingerprinting has been previously explored for authentication such as virgin olive oil, pine nut terpene fingerprint contains a notable fraction composed of monoterpenoids (Adelina et al.,2021; Rogachev et al., 2015). Given the interconnected biosynthetic pathways of mono- and sesquiterpenoids, they are likely equally influenced by genetic and environmental factors. Consequently, both could potentially contribute to the geographical and botanical differentiation of pine nuts, and for this reason the entire fraction was included for evaluation in this study. To confirm that class discrimination was consistently based on specific terpene patterns, and to ascertain whether both mono- and sesquiterpenoids contributed to the discrimination, we examined the highest significant regression coefficients from both PLS-DA models and tentatively identified the terpenoid structure of the corresponding chromatographic peaks, according to their elution order and full scan MS spectra. It is important to emphasize that the aim was not to conduct a comprehensive study of all discriminant variables or to move towards a targeted analysis. Instead, we focused on the most relevant variables to confirm their terpene nature and to gain insight into their general molecular structure, such as whether they were monoterpene or sesquiterpene hydrocarbons, or oxygenated derivatives.

To explore the variables that had the greatest impact on discriminating between pine nuts classes, we examined the highest significant regression coefficients of both of ESP/non-ESP, and CAT/CS PLS-DA models as explained in section 2.4.3. For both models, plotting the regression coefficients against the variables of the unfolded matrix revealed that the relevant coefficients were distributed along the entire EICs of ions with *m/z* 93, 95, and 119, while being concentrated towards the end of EICs of ions with *m/z* 159, 161, 189, and 204 (**Figure 2**). This is because the former fragment ions are common in both mono- and sesquiterpenoids eluting across the entire chromatogram (Maleknia et al., 2007; Tani et al., 2003; Vichi et al 2006; Vichi et al 2010), whereas the latter are specific of semi-volatile sesquiterpenes with higher retention times (Vichi et al., 2006; Vichi et al 2010).

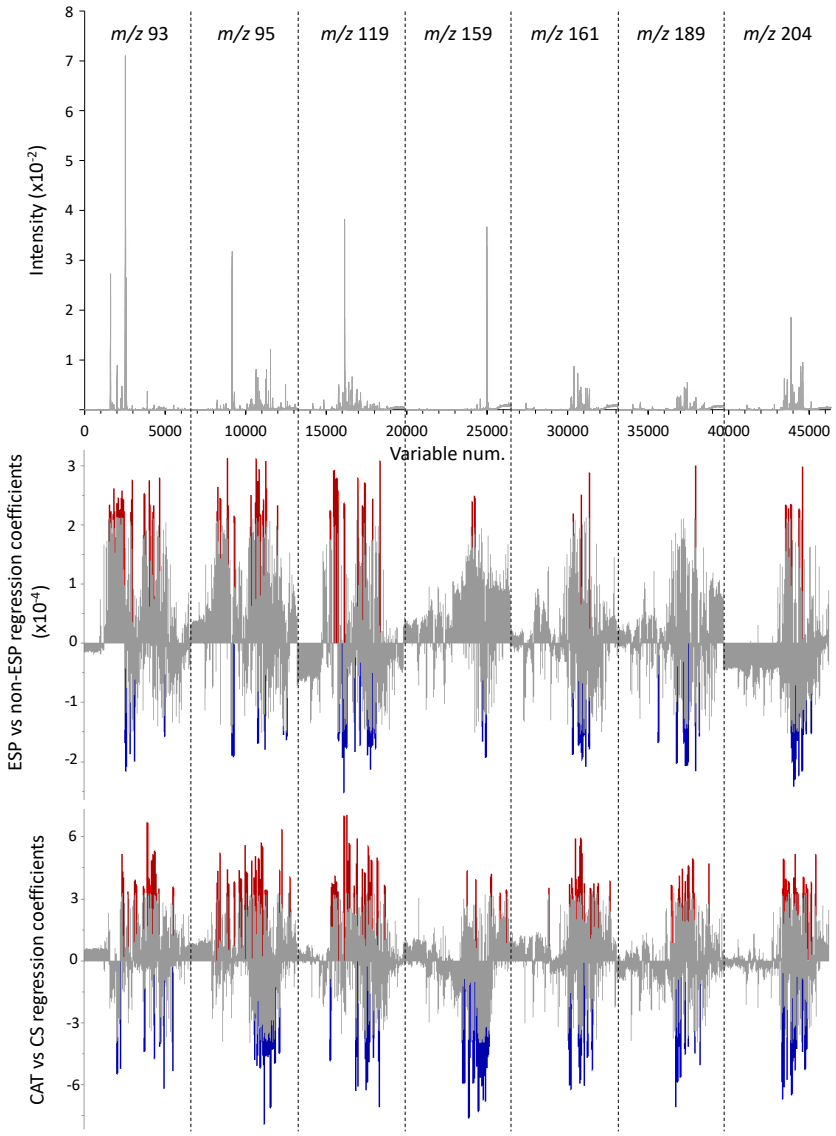


Figure 2. Regression coefficients of the PLS-DA ESP vs non-ESP and CAT vs CS models, plotted against the variables (acquisition points) of the unfolded matrix. For each model, the most relevant coefficients for the prediction of the ESP and CAT classes are highlighted in blue (negative coefficients) and those relevant for non-ESP and CS in red (positive coefficients).

At first glance, these outcomes suggest that both mono- and sesquiterpene families could be regarded as valuable markers for authenticating the botanical and geographical origins of pine nuts, endorsing the hypothesis that both mono- and sesquiterpenes would contribute to discrimination. Specifically, the ESP samples' highest regression coefficients were mostly found in the middle-final section of EICs, being slightly more abundant in EICs *m/z* 119 and 204 (**Figure 2**), and thus, probably attributable to sesquiterpenes. Conversely, non-ESP class was mainly distinguished by compounds detectable in across the entire EICs *m/z* 93, 95, 119 (**Figure 2**), likely including several monoterpenoids. Likewise, both mono- and sesquiterpene

compounds appeared to drive the discrimination between CAT and CS classes. As the most relevant CAT coefficients were in the middle-final section of most of EICs, particularly in EICs *m/z* 95, 159, 204, they were probably attributable to sesquiterpenes. The predominant coefficients distinguishing CS pine nuts, similarly to non-ESP samples, were distributed along the whole EICs at EICs *m/z* 93, 95, 119, several of them probably corresponding to monoterpenoids.

To gain more insight and further support these findings, we examined the compounds related to the most relevant variables in each model. To exemplify some of the tentatively identified compounds that mainly contribute to class distinction, **Figure 3** compares the EICs at *m/z* 93 and 204 corresponding to a non-ESP vs an ESP sample, and to a CS vs a CAT sample. Firstly, it is remarkable that several of these significant variables corresponded to minor compounds or not well-resolved chromatographic peaks, which might hinder their identification and quantification using traditional targeted approaches. This underscores the fingerprinting approach as a more suitable option for their analysis, confirming previous findings (Quintanilla-Casas et al., 2020).

Next, concerning the nature of compounds driving the discrimination between ESP and non-ESP samples, the tentative identification of relevant compounds suggested that the relevant compounds detected in EIC at m/z 93 included both monoterpene hydrocarbons, eluting in the initial part of the chromatogram (e.g. compounds with mass spectra attributable to compounds such as pinene, camphene, sabinene, carene, myrcene, and cymene, relevant for the non-ESP class; limonene and mentha triene, for the ESP class; myrcene and cymene, for the CS class, and pinene and carene, for the CAT class), and their oxygenated derivatives, with higher retention times (e.g. compounds with mass spectra attributable to limonene oxide, relevant for non-ESP and CAT classes; camphor, for non-ESP and CS classes; dihydrocarvone for CS; borneol for non-ESP). On the other hand, the tentative identification of relevant compounds in EIC at *m/z* 204 permitted to assume that several relevant compounds had a sesquiterpene structure (e.g. in the ESP/non-ESP model, compounds with mass spectra matching with that of copaene, junipene, cubebene, cadinene, muurolene and amorphene were relevant for the ESP class, and two not identified compounds whose spectra matched with those of various possible sesquiterpenes distinguished the non-ESP class. In the CAT/ CS model, compounds possibly corresponding to copaene, murolene and an

unidentified sesquiterpene distinguished CS sample, while others, likely junipene and cubebene were relevant for CAT ones). The EIC at *m/z* 204 was selected as an example because it corresponds to the molecular ion of sesquiterpene hydrocarbons (Vichi et al., 2006).

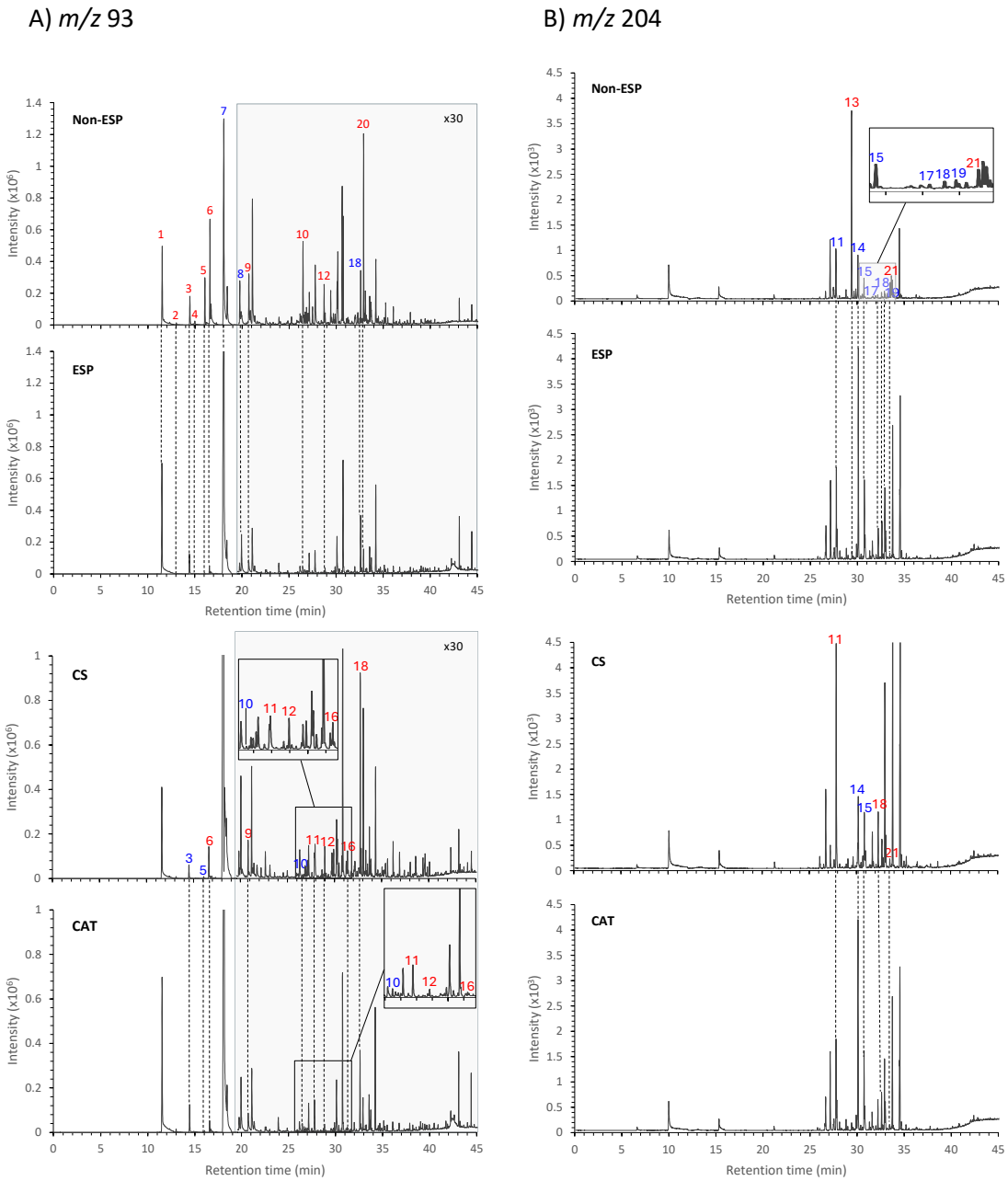


Figure 3. Extracted chromatograms of two representative ions (*m/z* 93, 204) and tentative identification of compounds corresponding to some of the most relevant variables. 1) α-pinene, 2) camphene, 3) β-pinene, 4) sabinene, 5) δ-carene, 6) myrcene, 7) limonene, 8) mentha triene isomer, 9) *p*-cymene, 10) limonene oxide isomer, 11) copaene isomer, 12) camphor, 13) NI sesquiterpene, 14) junipene, 15) cubebene isomer, 16) dihydrocarvone, 17) cadinene isomer, 18) murolene isomer, 19) amorphene, 20) borneol, 21) NI sesquiterpene. Blue: relevant for ESP, CAT; red: relevant for non-ESP, CS.

In summary, the examination of regression coefficients evidenced that numerous variables across all the acquired ions, corresponding to minor and major species, contributed significantly to class discrimination. Both monoterpene and sesquiterpene compounds, including hydrocarbons and their oxygenated derivatives, played a crucial role in the classification. Specifically, monoterpenes seemed to be more characteristic of the non-ESP and CS classes, whereas several compounds with sesquiterpene structure contributed equally to distinguish the origin of samples in both ESP/non-ESP and CAT/ CS models.

4. Conclusions

The volatile and semi-volatile terpene fingerprinting has proven to be a powerful method for authenticating pine nuts, with PLS-DA effectively identifying patterns related to their origin and species, while minimizing variables linked to factors such as harvest year or commercial producer, and therefore, our hypothesis was confirmed. This method provided a high efficiency (> 99%) in the discrimination of pine nuts of different species into ESP and non-ESP classes, and between pine nuts of the same species but from two nearby geographical regions, CAT and CS. Additionally, the preliminary multi-class PLS-DA origin model showed the potential of this method to authenticate multiple geographical origins, provided a sufficiently comprehensive and diverse sample set is used.

Finally, the exploration of PLS-DA regression coefficients confirmed that the compounds related to the variables primarily contributing to the discrimination have a mono- and sesquiterpene structure, including both terpene hydrocarbons and some oxygenated derivatives.

In conclusion, volatile terpene fingerprinting proved to be fast, efficient and straightforward, making it easily to apply to large number of samples in routine laboratories. It could serve as a valuable supporting screening tool for official controls, enhancing their effectiveness and ensuring consumer protection.

Acknowledgments

INSA-UB Maria de Maeztu Unit of Excellence (Grant CEX2021-001234-M) funded by MICIU/AEI/ERDF, EU. INSA-UB authors are part of the LiBiFOOD Consolidated Research Group (2021-SGR-00854) recognised by AGAUR (Generalitat de Catalunya). The authors would like to express their gratitude to INIA-CIFOR and CESEFOR for providing the *P. pinea* samples from Madrid and Castile and Leon.

Research data

Torres-Cobos, Berta; Nicotra, Soriana B; Asensio-Manzano, Cèlia; Aletà, Neus; Teixidó, Anna; Rovira, Mercè; Romero, Agustí; Guardiola Ibarz, Francesc; Vichi, Stefania; Tres Oliver, Alba, 2024, "Pine nut mono- and sesquiterpenoid fingerprints (Extracted Ion Chromatograms) obtained by HS-SPME-GC-MS ", https://doi.org/10.34810/data1730 CORA. Repositori de Dades de Recerca.

Funding

This work was developed in the context of the project TRACENUTS, PID2020-117701RB-I00 financed by MICIU/AEI/https://doi.org/10.130 39/501100011033. B. Torres-Cobos thanks the Spanish Ministry of Universities predoctoral fellowships FPU20/01454.

CRediT authorship contribution statement

B. Torres-Cobos: Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft. C. Asensio: Formal analysis, Data curation, Investigation. S. Nicotra: Data curation, Investigation. N. Aletà: Conceptualization, Resources, Writing – review & editing; A. Teixidó: Resources, Writing – review & editing. M. Rovira: Resources, Writing – review & editing. A. Romero: Resources, Writing – review & editing. F. Guardiola: Supervision, Writing – review & editing. S. Vichi: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing. A. Tres: Conceptualization, Methodology, Supervision, Writing – review & editing.

Declarations 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.

References

Adelina, N.M., Wang, H., Zhang, L., & Zhao, Y. (2021). Comparative analysis of volatile profiles in two grafted pine nuts by headspace-SPME/GC–MS and electronic nose as responses to different roasting conditions. *Food Research International,* 140, Article 110026. https://doi.org/10.1016/j.foodres.2020.110026

Arrabal, C., García-Vallejo, M. C., Cadahia, E., Cortijo. M., & Fernández de Simón, B. (2012). Characterization of two chemotypes of *Pinus pinaster* by their terpene and acid patternsin needles. *Plant Systematics and Evolution*, *298*, 511–522. JSTOR, http://www.jstor.org/stable/43498310 (accessed Apr 16, 2024).

Avula, B., Smillie, T. J., Wang, Y. H., Zweigenbaum, J., & Khan, I. A. (2015). Authentication of true cinnamon (*Cinnamon verum*) utilising direct analysis in real time (DART)-QToF-MS. *Food Additives & Contaminants: Part A*, *32*, 1–8. https://doi.org/10.1080/19440049.2014.981763

Awan, H. U. M., & Pettenella, D. (2017). Pine nuts: a review of recent sanitary conditions and market development. *Forests*, *8*, Article 367. https://doi.org/10.3390/f8100367

Bonari, G., Chytrý, K., Çoban, S., & Chytrý, M. (2020). Natural forests of *Pinus pinea* in western Turkey: a priority for conservation. *Biodiversity and Conservation*, *29*, 3877–3898. https://doi.org /10.1007/s10531-020-02052-z

Destaillats, F., Cruz-Hernandez, C., Giuffrida, F., & Dionisi, F. (2010). Identification of the botanical origin of pine nuts found in food products by gas−liquid chromatography analysis of fatty acid profile. *Journal of Agricultural and Food Chemistry*, *58*, 2082-2087. https://doi.org/10.1021/jf9041722

Destaillats, F., Cruz-Hernandez, C., Giuffrida, F., Dionisi, F., Mostin, M., & Verstegen, G. (2011). Identification of the botanical origin of commercial pine nuts responsible for dysgeusia by gas-liquid chromatography analysis of fatty acid profile. *Journal of Toxicology*, *2011*, Article 316789. https://doi.org/10.1155/2011/316789

Eriksson, L., Trygg, J., & Wold, S. (2008). CV-ANOVA for significance testing of PLS and OPLS® models. Journal of Chemometrics, 22, 594–600. https://doi.org/10.1002/cem.1187

Evaristo I., Batista D., Correia I., Correia P., & Costa R. (2013). Chemical profiling of Portuguese *Pinus pinea* L. nuts and comparative analysis with *Pinus koraiensis* Sieb. & Zucc. commercial kernels. *Options Méditerranéennes. Séries A: Mediterranean Seminars*, *105*, 99–104. http://om.ciheam.org/om/pdf/a105/00006787.pdf

Fardin-Kia, A. R., Handy, S. M., & Rader, J. I. (2012). Characterization of pine nuts in the U.S. market, including those associated with "pine mouth", by GC-FID. *Journal of Agricultural and Food Chemistry*, *60*, 2701–2711. https://doi.org/10.1021/jf205188m

Faria, J.M.S., & Rodrigues, A.M. (2021). Metabolomic variability in the volatile composition of essential oils from *Pinus pinea* and *P. pinaster*. *Biology and Life Sciences Forum*, *2*, Article 14. https://doi.org/10.3390/BDEE2021-09428

Handy, S., Parks, M. B., Deeds, J. R., Liston, A., De Jager, L. S., Luccioli, S., Kwegyir-Afful, E., Fardin-Kia, A. R., Begley, T. H., Rader,J. I., & Diachenko, G. W. (2011). Use of the chloroplast gene ycf1 for the genetic differentiation of pine nuts obtained from consumers experiencing dysgeusia. *Journal of Agricultural and Food Chemistry*, *59*, 10995−11002. https://doi.org/10.1021/jf203215v

International Nut and Dried Fruits (INC), 2023. Nuts & Dried Fruits Statistical Yearbook 2022/20223. https://inc.nutfruit.org/publications/ (accessed Apr 10, 2024).

International Organization for Standardization (ISO). (1991). ISO 7911:1991 Unshelled pine nuts —Specification.

Kim, E., Yang, S., Jeon, B. B., Song, E., & Lee, H. (2024). Terpene compound composition and antioxidant activity of essential oils from needles of *Pinus densiflora*, *Pinus koraiensis*, *Abies holophylla*, and *Juniperus chinensis* by harvest period. *Forests*, *15*, Article 566. https://doi.org/10.3390/f15030566

Loewe-Muñoz, V., Álvarez, A., & Navarro-Cerrillo, R. M. (2018). Morphometric and chemical fruit variability of selected stone pine trees (*Pinus pinea* L.) grown in non-native environments. *Plant Biosystems - An International Journal Dealing with All Aspects of Plant Biology*, *152*, 547–555. https://doi.org/10.1080/11263504.2018.1435587

Loewe, V., Navarro-Cerrillo, R. M., García-Olmo, J., Riccioli, C., & Sánchez-Cuesta, R. (2017). Discriminant analysis of Mediterranean pine nuts (*Pinus pinea* L.) from Chilean plantations by near infrared spectroscopy (NIRS). *Food Control*, *73*, 634–643. https://doi.org/10.1016/j.foodcont.2016.09.012

Maleknia, S. D., Bell, T. L., & Adams, M. A. (2007). PTR-MS analysis of reference and plant-emitted volatile organic compounds. *International Journal of Mass Spectrometry*, *262*, 203-210. https://doi.org/10.1016/j.ijms.2006.11.010

Marti, G., Boccard, J., Mehl, F., Debrus, B., Marcourt, L., Merle, P., Delort, E., Baroux, L., Sommer, H., Rudaz, S., & Wolfender, J. (2014). Comprehensive profiling and marker identification in non-volatile citrus oil residues by mass spectrometry and nuclear magnetic resonance, *Food Chemistry*, *150*, 235–245. https://doi.org/10.1016/j.foodchem.2013.10.103

Matsushita, T., Zhao, J. J., Igura, N., & Shimoda, M. (2018). Authentication of commercial spices based on the similarities between gas chromatographic fingerprints. *Journal of the Science of Food and Agriculture*, *98*, 2989–3000. https://doi.org/10.1002/jsfa.8797

Mikkelsen, A. Æ., Jessen, F., & Ballin, N. Z. (2014). Species determination of pine nuts in commercial samples causing pine nut syndrome. *Food Control*, *40*, 19–25. https://doi.org/10.1016/j.foodcont.2013.11.030

Moscetti, R., Berhe, D. H., Agrimi, M., Haff, R. P., Liang, P., Ferri, S., Monarca, D., & Massantini, R. (2021). Pine nut species recognition using NIR spectroscopy and image analysis. *Journal of Food Engineering*, *292*, Article 110357. https://doi.org/10.1016/j.jfoodeng.2020.110357

Mutke, S. (2022). Producción y comercio de piñones comestibles en el mundo. *Montes*, *147*, 40–43. https://www.revistamontes.net/Buscador.aspx?id=15180 (accessed Apr 14, 2024Z.

Mutke, S., Pastor, A., & Picardo, A. (2013). Toward a traceability of European pine nuts "from forest to fork". *Options Méditerranéennes. Séries A: Mediterranean Seminars*, *105*, 105–109. http://om.ciheam.org/om/pdf/a105/00006788.pdf

Nam, S. L., de la Mata, A. P., Dias, R. P., Harynuk, J. J. (2020). Towards standardization of data normalization strategies to improve urinary metabolomics studies by GC× GC-TOFMS. *Metabolites*, *10*(9), 376. https://doi.org/10.3390/metabo10090376

Nielsen, N. P. V., Carstensen, J. M., & Smedsgaard, J. (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, *805*, 17–35. https://doi.org/10.1016/S0021-9673(98)00021-1

Nikolic, M. V., Jakovljevic, V. L., Bradic, J. V., Tomovic, M. T., Petrovic, B. P., & Petrovic, A. M. (2022). Korean and siberian pine: review of chemical composition and pharmacological profile. *Acta Poloniae Pharmaceutica*, *79*, 785-797. https://doi.org/10.32383/appdr/161040

Quintanilla-Casas, B., Bertin, S., Leik, K., Bustamante, J., Guardiola, F., Valli, E., Bendini, A., Gallina-Toschi, T., Tres, A., & Vichi, S. (2020). Profiling versus fingerprinting analysis of sesquiterpene hydrocarbons for the geographical authentication of extra virgin olive oils. *Food Chemistry*, *307*, Article 125556. https://doi.org/10.1016/j.foodchem.2019.125556

Quintanilla-Casas, B., Torres-Cobos, B., Guardiola, F., Servili, M., Alonso-Salces, R. M., Valli, E., Bendini, A., Gallina Toschi, T. G., Vichi, S., & Tres, A. (2022a). Geographical authentication of virgin olive oil by GCMS sesquiterpene hydrocarbon fingerprint: Verifying EU and single country label-declaration. *Food Chemistry*, *378*, Article 132104. https://doi.org/10.1016/j.foodchem.2022.132104

Quintanilla-Casas, B., Torres-Cobos, B., Guardiola, F., Romero, A., Tres, A., & Vichi, S. (2022b). Geographical authentication of virgin olive oil by GC-MS sesquiterpene hydrocarbon fingerprint: Scaling down to the verification of PDO compliance. *Food Control*, *139*, 109055. https://doi.org/10.1016/j.foodcont.2022.109055

Reed, R. I. (1963). 13 - Mass spectra of terpenes. In F. W. McLafferty (Ed.). Mass spectrometry of organic ions. New York, Academic Press, pp 637-699. ISBN: 9780323142779.

Regulation (EU) No 2023/2429 of the European Parliament and of the Council of 17 August 2023 as regards marketing standards for the fruit and vegetables sector, certain processed fruit and vegetable products and the bananas sector, pp. 1–77. http://data.europa.eu/eli/reg_del/2023/2429/oj

Ríos-Reina, R., Callejón, R. M., & Amigo, J. M. (2021). Feasibility of a rapid and non-destructive methodology for the study and discrimination of pine nuts using near-infrared hyperspectral analysis and chemometrics, *Food Control*, *130*, Article 108365. https://doi.org/10.1016/j.foodcont.2021.108365

Rogachev, A. D., & Salakhutdinov, N. F. (2015). Chemical composition of *Pinus sibirica* (Pinaceae). *Chemistry & Biodiversity*, *12*, 1-53. https://doi.org/10.1002/cbdv.201300195

Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H. P., & Hollender, J. (2014). Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environmental Science & Technology*, *48*, 2097–2098. https://doi.org/10.1021/es5002105

Tani, A., Hayward, S., & Hewitt, C. N. (2003). Measurement of monoterpenes and related compounds by proton transfer reaction-mass spectrometry (PTR-MS). *International Journal of Mass Spectrometry*, *223*, 561-578. https://doi.org/10.1016/S1387-3806(02)00880-1

Torres-Cobos, B., Quintanilla-Casas, B., Romero, A., Ninot, A., Alonso-Salces, R. M., Toschi, T. G., Bendini, A., Guardiola, F., Tres, A., & Vichi, S. (2021). Varietal authentication of virgin olive oil: Proving the efficiency of sesquiterpene fingerprinting for Mediterranean Arbequina oils. Food Control, *128*, Article 108200. https://doi.org/10.1016/j.foodcont.2021.108200

Torres-Cobos, B., Quintanilla-Casas, B., Rovira, M., Romero, A., Guardiola, F., Vichi, S., & Tres, A. (2024). Prospective exploration of hazelnut's unsaponifiable fraction for geographical and varietal authentication: A comparative study of advanced fingerprinting and untargeted profiling techniques. *Food Chemistry*, *441*, Article 138294 https://doi.org/10.1016/j.foodchem.2023.138294

UNECE (2013), UNECE Standard DDP-12 concerning the marketing and commercial quality control of Pine Nut Kernels. 2013 Editions. United Nations, New York and Geneva. https://unece.org/sites/default/files/2023-08/12_PineNutKernels_e.pdf

Vichi, S., Guadayol, J. M., Caixach, J., López-Tamames, E., & Buxaderas, S. (2006). Monoterpene and sesquiterpene hydrocarbons of virgin olive oil by headspace solid- phase microextraction coupled to gas chromatography/mass spectrometry. *Journal of Chromatography A*, *1125*, 117–123. https://doi.org/10.1016/j.chroma.2006.05.029

Vichi, S., Lazzez, A., Grati-Kamoun, N., López-Tamames, E., & Buxaderas, S. (2010). Evolution of sesquiterpene hydrocarbons in virgin olive oil during fruit ripening. *Journal of Agricultural and Food Chemistry*, *58*, 6972–6976. https://doi.org/10.1021/jf100497c

Vichi, S., Riu-Aumatell, M., Mora-Pons, M., Buxaderas, S., & López-Tamames, E. (2005) Characterization of volatiles in different dry gins. *Journal of Agricultural and Food Chemistry*, *53*, 10154–10160. https://doi.org/10.1021/jf058121b

Vichi, S., Tres, A., Quintanilla-Casas, B., Bustamante, J., & Guardiola, F. (2018). Sesquiterpene hydrocarbons, a promising tool for virgin olive oil geographical authentication. In M. Kontominas (Ed.). *Authentication and Detection of Adulteration of Olive Oil*. New York, NY: Nova Science Publishers, Inc. ISBN: 978-1-53614-596-0

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), 209-212. https://doi.org/10.1080/01621459.1927.10502953

CHAPTER 7.
GENERAL DISCUSSION

This chapter provides a comprehensive discussion of the global results obtained in this doctoral thesis, focusing on the common outcomes derived from the strategies developed for verifying food varietal and geographical origin rather than the specific results of each individual study. The discussion is divided into five sections: geographical and varietal authenticity markers, analytical approaches, instrumental techniques, data treatment and chemometric models, and applications and usefulness of the developed methodologies for enhancing food geographical and varietal authenticity. The first four sections correspond to the main aspects considered in the development of reliable food authentication tools. Finally, the last section addresses the applicability and future prospects of the developed methods.

7.1 Geographical and varietal authenticity markers

An authentication model needs to be based on reliable markers, selected in agreement with scientific knowledge that supports that this marker can vary between authentication categories. In the present thesis, this implied that markers should vary between food samples of different origins and cultivars/species, enabling the verification of product claims regarding provenance and varietal origin. Several markers studied in this thesis, including stable isotopes, SH, UF and TAGs, achieved this intended aim. These markers have proven to be directly linked to the food product's growing region and cultivar, with minimal influence from other factors beyond environmental and genetic conditions. However, their suitability varied depending on the commodity and the authentication aspect to be verified. The efficiency of each tested marker for the varietal and geographical authentication of VOO, hazelnuts, and pine nuts are thoroughly discussed in the following subsections.

7.1.1 Isotopic markers

To verify the declared geographical origin of OO, hazelnuts, and pine nuts, various isotopic markers have been studied and authentication models have been built upon them. Three types of isotopic ratios have been explored, both individually and combined: light bioelement stable isotopes ($\delta^{13}C$, $\delta^{2}H$, $\delta^{15}N$, $\delta^{18}O$, $\delta^{34}S$) as markers to verify OO geographical origin ($\delta^{13}C$, $\delta^{2}H$, $\delta^{18}O$) (Publication 2); their combination with compound-specific isotopic ratios, such as $\delta^{13}C$ and $\delta^{2}H$ of FAME for hazelnut

authentication (Publications 4 and 5); and heavy geoelements, specifically Sr ratios, for hazelnut (Publication 4) and pine nut origin (Publication 10).

In agreement with previous literature (section 1.3.2), stable isotopes of light elements were strongly influenced by climatic, environmental and geological factors of the region where agricultural products were grown (Publications 2, 4 and 5). Specifically, correlations between $\delta^{2}H$ and $\delta^{18}O$ with the isotopic ratios of the precipitations at the specific locations were shown in Publications 4 and 5. Additionally, Publication 2 highlighted the effect of temperature and precipitation on the classification accuracy of the isotopic models. This close relationship with pedoclimatic factors, combined with the minimal influence of other variables on the $\delta^{13}C$, $\delta^{2}H$ and $\delta^{18}O$, confirmed the suitability of these isotopes as markers for the geographical authentication of agricultural products [83-87,97]. However, among the bulk isotope values evaluated within this thesis, $\delta^{18}O$ demonstrated the highest discriminatory power in authenticating the provenance of VOO and hazelnuts (Publications 2, 4 and 5). Similarly, the $\delta^{13}C$ and $\delta^{2}H$ of FAMEs were also linked to environmental and climatic conditions, proving useful for discriminating among samples from different regions of origin and enhancing classification accuracy (Publications 4 and 5).

On the other hand, $\delta^{15}N$ and $\delta^{34}S$, although potentially related to geographical factors, were found to be inadequate for verifying the geographical origin of agri-food products as they were significantly influenced by agronomic practices, particularly fertilisation (Publication 4).

The $^{87}Sr/^{86}Sr$, a heavy geoelement whose main source in plants is the soil, showed promising results in the discrimination of hazelnuts from different origins and was in accordance with the geological characteristics of the production areas. Although studies have reported the potential impact of fertilizers on the Sr isotopic composition of agricultural products [258,259], this factor did not appear to critically affect the results in the studied case (Publication 4). However, due to the laborious sample preparation steps and higher operating costs involved in Sr isotopic analysis, its application was deemed inadequate for routine analysis and more suited as a confirmatory tool for uncertain samples. In light of the promising results obtained from the analysis of $^{87}Sr/^{86}Sr$ in hazelnuts, the Sr isotopic ratio was also studied as a

geographical marker of pine nuts (Publication 10). The influence of agronomic practices was expected to be negligible in this case, as pine nuts are collected from naturally grown forests. Although significant differences were found between Asian and Mediterranean pine nuts, these differences were insufficient to achieve a clear distinction between the origins. In fact, the heterogeneous bedrock across the extensive study areas resulted in considerable variability in the Sr ratios within each region, leading to an overlap among the categories. These findings underscore the limitation of using a single isotopic marker to distinguish samples from different provenances.

Apart from the suitability of isotopes for each authentication issue, it is noteworthy to mention various general aspects that concern all the isotopic markers examined. As reported in section 1.4.4, the primary strengths of isotopic analysis include its robustness, high reproducibility and precision, largely due to the use of CRMs to correct bias. These features facilitate straightforward transferability and implementation of these methods by accredited laboratories [81] and support their use in official control methods [84]. However, isotopic markers faced limitations in discriminating food origin compared to other markers (SH, UF, TAGs). This was primarily due to the limited ability of isotopic markers to differentiate between geographically close regions with similar pedoclimatic conditions, as observed with VOO and hazelnuts (Publications 2 and 5), or from very broad geographical areas, as in the case of pine nuts (Publication 10). Additionally, stable isotopes are highly sensitive to the harvest year, which hindered the classification of VOO samples from different harvest years using a single model (Publication 2). These drawbacks may be attributed to the limited analytical information provided by targeted methods compared to untargeted ones (Publication 1), ultimately reducing the efficiency of the classification. The performance of the multi-isotope approach could be enhanced by combining it with other markers, such as mineral profile [97,104] or metabolites [205,206]. This strategy has proven effective in pine nut authentication, where elements like Rb and Sr showed promising potential in distinguishing Mediterranean pine nuts from the Asian ones (Publication 10).

7.1.2 Metabolic markers

Volatile and semi-volatile terpene compounds were tested as markers for botanical and geographical food authenticity. Specifically, SH fingerprinting yielded excellent results in verifying VOOs provenance (Publication 2) and cultivar (Publication 3). Additionally, when combined with the monoterpene fraction, they proved to be an excellent tool for authenticating pine nuts from different species and origins (Publication 10). However, in the case of hazelnuts, these markers were unsuitable for authentication due to their minimal concentration in the kernels and oil (section 5.3.1).

The success of these terpene markers is attributed to the influence of genetic and environmental factors on their composition in plants, enabling a direct correlation with the plant cultivar and growing area. Moreover, they are relatively stable to oxidation and minimally affected by storage and processing conditions. Therefore, mono- and sesquiterpene compounds meet all the criteria of appropriate authenticity markers. As long as their concentration in the plant product is sufficiently high to be detected and measured, they are powerful botanical and geographical authenticity tools for agricultural products.

Since SH were not adequate markers for hazelnut authenticity, other options were explored. For this purpose, UF and TAGs were selected as promising candidates as several studies have reported variations in their composition among hazelnuts from different origins and cultivars [44, 133,153-155]. Although **UF** constitutes a minor part of the lipid fraction, it contains several secondary metabolites whose concentration and composition are influenced by genetic and pedoclimatic factors. Additionally, most of them are relatively stable non-volatile species contributing to produce reliable classification models. The studies conducted in this thesis (Publications 7 and 8) have demonstrated that models based on the UF fingerprint can simultaneously verify hazelnut cultivar and origin with high accuracy, sensitivity, and specificity, making them reliable tools for geographical and varietal authentication of hazelnuts. However, their analysis is laborious and time-consuming, requiring multiple sample preparation steps which makes it difficult to apply for routine screening. Instead, it could serve as a valuable confirmatory method to support fast-screening tools.

In this context, **TAGs**, the main constituents of the hazelnut lipid fraction, were tested to enhance analytical efficiency in hazelnut authentication (Publication 8). Their composition and FA arrangement are determined by genetics and environmental factors [44,133] and are not easily degraded or affected by processing or storage procedures. For this purpose, it was essential to dispose of an analytical method that required minimal sample preparation and that employed affordable and automatable instrumentation. Thus, a method was successfully developed to analyse TAGs in highly lipid-rich food matrices by HT-GC-MS (Publication 6). Then, this analytical method was applied to analyse TAGs in oil extracted from hazelnuts, without requiring any further optimization (Publication 8).

In order to develop the hazelnut authentication models, TAG fingerprint was determined in a large sample set of hazelnut samples encompassing a wide variety of cultivars, harvest years, and regions of origin (Publication 8). In general, the model achieved satisfactory results with good classification accuracy comparable to other more laborious methods. Nevertheless, UF clearly outperformed TAGs in both cultivar and origin discrimination, displaying unmatched effectiveness in authenticating hazelnut. For this reason, a combined strategy was applied integrating both TAG and UF models. This approach significantly reduced the analytical workload while maintaining high classification accuracy, proving to be a powerful tool for both varietal and geographical authentication of hazelnuts (Publication 8).

Finally, a **broader approach** was employed to simultaneously study a **wide range of metabolic analytes** by analysing hazelnut samples using **NIR** and **MIR** spectroscopic techniques (Publication 9). The outcomes of this study showed that the untargeted analysis of multiple metabolites provided high predictive accuracy in classifying hazelnuts according to their geographical area of origin and cultivar, matching the accuracy obtained with the UF fingerprinting method. Additionally, this method offered valuable insights into the key compounds driving hazelnut discrimination, primarily proteins and lipids, proving to be a fit-for-purpose screening method for verifying the varietal and geographical origin of hazelnuts.

Overall, across all the conducted studies, methods based on metabolic markers outperformed isotopic approaches for the geographical authentication of VOO,

hazelnuts, and pine nuts. Metabolic approaches proved more efficient, achieving accurate classification while simultaneously verifying both the varietal and geographical origins of food products, confirming hypotheses 4.1 and 4.2. Among the metabolic methods, UF fingerprinting and NIR spectroscopy were more effective than MIR and TAG fingerprinting. However, combining rapid TAG fingerprinting with the UF method offered a good balance between efficiency and workload. Lastly, NIR spectroscopy was just as effective as these methods, with even less effort, rendering it the fastest approach for hazelnut geographical and varietal authentication.

7.2 Analytical approaches

Although authenticity markers are key in an authentication model as they hold the chemical information necessary to discriminate various categories (origins, cultivars), the overall success of an authentication strategy relies on their combination with several other factors: the analytical approach, the method and instrumental technique, the data treatment, and the chemometric tools employed. Regarding the analytical approach, as reviewed in section 1.4.1, the state-of-the-art in food authentication involves the use of untargeted approaches, as they incorporate more analytical information in determining food authenticity. Nevertheless, for most markers, targeted approaches are still advantageous in terms of transferability.

7.2.1 Isotopic targeted approach

A targeted approach was followed in all the isotopic studies conducted, focusing on the identification and quantification of various isotopic ratios (Publications 2, 4, 5 and 10). This targeted approach was not applied to metabolomic methods, as previous research indicated it to be less effective than untargeted analysis for food authentication based on a large number of metabolic markers [168,Publication 1].

Targeted approach facilitates reproducibility and transferability among different laboratories and control bodies, which are key elements for the adoption of these tools as reference methods by regulatory authorities. However, although quite satisfactory results were obtained in the geographical and varietal discrimination of VOOs and hazelnuts, this approach was consistently outperformed by untargeted methods.

7.2.2 Metabolic untargeted profiling approach

Unlike targeted approaches, which analyse a predefined limited number of compounds, providing fewer comprehensive data for fraud detection thereby diminishing authentication efficiency, the untargeted approaches were more effective for complex analytical problems such as verifying the geographical and botanical origin of agricultural food products. For instance, in assessing VOO purity, the untargeted methods outperformed the official targeted method by enabling the detection of small quantities of adulterant high-oleic vegetable oils in VOO (Publication 6).

One of the untargeted approaches applied in this thesis has been the untargeted profiling approach using Parallel factor analysis 2 (PARAFAC2) models applied to three-way GC-MS data (samples × retention time × m/z) of the UF of hazelnuts (Publication 7). This tool is integrated into the software PARADISe (PARAllel factor analysis 2 Deconvolution and Identification System), which enhances the robustness of the results by reducing user interaction and optimizing analysis time (Publication 1). PARAFAC2 deconvoluted the Total Ion Chromatograms of the UF into a peak table containing the relative concentrations of pure compounds, and their mass spectra. In total, 155 compounds were extracted. By this, it addressed common chromatographic data issues such as co-eluting peaks, low signal-to-noise ratio, and retention time shifts.

This approach offered several advantages, the main one being its ability to provide chemical interpretable results. This makes it a suitable tool for analysing samples with unknown compositions, as it provides pure mass spectra that simplify the identification of analytes. For example, it revealed that sterols, along with linear and terpene alcohols, were the main compounds driving the discrimination of hazelnuts (Publication 7). Consequently, it also avoided the time-consuming steps of identification and quantification of targeted approaches while preserving all the valuable information. Additionally, conventional strategies for targeted methods can be adapted to evaluate the performance of untargeted profiling, facilitating its transferability between laboratories. However, as demonstrated in Publication 7 and supported by other literature data [260], the fingerprinting approach proved superior to untargeted profiling. Fingerprinting was more effective in extracting detailed

information from chromatographic data, including minor discriminant species, which was crucial for the development of authentication models. Therefore, fingerprinting was chosen for the subsequent studies on hazelnut geographical and varietal authentication (Publication 8).

7.2.3 Metabolic untargeted fingerprinting approach

The fingerprinting approach has been applied in several studies throughout this thesis to analyse GC-MS data (Publications 2, 3, 6, 7, 8, and 11) and spectroscopic data (Publication 9), with the aim of developing authentication models.

In the case of GC-MS, a three-way data matrix (samples × retention time × *m/z*) was obtained. While this type of data typically requires complex multi-way chemometric algorithms, a simpler approach was applied by unfolding the data into a two-way matrix (samples × retention time). This was achieved through the concatenation of the acquired *m/z* extracted ion chromatograms in the row-wise direction (Publication 1). This simplification allowed for the application of bilinear methods.

Besides, a key aspect of fingerprinting analysis is **signal preprocessing**. Regarding **chromatographic** fingperinting, ithe present studies have shown that applying baseline correction and normalization prior to alignment can significantly enhance performance. For instance, in the TAG fingerprinting (Publication 4), baseline correction was implemented using asymmetric least squares, a method proven to be highly effective for chromatographic data [175]. This correction was essential for addressing baseline level variations among samples and mitigating the drift caused by column bleeding, which is particularly significant at the high temperatures used in the HT-GC-MS method. Normalization was also applied to chromatographic fingerprinting data (Publication 2, 4 and 11) to correct intensity variations between batches and balance weights among samples. Finally, the most critical factor in chromatographic fingerprinting is the retention time shift among samples, an issue not usually encountered in other fingerprinting techniques such as spectroscopy. To address this issue, the COW algorithm was applied [183]. This flexible alignment technique adjusts chromatographic profiles by shifting, stretching, and compressing the peaks without causing peak swapping. The COW method also allows for the automatic optimization of

warping parameters, such as slack size and segment length, which enhances its implementation and effectiveness (**Figure 17**).

However, for those datasets that included multiple batches (Publications 2-8 and 11) a preliminary prealignment of each analytical batch was necessary before performing the general alignment across all batches. This step was crucial to reduce computational cost and ensure optimal results.

This signal preprocessing was fundamental to ensure reliable comparisons among samples, which is essential for developing accurate and robust authentication models.

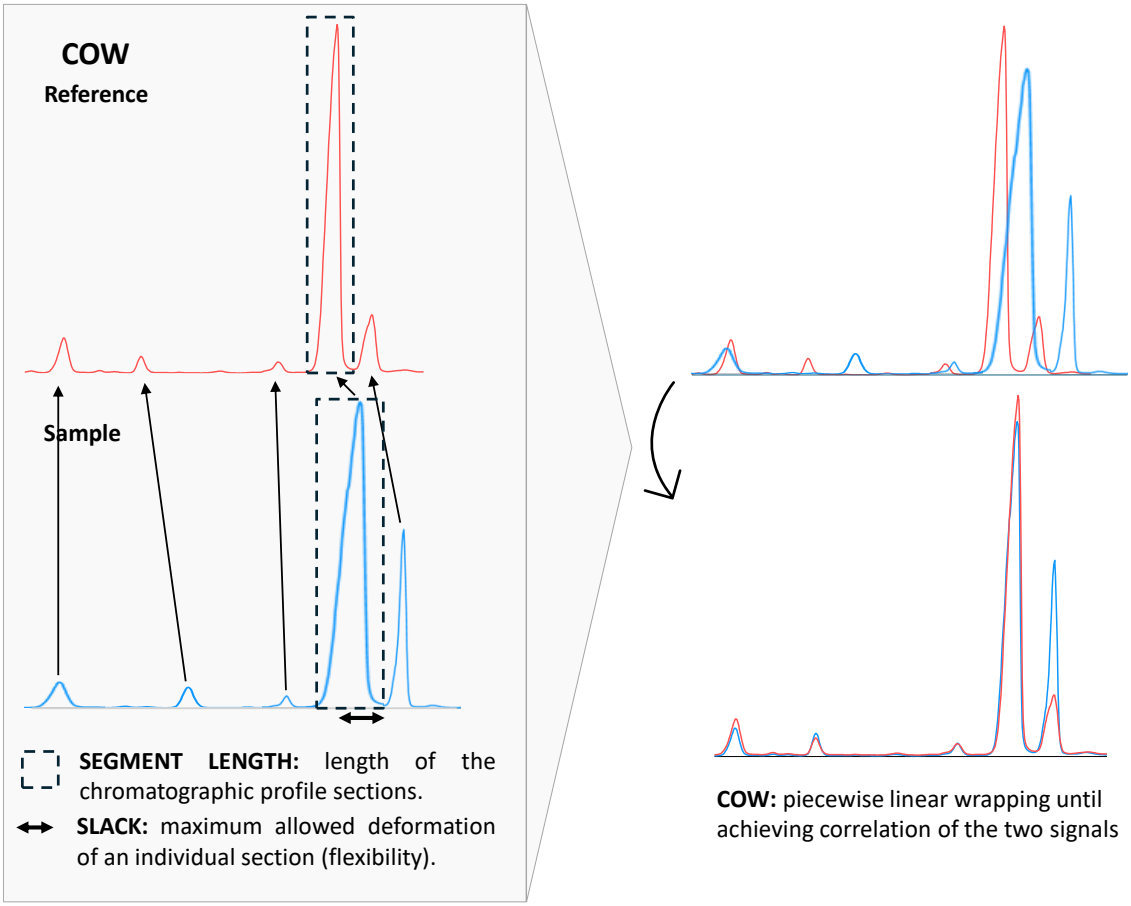


Figure 17. Cow alignment diagram. COW: correlation optimized wrapping.

For **spectroscopic data** fingerprints, no prior alignment is required due to the absence of time shifts among samples. However, other types of signal preprocessing (derivatives, Savitzky–Golay smoothing, SNV transformation) were applied to address instrumental issues and enhance the signal-to-noise ratio (Publication 9).

Overall, the fingerprinting approach demonstrated exceptional effectiveness in extracting valuable sample information for developing robust authentication models. It was especially advantageous for complex data involving multiple species and minor compounds, where identification and quantification were challenging due to low concentrations, co-elution, or similar instrumental responses. This was particularly true for SH, which, despite their high structural variability, exhibit very similar mass spectra (Publication 2, 3 and 11). This approach has also demonstrated successful in TAG analysis (Publications 6 and 8), surpassing targeted official method and proving to be a fit-for-purpose screening tool capable of analysing a large number of samples.

Moreover, fingerprinting was the most suitable approach for the hazelnut UF data (Publication 7), outperforming untargeted profiling by providing higher sensitivity and classification accuracy. Its success in this case relies on the higher sensitivity of the SIM acquisition mode in the fingerprinting approach, which allowed for the detection of minor relevant species that were overlooked by the untargeted profiling's full scan acquisition. Therefore, the selection of the ion for SIM based on the prior knowledge of the chemical compound families present in the samples (those in the UF fraction in this case) was key for the successfulness of the fingerprinting approach.

On the other hand, the fingerprinting approach was also applied to spectroscopic data obtained from the analysis of hazelnuts by NIR and MIR techniques with exceeding outcomes (Publication 9). This approach is widely used in the spectroscopic field because it allows for fast, simple, and simultaneous analysis of a large variety of compounds found in food samples, such as lipids, proteins, and carbohydrates [173]. Moreover, unlike other fingerprinting techniques like GC-MS, spectroscopic techniques offer higher reproducibility [Publication 1,261,262] which eases their transferability across laboratories.

7.3 Instrumental techniques

The instrumental technique employed is a key factor in assessing the availability and applicability of an authentication method. Ideally, this technique should be fast, cost-effective, and widely accessible to many laboratories to facilitate the implementation

of the authentication method. The main instrumental techniques used in the conducted studies are thoroughly discussed in the following sections.

7.3.1 IRMS and MC-ICP-MS

The most common technique for stable isotope analysis is IRMS [84,88]. When coupled with an elemental analyser (EA), it allows for the straightforward analysis of multiple light bio-elements ($^{13}C/^{12}C$, $^{2}H/^{1}H$, $^{15}N/^{14}N$, $^{18}O/^{16}O$ and $^{34}S/^{32}S$) in solid and liquid samples with minimal sample preparation [84,88,Publication 4].

In addition, CSIA has become increasingly popular, as it provides additional and unique isotopic information about the food sample, reflecting factors such as geographical origin and cultivation practices [88,101,102,263,Publication 7]. CSIA is performed by combining IRMS with a separation technique such as GC, which isolates individual compounds from a complex mixture and determines their isotopic ratios.
For heavier elements like Sr, which cannot be converted to gases, MC-ICP-MS is used. This high-resolution technique provides precise measurements of isotopic ratios [88,113].

These techniques have been applied to the isotopic analysis of VOO, hazelnuts and pine nuts, for verifying geographical authenticity (Publications 2, 4, 5 and 10). They offer high precision, sensitivity, and accuracy in measure, require only small sample sizes, and are applicable to almost any type of food or beverage [84,88]. By reporting stable isotope abundances relative to certified reference materials, these methods achieved high measurement precision under repeatability conditions [81,84,Publication 2]. This unifies isotope measurements, resulting in a robust and easily transferable methodology.

However, the routine application of these techniques for food geographical authentication faces several challenges. The high cost of instrumentation and maintenance, along with the complex and labour-intensive sample preparations required in some cases—particularly for Sr analysis by MC-ICP-MS—pose significant barriers. For instance, in the geographical and varietal authentication of hazelnuts, the Sr isotope was excluded in favour of more cost-effective bioelements, which offered comparable discrimination accuracy and are more easily applicable to routine analysis

(Publications 4 and 5). Additionally, operating these instruments requires specialized knowledge and trained personnel, and the potential for contamination during sample preparation or analysis demands strict laboratory protocols.

7.3.2 GC-MS

Most of the metabolomic methods developed in this thesis analyse markers using GC coupled with low-resolution MS. This technique is commonly available in control laboratories, and, implementing these methods generally requires minimal investment in instrumentation. Additionally, this technique is easily automatable, requires minimal sample manipulation, is environmentally friendly due to being solvent-free, and offers high sensitivity, especially in SIM acquisition mode. When coupled with HS-SPME, it enables fast and sensitive analysis of volatile and semi-volatile compounds from small amounts of solid and liquid samples without any prior preparation.

Moreover, the mass spectra obtained in full scan acquisition mode (Publication 7) and the available libraries facilitate the identification of specific compounds or compound families that contribute to the differentiation among samples. Additionally, they enable to perform a “back to target” approach, by identifying the most relevant compounds in discrimination and designing a targeted method focused on the analysis of these few key target compounds.

In the conducted works, the analysis of VOO, hazelnuts and pine nuts through GC-MS provided comprehensive molecular-level information essential in developing varietal and geographical authentication models (Publications 2-8 and 11). GC-MS-based methods have proven more advantageous for generating data to build authentication models, likely due to the greater amount of information provided. For instance, HT-GC-MS for TAG analysis resulted more efficient than HPLC-RID and even FIA-HESI-HRMS for differentiating based on subtle differences in TAG composition (Publication 6).

This underscores the suitability of GC-MS as an effective technique for authentication screening methods in routine controls. However, its successfulness also depends on the specific extraction methods needed to obtain the authenticity markers and the subsequent data treatment.

7.3.3 Spectroscopic techniques

Lastly, spectroscopic techniques, including NIR and MIR, were applied to analyse hazelnuts for geographical and varietal authentication. Specifically, three instrumental approaches were tested: NIR spectroscopy with a benchtop instrument, NIR spectroscopy using a handheld device, and MIR spectroscopy (Publication 9).

These spectroscopic methods are significantly faster than the other techniques explored, enabling the direct analysis of samples (whether whole kernels or ground) in just minutes, or even less. Their operational systems are straightforward and simple, and since no solvents are used in either sample preparation or instrumental analysis, these methods are considered green techniques. Additionally, they are cost-effective and non-destructive, preserving the sample for further analysis if needed. These intrinsic characteristics make spectroscopic techniques highly adaptable for on-site and online applications, making them ideal for routine analysis.

On the other hand, these techniques are less sensitive and specific, often leading to overlapping bands and complex spectra, particularly when analysing intricate mixtures. This can hinder interpretability and complicate the differentiation of sample components. For instance, NIR spectra of hazelnuts revealed bands that could be related to lipids, proteins, complex carbohydrates and compounds containing hydroxy groups, that were relevant for the discrimination between cultivars and origins. Additionally, NIR and MIR spectroscopies are sensitive to environmental conditions such as temperature, humidity, and pressure, as well as physical properties of the sample like particle size, which can introduce variability in the results. Therefore, it is essential to ensure that samples are analysed under consistent conditions: at the same temperature (ideally room temperature), in a dry environment, and with a homogeneous particle size for ground samples.

The application of NIR spectroscopy in hazelnut authentication enabled the rapid analysis of a broad spectrum of chemical compounds and yielded highly successful, validated results for simultaneously verifying the geographical and botanical origin of hazelnuts, demonstrating its effectiveness as a fit-for-purpose screening technique (Publication 9).

7.4 Data treatment and chemometric models

7.4.1 Univariate statistics

Various descriptive statistics including mean, median, standard deviation, interquartile range, Shapiro-Wilk test, parametric tests (t-test, one-way ANOVA), and non-parametric Kruskal-Wallis and Mann-Whitney U tests, were applied to the isotope data of VOOs (Publication 2), hazelnuts (Publications 4 and 5) and pine nuts (Publication 11). These methods were employed to describe the data, assess differences in isotopic ratios across different origins, evaluate the impact of harvest year, and investigate the influence of environmental factors such as precipitation and temperature (Publications 2, 4 and 5), as well as the impact of agricultural practices such as fertilisation, and geological factors (Publication 6).

Although these methods were useful for examining the relationships between isotopic ratios and geographical, climatic, geogenic, and anthropogenic factors, their results were inconclusive for reliably determining the geographical origin of the studied food products. Significant differences in isotopic ratios were observed across regions for VOOs (Publication 2), hazelnuts (Publications 6 and 7) and pine nuts (Publication 10); however, this alone was insufficient to accurately discriminate the provenance of the samples due to the overall dispersion of their isotopic values, which often resulted in considerable overlap. As a result, single thresholds for these isotope markers could not efficiently distinguish sample classes. This highlighted the importance of exploring multivariate classification methods to evaluate more accurately their discrimination potential.

7.4.2 Unsupervised methods

Unsupervised methods, particularly PCA, were employed to conduct the initial exploration of quantitative (isotopes), profiling (UF, TAGs) and two-way fingerprinting (SH, UF, TAGs, spectroscopic) data. Their application aimed to examine the natural distribution of the samples, identify possible trends, and detect outliers according to the parameters Hotelling's T^2 and Q residuals [190-192]. Score plots facilitated the investigation of sample clustering, helping to determine which factors—such as cultivar, origin, batch effect, or harvest year—might influence this distribution and to

assess whether significant overlap existed among sample groups (Publications 3 and 7). However, due to the data's complexity and the high presence of non-informative variables (noise), principal components often proved insufficient to fully explain the variance between different categories, especially when chromatographic fingerprints were involved.

On the other hand, PARAFAC2, another unsupervised method, was applied to deconvolute the three-way GC-MS data of the hazelnut UF as an unsupervised method (Publication 7). This method provided the resolved mass spectra and the peak table containing the relative concentration of each analyte. The resulting data matrix, which included the peak areas for each sample, was then employed to develop effective supervised classification models. Moreover, the pure mass spectra obtained through PARAFAC2 method eased the chemical interpretation of the subsequent model facilitating the identification of the chemical families that played a crucial role in the classification of samples according to their origin or cultivar.

Given that the primary objective of the studies was to predict qualitative variables, such as the cultivar and origin, supervised methods were applied in most studies of this thesis for developing validated classification models. Therefore, model calibration, which included outlier detection, preprocessing methods, and the selection of the optimal number of components, was carried out using the supervised technique.

7.4.3 Supervised methods: Classification models

Discriminant (PLS-DA) and class modelling techniques (SIMCA) were assayed as supervised classification techniques throughout all PoC studies due to their suitability for building classification models with categorical variables (section 1.4.3). As SIMCA relies on a well-represented sample category rather than requiring a representative sample set of multiple categories, it is well-suited for classification in cases where the non-authentic category is not finite. This makes it effective for determining whether a sample belongs to the defined authentic category or not. However, SIMCA models did not perform as effectively as expected (data not shown). This might be because SIMCA, being PCA-based, struggles with data containing a significant amount of non-informative variables or noise, as it is the case of chromatographic fingerprints.

Moreover, the presence of closely related or highly similar classes may further challenge discrimination using SIMCA models. In contrast, PLS-DA is better suited to handle noisy data and is particularly effective with datasets that include a large number of highly correlated variables, characteristics frequently encountered in the studies presented.

Binary (only two categories) or multi-class (more than two categories) PLS-DA models were developed depending on the authentication purpose. **Binary PLS-DA** models were used to authenticate one specific cultivar or region of origin from all the others: Arbequina vs other VOO cultivars (Publication 3), TG vs other hazelnut cultivars (Publications 7-9); Italian VOO vs VOO from other countries (Publication 2), Spanish pine nuts vs pine nuts from other provenances (Publication 11).

Special efforts were made to incorporate a wide range of samples to ensure their variability was as representative as possible of the real market, especially for the non-authentic class, which is the most challenging to represent due to its broad scope.. However, in some cases, this proved especially challenging (e.g., non-EU samples: covering all possible countries; non-Arbequina, non-TG: covering all other cultivars worldwide). As a result, the interpretation of the study's findings takes this limitation into account, acknowledging that the developed models are intended for use under conditions similar to those in which they were developed.

Multi-class PLS-DA models were applied to differentiate among different cultivars or geographical origins, whether they were countries or regions close to each other or further away: VOO from three Italian regions CAL/SIC/PUG (Publication 2), seven VOO cultivars (Publication 3), hazelnuts from four countries of origin CHL/ESP/GEO/ITA (Publications 6-9) and pine nuts from four countries of origin CHN/TUR/RUS/ESP (Publication 11).

Nonetheless, in cases where the sample size for one or more specific categories was insufficient to adequately represent their variability and enable external validation (Publication 7), or when the model struggled to distinguish between very similar categories (e.g. geographically close regions with similar pedoclimatic conditions) (Publication 5), these categories were merged into a broader, more general group

provided that this grouping was meaningful for the application to authenticate the product (e.g., producing areas with similar market prices and risk of mislabelling or those that can be merged due to geographical proximity). In light of this, the PoC study in Publication 7, adjusted the multi-class PLS-DA approach for determining geographical origin (CHL/ESP/ITA) to binary classification models (EU vs. non-EU and ESP vs. ITA) due to the limited number of tested samples, aiming to enhance performance. Later, when the sample set was enriched (Publication 8), multi-class PLS-DA origin models (CHL/ESP/GEO/ITA) were successfully developed and validated.

One of the criticisms that discriminant techniques have received when used in authentication is that they are expected to always classify samples into one of the categories included in the model, even if the sample is extraneous. However, outlier detection would rule out the possibility of classifiying extraneuous samples in one of the categories. Moreover, the classification criteria applied for all models in this thesis, included the possibility of a 'no class' answer. In PLS-DA binary models, categories are expressed as PLS dummy variables (1 for the non-authentic category and 0 for the authentic category). Since multi-class PLS-DA models work as multiple binary models of each category against the others, a dummy Y matrix holds as many classification vectors as categories, and each vector has values of 1 for the specific category and 0 for all the other categories. In this case, the PLS predicted value (PV) of each sample was used for classification into one category or the other according to the highest PV criteria, provided it was above the classification threshold, usually set at 0.5 unless otherwise indicated. Samples with PVs below the classification threshold for all categories were not assigned to any category. This means that extraneous samples would fall into the 'no class' category, thereby overcoming one of the main drawbacks of using discriminant techniques for authentication. In any case, it is important to bear in mind that models are to be applied to samples similar to those used in model validation. In this respect, EU vs non-EU models should be further tested with samples from countries not included in the sample sets.

Before model building, the analysed sample set was divided into **training and validation sets** following a stratified random sampling strategy, with 80% of samples from each category allocated to the training set and 20% to the validation set. The

generated training and validation sets preserved the balance in the proportions of the original sample set. The 80/20 balance is usually used in model validation as it leaves a sufficient part of samples to cover all possible natural variability for model calibration, while disposing of a certain number of samples to obtain reliable external validation. This process was repeated multiple times (ranging from 3 to 7 iterations) to assess the impact of the sample set composition and enhance the robustness of the external validation. This iterative process assured that external validation was robust and not dependent on influential samples that could have led to biased results. In each iteration, PLS-DA models were developed, calibrated and internally validated by leave-10%-out cross-validation exclusively using the samples in training set of that iteration. During calibration, outlier detection was conducted by evaluating Hotelling's T^2 and Q residuals in a model with a low number of LVs to prevent noise from hindering the identification of anomalous samples. The optimal preprocessing method and number of LVs were then selected based on the criterion of the lowest root mean squared error of cross-validation (RMSEcv). Finally, ANOVA (p-value) on the cross-valdiated predicteve residuals, along with permutation tests (n = 20), were performed to assess potential model overfitting.

A general scheme of the chemometric techniques applied to build classification models is represented in **Figure 18**.

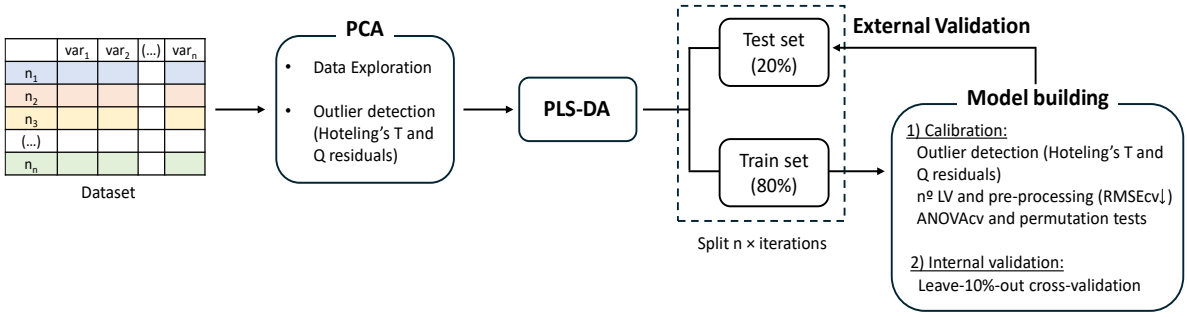


Figure 18. General scheme of the chemometric techniques applied to build PLS-DA classification models. PCA: principal component analysis, PLS-DA: partial least squares discriminant analysis, LV: latent variables, RMSEcv: root mean squared error of cross-validation.

To further enhance the performance of the models, in Publications 2, 5 and 8 **classification thresholds** were optimized by analysing receiver operating characteristics (ROC). ROC curves were generated with the PVs obtained in the leave-10%-out cross-

validation plotting the sensitivity (true positives/ [true positives + false negatives]) and 1-specificity (1 - [true negatives/ (true negatives + false positives)]) values obtained when the PV threshold to assign samples to a diagnostic category varies [264]. ROC analysis was applied to PV values from each PLS-DA model and category. The optimised thresholds were those leading to the maximum value of sensitivity + specificity (Publications 2, 5 and 8).

In addition, threshold optimization proved valuable in establishing an uncertainty range for classification, rather than relying on a single threshold, as demonstrated in Publication 8. Samples with PVs falling within the uncertainty range were treated as boundary samples and were not assigned to any class. This approach was integrated into a combined TAG and UF strategy. Initially, samples were analysed using the fast screening method for TAGs, and those within the uncertainty range were further assessed using the UF fingerprinting method. This combined strategy significantly enhanced analytical efficiency by reducing workload and achieved a high classification accuracy.

It is important to note that classification thresholds should be updated whenever models are enriched with additional samples. In this context, in future implementation of these models, the responsibility for setting classification thresholds should be entrusted to the corresponding oversight authorities. These authorities may adjust the thresholds according to the specific control objectives, balancing the acceptable risk of misclassification. For instance, in a screening approach, thresholds could be adjusted to minimize false negatives, even if it leads to an increase in false positives, ensuring that no fraudulent sample escape detection. In adulteration issues, the authentic and the non-authentic class are clearly defined by the authentication problem itself. However, for certain geographical and varietal authentication models, a prior agreement should be established to define the "authentic category," because this one will define the false negatives. The “authentic class” is typically the most expensive or valuable product susceptible to fraud (e.g., Italian VOOs, Spanish and Italian hazelnuts, Spanish *P. pinea* pine nuts), and the "non-authentic category" would then comprise lower-cost products that are falsely marketed as authentic to gain economic profit.

Finally, to ensure the reliability of the models and prevent overfitting or overly optimistic outcomes, conducting **external validation** is essential. This process involves predicting the class of samples that were never included in the model development—those in the designated validation set. External validation was performed in all the studies (Publications 2-9, and 11), excluding PoC studies that lacked a sufficient number of samples to accurately represent the categories (Publication 4 and Publication 10). Successful external validation results were achieved for most of the developed methods (Publications 2, 3, 5,8, 9 and 11) for food geographical and varietal authentication, demonstrating high classification accuracy and strong model fit.

PLS-DA models, when based on metabolic markers such as SHs, TAGs, and UF— which are influenced by both genetic factors and pedoclimatic conditions—enable the identification of characteristic patterns related to the food property of interest, whether it be cultivar or origin. These models effectively minimize the impact of variables unrelated to the property being verified. Therefore, depending on the variable selected for supervising the pattern recognition analysis, PLS-DA models applied to the same markers and data can authenticate the varietal or geographical origin of the samples (Publications 2, 3, 7-9 and 11).

7.4.4 Model exploration

To understand the foundation of the models' predictions, it is crucial to evaluate the underlying chemical information. Therefore, to ensure the reliability of the developed models and confirm that their predictive accuracy is rooted in valid chemical and biological information, the models' regression coefficients were examined.

The regression coefficients of the PLS-DA models were studied to tentatively identify those variables that mainly contributed to the discrimination among classes. The jack-knife standard error of cross-validation (SEcv) was used to evaluate the significance of the regression coefficients, considering as significant those with values higher than their corresponding SEcv. Out of the significant variables, only the ones with the highest absolute values were considered, and the corresponding compounds or compound families were tentatively identified, based on their mass spectra and elution order (Publications 3, 7 and 11) or on the band assignation (Publication 9). Additionally,

models' regression coefficients were also employed to select the relevant isotopic variables in the PoC study (Publication 4) that were further determined in the large-sample set study (Publication 5).

The analysis of the regression coefficients in the VOO SH fingerprinting model (Publication 3) revealed that a significant number of variables, spread across the entire chromatogram and all acquired ions, were relevant for distinguishing between VOO cultivars. Notably, many of the highest regression coefficients corresponded to minor SH compounds or poorly resolved chromatographic peaks (**Figure 19**). Consequently, their identification and quantification using traditional targeted approaches would have been challenging.

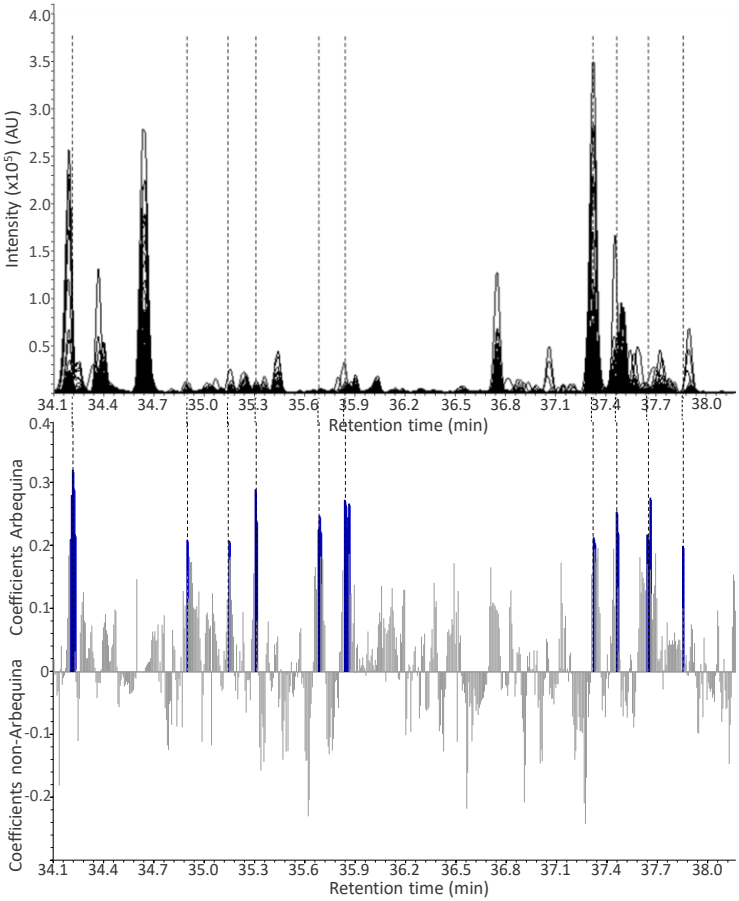


Figure 19. Section of the ion *m/z* 93 chromatogram of the VOO sesquiterpene fingerprinting data plotted against the regression coefficients of the PLS-DA Arbequina model (relevant variables for the Arbequina class highlighted in blue) (Publication 3).

Similar results were obtained examining the regression coefficients of the pine nut SH models (Publication 11). In this case, both mono- and sesquiterpene compounds

contributed to the discrimination. These findings underscore the suitability of the fingerprinting approach, which facilitates the analysis of minor species and coeluting peaks, offering a unique advantage over traditional targeted methods.

Regarding the hazelnut varietal and geographical authentication by UF models (Publication 7), the regression coefficients from the PLS-DA models developed using both SIM fingerprinting and untargeted profiling data were compared to identify differences in the key discriminant compounds. Results revealed that, while some compounds were important for discrimination in both approaches, each method also identified unique relevant variables. This indicates that the information provided by each approach varies due to their distinct operational modes. The greater sensitivity of the SIM fingerprinting method enabled the detection of minor compounds that significantly contributed to sample classification, and that could not be detected using the untargeted profiling's full scan acquisition. Moreover, the comparison of regression coefficients revealed that in that study, restricting data acquisition to specific ions in the SIM fingerprinting approach did not hinder the detection of discriminant compounds. Indeed, the compounds identified as relevant only by the untargeted profiling models were also detected by the SIM method but were not considered relevant for the classification.

On the other hand, the untargeted profiling approach provided the pure mass spectra facilitating the straightforward identification of the most relevant compounds as indicated by the regression coefficients of the models—an advantage not provided by the SIM fingerprinting method. This makes untargeted profiling particularly adequate for analysing samples with unknown compositions, where an easy identification of discriminant compounds is essential. However, when there is prior knowledge of the compound families present in the samples, as it happened in our UF study (Publication 7), selecting specific ions makes the SIM fingerprinting method more suitable, as it delivers richer information and enables a more efficient discrimination.

For the models developed on NIR and MIR spectroscopic data of hazelnuts, the analysis of the regression coefficients revealed that each model relied on different spectral information (Publication 9). Despite the challenges in associating spectral bands with specific components—due to the low specificity of the techniques, complex spectra,

and overlapping bands—it was possible to correlate the key variables with the main compositional elements of the samples. The findings indicated that lipids and proteins were the primary drivers of hazelnut cultivar and origin differentiation. This agreed with the previous successfulness of both TAG and UF markers, that had already suggested these spectroscopic techniques as suitable candidates.

In all studies where PLS-DA models were applied to the same authenticity markers for verifying the geographical or botanical origin of food (Publications 3, 7, 9, 11) it was observed that the relevant variables for sample discrimination varied depending on whether provenance or cultivar was selected as the supervising variable. This finding confirms that genetic and environmental factors have distinct impacts on the metabolic fingerprints of the analysed markers and that PLS-DA is able to extract and use this information for the discrimination of different geographical or varietal groups.

Finally, and as previously mentioned, the examination of the model's regression coefficients can be used to select variables with the highest discriminant power to optimize the classification models, as demonstrated in Publication 4. In this PoC study multiple isotopic markers ($\delta^{13}C$, $\delta^{2}H$, $\delta^{18}O$, $\delta^{15}N$, $\delta^{34}S$, $^{86}Sr/^{87}Sr$, and the $\delta^{13}C$ and $\delta^{2}H$ of the FAMEs) were evaluated to verify the provenance of hazelnuts. The regression coefficients revealed that, after excluding isotopes potentially influenced by agronomic practices, $\delta^{18}O$, and $\delta^{13}C$ and $\delta^{2}H$ of the FAMEs, significantly contributed to the geographical discrimination of hazelnuts (**Figure 20**). These findings were the basis for developing more robust hazelnut geographical authentication models using large-scale datasets (Publication 5).

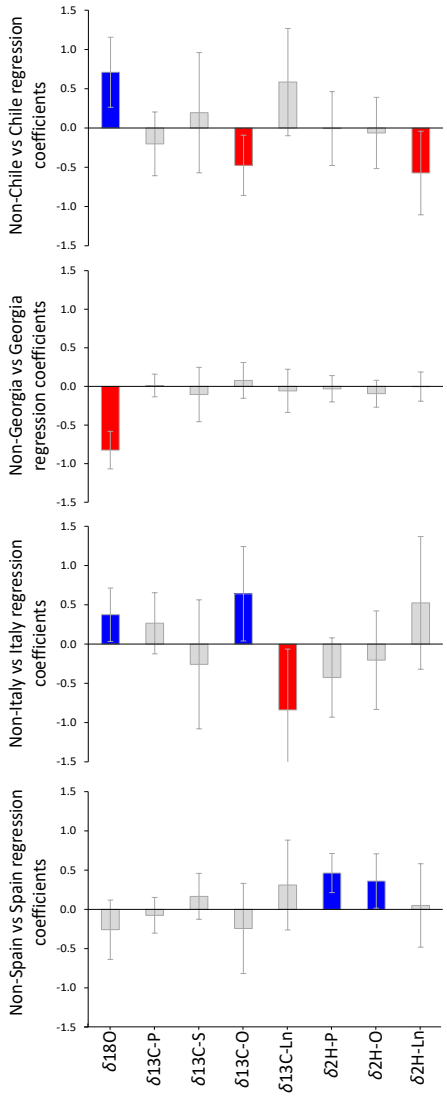


Figure 20. Regression coefficients of the hazelnuts PLS-DA origin model developed with the variables $\delta^{18}O$, and $\delta^{2}H$, $\delta^{13}C$ of the main fatty acid methyl esters (FAMEs) (P: palmitic, S: stearic, O: oleic, Ln: linoleic). Significant coefficients are highlighted in blue (specific country) or red (other countries) (Publication 4).

7.5 Applications and usefulness of the developed methodologies for enhancing food geographical and varietal authenticity

7.5.1 Global overview and current application

Targeted isotopic methods have demonstrated satisfactory discrimination power with just a few isotopic markers for the geographical authentication of VOO and hazelnuts, provided the regions in question exhibit significant pedoclimatic differences (Publications 2 and 5), or clear geological differences (Publication 10). However, they failed in discriminating between climatologically similar regions, such as ITA and ESP,

and in discriminating wide areas with highly variable geology, such as RUS and CHN pine nuts (Publication 10). Therefore, for pine nuts, the analysis of Sr ratios was unsuccessful in verifying their provenance. Isotopic methods consistently exhibited lower prediction accuracy compared to untargeted metabolomic methods (Publications 2, 8 and 11).

Despite these limitations, isotopic methods offer straightforward transferability, versatility, and broad applicability, facilitating their implementation across different laboratories. Consequently, they could serve as supporting screening tools to official control methods to verify the geographical origin of VOO and hazelnuts if the regions under scrutiny have distinct geographical and climatic features.

The methodologies developed for the authentication of VOO, hazelnuts, and pine nuts based on metabolic markers have proven highly successful, particularly the fingerprinting techniques. These methods have significantly outperformed isotopic targeted approaches for all analysed agrifood products (Publications 2, 5, 8, 9, 10 and 11), even in the verification of the geographical origin where stable isotopes are widely accepted markers. In fact, metabolic markers enable simultaneous verification of both varietal and geographical origin, offering exceptional classification accuracy, sensitivity, specificity, and robust model fitting, with minimal interference from other factors than the property being verified.

For VOO and pine nuts, SH fingerprinting has emerged as the most effective tool for varietal and geographical authentication (Publications 2, 3 and 11). This method is both fast and straightforward, allowing direct analysis of samples without the need for prior sample treatment, using the automatable and widely available GC-MS technique. Models build using SH fingerprinting data have demonstrated the highest classification accuracy in external validation across all studies conducted on pine nuts and VOO, effectively authenticating both provenance and cultivar.

For hazelnuts, where the minimal SH concentration renders these compounds unsuitable as authenticity markers, UF and NIR spectroscopy fingerprinting were the most effective methods for varietal and geographical authentication, achieving similar rates of correct classification in external validation (> 94% for cultivar and > 96% for

origin) (Publications 8 and 9). However, UF fingerprinting involves a more laborious analytical procedure, making it better suited as a confirmatory tool. As demonstrated in Publication 8, it was more time-efficient to verify challenging boundary samples when used in combination with faster screening methods such as TAG fingerprinting.

The NIR fingerprinting presented lower analytical sensitivity and selectivity than the UF fingerprint obtained through GC-MS, limiting the chemical interpretability of the models and complicating the identification and differentiation of sample components, thereby making its use as a confirmatory method significantly more challenging. Nonetheless, it remains a fast, simple, versatile and cost-effective method that allows for the direct analysis of solid samples in a short time. This makes it the most suitable screening tool to verify hazelnuts varietal and geographical origin in routine analysis. Combining the rapid NIR fingerprinting screening method with UF fingerprinting as a confirmatory tool could be the most effective and efficient approach for verifying hazelnut varietal and geographical origin and should be considered for future evaluation.

However, the primary limitation of these untargeted fingerprinting approaches is the absence of standardized validation protocols, which complicates their transferability and prevents their use as official reference methods. Until this issue is resolved, these methods can serve as a guidance for inspections and conformity checks, supporting the work of official control bodies and enhancing risk analysis for VOO, hazelnuts and pine nuts. In this framework, an initial screening using untargeted fingerprinting could efficiently process a large number of samples, offering an early indication of whether they conform to the declared varietal and geographical origins or not. Suspicious or borderline samples could then be subjected to further analysis using official control methods, if available, or could prompt traceability procedures and document verification. This approach would greatly improve the effectiveness of existing official controls. Additionally, these methods could also be implemented as self-monitoring tools within the private sector.

Finally, regardless of whether metabolic or isotopic methods are used, the discriminant models developed for varietal and geographical authentication should be regularly

updated by incorporating samples from new harvest years, cultivars/species, producers, and regions of origin.

7.5.2 Future prospects

The performance of isotopic methods as tools for the geographical authentication of VOO, hazelnuts, and pine nuts could be significantly enhanced by combining them with other authenticity markers, such as mineral elemental profiles, FAs, or additional isotopes [84,85,90,97,100]. In fact, the results of Publication 10 already suggest this for Rb and Sr measurement in pine nuts. Furthermore, the development and use of accessible, large, and comprehensive databanks—regularly updated with samples from various regions, harvest seasons, cultivars, and producers—would serve as an invaluable resource for comparison and model building. Such databanks would greatly improve the effectiveness of isotopic methods in the geographical authentication of agrifood products [84].

Given the outstanding results achieved with untargeted fingerprinting methods and their great potential for authenticating the geographical and varietal origin of VOO, hazelnuts, and pine nuts, there is a pressing need to improve their transferability to advance towards their applicability in routine official control. To accomplish this purpose, it is essential to establish standardized procedures with clear and precise guidelines for validating untargeted fingerprinting methods. These procedures should include:

- Verification of the repeatability and reproducibility of analytical data, in this case chromatographic or spectroscopic fingerprint, obtained through different instruments [265,266].
- Validation of authentication models through external validation by predicting new samples, not included in the model development. This process should account for the natural variability of the food being analysed and assess model stability by testing samples analysed by different operators and instruments [200].
- Protocols for interlaboratory studies to evaluate the method's robustness and ensure consistent results across various laboratories.

Finally, to maintain the validity of the authentication method over time, it is crucial to regularly monitor its performance. This should involve analysing blind samples, conducting proficiency tests, and using reference samples [81]. Additionally, incorporating new samples that include additional authentication traits or categories will be useful to track changes in the agrifood product and help ensure continued accuracy and relevance.

CHAPTER 8.
CONCLUSIONS

Conclusions regarding the enhancement of olive oil geographical and varietal authenticity

1. Comparison of isotopic analysis and SH fingerprinting for authenticating VOO provenance

The **systematic comparison** of untargeted **SH fingerprinting and targeted isotopic profiling** methods for verifying the origin of Italian VOO, revealed that **SH fingerprinting outperforms the isotopic method in sensitivity, specificity and accuracy**, thus establishing it as a **more effective tool for the geographical authentication of VOO**.

2. Development of an untargeted metabolic method to verify VOO cultivar

Varietal authentication models developed using **SH fingerprinting of VOO** obtained through HS-SPME-GC–MS and combined with PLS-DA **effectively verified olive cultivars** beyond geographical, agronomic, and technological factors, confirming **SH fingerprinting as a reliable tool for accurate VOO provenance and varietal authentication** in a single analysis.

Conclusions regarding the enhancement of hazelnut geographical and varietal authenticity

3. Development of a targeted isotopic method to verify hazelnut geographical origin

3a. The **most promising isotopic markers** for authenticating hazelnut geographical origins were **bulk $\delta^{18}O$, and $\delta^{2}H$ and $\delta^{13}C$ of the main FAMEs**, as they were minimally affected by external factors like fertilisation and effectively discriminate between distinct regions, with **$^{87}Sr/^{86}Sr$** serving as a valuable **confirmatory tool**.

3b. The development of a **multi-isotopic method**, integrating **the most discriminant bulk and compound-specific isotopic ratios with chemometric analysis**, effectively authenticated hazelnuts from geographical areas with distinct climatic conditions but not from those with similar climates, demonstrating that **this method is best suited to distinguish regions with clear pedoclimatic differences**.

4.1 Development of an untargeted metabolic method to simultaneously verify hazelnut cultivar and origin

4.1a The development of a method based on hazelnut UF and PLS-DA confirmed that **UF is a promising tool** for both **geographical and varietal authentication**, with **fingerprinting outperforming untargeted profiling** in data treatment.

4.1b A **fingerprinting method** combining **HT-GC-MS** with PLS-DA was developed to analyse the **TAG fraction** of lipid samples, highlighting its **superiority in detecting subtle differences in TAG** composition compared with other methods.

4.1c Authentication models based on **UF fingerprinting**, though **labour-intensive**, achieved **high sensitivity, specificity, and accuracy** in distinguishing hazelnuts by **origin and cultivar**, while **TAG fingerprinting**, though **less efficient**, provided a **faster and simpler alternative** suitable for large-scale screening.

4.1d Integrating TAG fingerprinting for initial screening with **UF fingerprinting** for verifying uncertain samples significantly **enhanced analytical efficiency** and maintained **high classification accuracy**, proving to be a **powerful and effective** tool for authenticating both the **varietal and geographical origin** of hazelnuts.

4.1e NIR emerged as the most suitable among the spectroscopic techniques tested, offering rapid and straightforward analysis with high accuracy, specificity and sensitivity, making it an ideal **screening tool** for authenticating hazelnuts **provenance and varietal origin in routine analysis**.

4.2 Comparison of metabolic and isotopic methodologies for authenticating hazelnut geographical and varietal origin

4.2a Despite the straightforward transferability of the **isotopic method**, its **discriminative capacity is limited** by similarities in climate and geography among regions, leading to **less** reliable and **effective** authentication **than metabolic untargeted methods**.

4.2b Metabolic fingerprinting methods, particularly **UF and NIR**, demonstrated **superior performance**, delivering the highest classification accuracy **for distinguishing both hazelnut cultivar and provenance**.

Conclusions regarding the enhancement of pine nut geographical and varietal authenticity

5. Development of a targeted isotopic method to verify pine nut geographical origin

5a The analysis of **$^{87}Sr/^{86}Sr$ isotopic ratio** in pine nuts from various origins, species, producers, and harvest years **failed to differentiate Spanish *P. pinea* from Asian pine nuts**, underscoring that a **single isotopic marker is insufficient** to address the complex challenge of **verifying the geographical origin** of an agrifood product.

5b The exploratory analysis of **Sr and Rb content** effectively differentiated between Asian and Mediterranean pine nuts, with partial distinction between the two Spanish regions studied, indicating that these markers **are promising tools for pine nut geographical authentication**.

6. Development of an untargeted metabolic method to verify pine nut species and origin

A method using **volatile and semi-volatile terpene fingerprinting** obtained by HS-SPME-GC–MS was successfully developed, enabling robust PLS-DA models that effectively **distinguished Spanish *P. pinea* from Asian pine nuts,** and differentiated between ***P. pinea* from two Spanish regions**, proving to be a **direct, rapid, and efficient tool for pine nut geographical and botanical authentication**.

As a general conclusion, the integration of suitable authenticity markers with cutting-edge analytical approaches, high-throughput instrumental techniques and chemometric tools has led to the **development of robust, reliable, and efficient authentication methods** that will contribute to enhancing the **geographical and varietal authenticity of VOO, hazelnuts, and pine nuts**.

CHAPTER 9.
REFERENCES

List of scientific contributions derived from the present thesis

Publication 1: Quintanilla-Casas, B., Torres-Cobos, B., Bro, R., Guardiola, F., Vichi, S., Tres, A. The volatile metabolome - GC-MS approaches in the context of food fraud. *Accepted for publication in Current Opinion in Food Science 2024.*

Publication 2: Torres-Cobos, B., Bontempo, L., Roncone, A., Quintanilla-Casas, B., Servilli, M., Guardiola, F., Vichi, S., Tres, A. Ground-breaking comparison of target stable isotope ratios vs. emerging sesquiterpene fingerprinting for authenticating VOO origin. *Under review in Food Chemistry 2024*.

Publication 3: Torres-Cobos, B., Quintanilla-Casas, B., Romero, A., Ninot, A., Alonso-Salces, R. M., Toschi, T. G., Bendini, A., Guardiola, F., Tres, A., Vichi, S. (2021). Varietal authentication of virgin olive oil: Proving the efficiency of sesquiterpene fingerprinting for Mediterranean Arbequina oils. *Food Control*, *128*, 108200. https://doi.org/10.1016/j.foodcont.2021.108200.

Publication 4: Torres-Cobos, B., Rosell, M., Soler, A., Rovira, M., Romero, A., Guardiola, F., Vichi, S., Tres, A. (2024). Investigating isotopic markers for hazelnut geographical authentication: Promising variables and potential applications. *Food Chemistry*, *449*, 139083. https://doi.org/10.1016/j.foodchem.2024.139083.

Publication 5: Torres-Cobos, B., Rosell, M., Soler, A., Rovira, M., Romero, A., Guardiola, F., Vichi, S., Tres, A. (2024). A multi-isotopic chemometric approach for tracing hazelnuts origin. *Foods*, *13*(21), 3399. https://doi.org/10.3390/foods13213399.

Publication 6: Torres-Cobos, B., Quintanilla-Casas, B., Vicario, G., Guardiola, F., Tres, A., Vichi, S. (2023). Revealing adulterated olive oils by triacylglycerol screening methods: Beyond the official method. *Food Chemistry*, *409*, 135256. https://doi.org/10.1016/j.foodchem.2022.135256.

Publication 7: Torres-Cobos, B., Quintanilla-Casas, B., Rovira, M., Romero, A., Guardiola, F., Vichi, S., Tres, A. (2024). Prospective exploration of hazelnut's unsaponifiable fraction for geographical and varietal authentication: A comparative study of advanced fingerprinting and untargeted profiling techniques. *Food Chemistry*, *441*, 138294. https://doi.org/10.1016/j.foodchem.2023.138294.

Publication 8: Torres-Cobos, B., Nicotra, S. B., Rovira, M., Romero, A., Guardiola, F., Tres, A., Vichi, S. (2025). Meeting the challenge of varietal and geographical authentication of hazelnuts through lipid metabolite fingerprinting. *Food Chemistry, 463*(2), 141203. https://doi.org/10.1016/j.foodchem.2024.141203.

Publication 9: Torres-Cobos, B., Tres, A., Vichi, S., Guardiola, F., Rovira, M., Romero, A., Baeten, V., Fernández-Pierna, J. A. (2025). Comparative analysis of spectroscopic methods for rapid authentication of hazelnut cultivar and origin. *Spectrochimica Acta, Part A: Molecular and Biomolecular Spectroscopy*, *326*, 125367. https://doi.org/10.1016/j.saa.2024.125367

Publication 10: Torres-Cobos, B., Rosell, M., Soler, A., Nicotra, S. B., Aletà, N., Teixidó, A., Rovira, M., Romero, A., Guardiola, F., Tres, A., Vichi, S. (2024). Evaluating Sr isotope ratios, and Sr and Rb elemental analysis for pine nuts geographical authentication. *Manuscript under preparation 2024*.

Publication 11: Torres-Cobos, B., Nicotra, S. B., Asensio-Manzano, C., Aletà, N., Teixidó, A., Rovira, M., Romero, A., Guardiola, F., Vichi, S., Tres, A. Mono- and sesquiterpenoid fingerprinting: a powerful and streamlined solution for pine nut authentication. *Under review in Food Chemistry 2024*.

List of references excluding those cited in the publications

1. Armstrong, D. J. (2009). Food chemistry and US food regulations. *Journal of agricultural and food chemistry*, *57*(18), 8180–8186. https://doi.org/10.1021/jf900014h.
2. Valadés, P. S. P., de la Paz, M. P., de Andrés Copa, P., Ribota, O. G., & Borda, I. A. (2003). Toxic oil syndrome: Survival in the whole cohort between 1981 and 1995. *Journal of Clinical Epidemiology*, *56*(7), 701-708. https://doi.org/10.1016/S0895-4356(03)00119-7.
3. Lees, M., Morin, J. F. (Eds.). (2018). *FoodIntegrity handbook: A guide to food authenticity issues and analytical solutions*. Eurofins Analytics France. ISBN: 978-2-9566303-1.
4. Robson, K., Dean, M., Brooks, S., Haughey, S., Elliott, C. (2020). A 20-year analysis of reported food fraud in the global beef supply chain. *Food Control*, *116*, 107310. https://doi.org/10.1016/j.foodcont.2020.107310..
5. Spink, J., Moyer, D. C. (2011). Defining the public health threat of food fraud. *Journal of food science*, *76*(9), R157–R163. https://doi.org/10.1111/j.1750-3841.2011.02417.x
6. Brooks, C., Parr, L., Smith, J. M., Buchanan, D., Snioch, D., Hebishy, E. (2021). A review of food fraud and food authenticity across the food supply chain, with an examination of the impact of the COVID-19 pandemic and Brexit on food industry. *Food Control*, *130*, 108171. https://doi.org/10.1016/j.foodcont.2021.108171.
7. Visciano, P., Schirone, M. (2021). Food frauds: Global incidents and misleading situations. *Trends in Food Science & Technology*, *114*, 424–442. https://doi.org/10.1016/j.tifs.2021.06.010.
8. Global Food Safety Initiative. (2018). Tackling food fraud through food safety management systems. Available online: https://mygfsi.com/wp-content/uploads/2019/09/Food-Fraud-GFSI-Technical-Document.pdf (accessed on 29 May 2024).
9. Spink, J. W. (2023). The GFSI food fraud prevention compliance development & requirements: A ten-year review. *Trends in Food Science & Technology*, 138, 766–773. https://doi.org/10.1016/j.tifs.2023.07.014.
10. Spink, J. W. (2019). The current state of food fraud prevention: overview and requirements to address 'How to Start?' and 'How Much is Enough?'. Current Opinion in Food Science, 27, 130–138. https://doi.org/10.1016/j.cofs.2019.06.001.
11. Codex Alimentarius Commission. Joint FAO/WHO Food Standards Programme. (2018). *Discussion paper on food integrity and food authenticity (CX/FICS 18/24/7)*. Available

online: https://www.fao.org/fao-who-codexalimentarius/sh-proxy/en/?lnk=1&url=https%253A%252F%252Fworkspace.fao.org%252Fsites%252Fcodex%252FMeetings%252FCX-733-24%252FWorking%2BDocuments%252Ffc24_07e.pdf (accessed on 4 June 2024).

12. Bitzios, M., Lisa, J. A. C. K., Krzyzaniak, S. A., Mark, X. U. (2017). Country-of-origin labelling, food traceability drivers and food fraud: Lessons from consumers' preferences and perceptions. *European Journal of Risk Regulation*, *8*(3), 541-558. https://doi.org/10.1017/err.2017.27.
13. Wu, W., Zhang, A., van Klinken, R. D., Schrobback, P., Muller, J. M. (2021). Consumer trust in food and the food system: a critical review. *Foods*, *10*(10), 2490. https://doi.org/10.3390/foods10102490.
14. Fernández-Ferrín, P., Bande, B., Martín-Consuegra, D., Díaz, E., Kastenholz, E. (2020). Sub-national consumer ethnocentrism and the importance of the origin of food products: an exploratory analysis. *British Food Journal*, *122*(3), 995-1010. https://doi.org/10.1108/BFJ-09-2019-0746.
15. Cabrera, E. R., Arriaza, M., Rodríguez-Entrena, M. (2015). Is the extra virgin olive oil market facing a process of differentiation? A hedonic approach to disentangle the effect of quality attributes. *Grasas y Aceites*, *66*(4), e105–e105. https://doi.org/10.3989/gya.0253151.
16. Cicerale, S., Liem, G., Keast, R. S (2016). Consumer perception, attitudes, liking and preferences for olive oil. In D. Boskou, M. Clodoveo (Eds.). Consumer perception, attitudes, liking and preferences for olive oil. *Products from olive tree*, 157–171. https://doi.org/10.5772/64554.
17. Kumar, S., Murphy, M., Talwar, S., Kaur, P., Dhir, A. (2021). What drives brand love and purchase intentions toward the local food distribution system? A study of social media-based REKO (fair consumption) groups. *Journal of Retailing and Consumer Services*, *60*, 102444. https://doi.org/10.1016/j.jretconser.2021.102444.
18. Thøgersen, J., Pedersen, S., Aschemann-Witzel, J. (2019). The impact of organic certification and country of origin on consumer food choice in developed and emerging economies. *Food Quality and Preference*, *72*, 10-30. https://doi.org/10.1016/j.foodqual.2018.09.003.
19. Charlebois, S., Schwab, A., Henn, R., Huck, C. W. (2016). Food fraud: An exploratory study for measuring consumer perception towards mislabeled food products and influence on self-authentication intentions. *Trends in food science & technology*, *50*, 211-218. https://doi.org/10.1016/j.tifs.2016.02.003.

20. Spink, J., Moyer, D. C., Speier-Pero, C. (2016). Introducing the food fraud initial screening model (FFIS). *Food control*, *69*, 306-314. https://doi.org/10.1016/j.foodcont.2016.03.016.
21. Gizaw, Z. (2019). Public health risks related to food safety issues in the food market: a systematic literature review. *Environmental health and preventive medicine*, *24*, 1-21. https://doi.org/10.1186/s12199-019-0825-5.
22. Danezis, G. P., Tsagkaris, A. S., Camin, F., Brusic, V., Georgiou, C. A. (2016). Food authentication: Techniques, trends & emerging approaches. *TrAC Trends in Analytical Chemistry*, *85*, 123-132. https://doi.org/10.1016/j.trac.2016.02.026.
23. Katerinopoulou, K., Kontogeorgos, A., Salmas, C. E., Patakas, A., Ladavos, A. (2020). Geographical origin authentication of agri-food products: a review. *Foods*, *9*, 489. https://doi.org/10.3390/foods9040489.
24. Official Journal of the European Communities. (2002). European Commission *Regulation No 178/2002 on laying down the general principles and requirements of food law, establishing the European Food Safety Authority and laying down procedures in matters of food safety*, pp. 1–24.
25. European Commission. General Food Law. Available online: https://ec.europa.eu/food/safety/general_food_law_en (accessed on 4 June 2024).
26. European Commission. Rapid Alert System for Food and Feed (RASFF). Available online: https://food.ec.europa.eu/safety/rasff_en (accessed on 4 June 2024).
27. Official Journal of the European Union. (2019). European Commission implementing *Regulation No 2019/1715 on laying down rules for the functioning of the information management system for official controls and its system components (the IMSOC Regulation)*, pp. 37–96.
28. Official Journal of the European Union. (2017). European Commission *Regulation No 2017/625 on official controls and other official activities performed to ensure the application of food and feed law, rules on animal health and welfare, plant health and plant protection products*, pp. 1–142.
29. European Commission. The Alert and Cooperation Network. Available online: https://food.ec.europa.eu/safety/acn_en (accessed on 4 June 2024).
30. European Commission. Standard operating procedures of the Alert and Cooperation Network (ACN), governing RASFF, AAC and FFN (version 3.1). https://food.ec.europa.eu/document/download/47cd12e2-6647-457f-a656-3ef665f50af2_en?filename=rasff_reg-guid_sops_en.pdf (accessed on 4 June 2024).

31. European Union. (2023). *2022 Annual Report. Alert and Cooperation Network*, Luxembourg. Available online: https://food.ec.europa.eu/document/download/499ffcf1-6c99-43ec-8905-5ff3e812eeb2_en?filename=acn_annual-report_2022.pdf (accessed on 4 June 2024).
32. European Union. (2022). *2021 Annual Report. Alert and Cooperation Network*, Luxembourg. Available online: https://food.ec.europa.eu/document/download/e8b14245-1f30-4f2b-bf5c-5e70e525e753_en?filename=acn_annual-report_2021-final.pdf (accessed on 4 June 2024).
33. European Union. (2021). *2020 Annual Report. The EU Agri-Food Fraud Network and the Administrative Assistance and Cooperation System*, Luxembourg. Available online: https://food.ec.europa.eu/document/download/5135ace4-2a9d-4bf7-afad-574621b43b1c_en?filename=ff_ffn_annual-report_2020_1.pdf (accessed on 4 June 2024).
34. European Union. (2020). *2019 Annual Report. The EU Agri-Food Fraud Network and the Administrative Assistance and Cooperation System*, Luxembourg. Available online: https://food.ec.europa.eu/document/download/a586adf7-74d8-4768-95c5-a653a7804af5_en?filename=ff_ffn_annual-report_2019.pdf (accessed on 4 June 2024).
35. European Union. (2019). *2018 Annual Report. The EU Food Fraud Network and the System for Administrative Assistance - Food Fraud*, Luxembourg. Available online: https://food.ec.europa.eu/document/download/4c30d7e4-cabc-4124-abca-ef7e642c78f5_en?filename=ff_ffn_annual-report_2018.pdf (accessed on 4 June 2024).
36. International Olive Council (IOC), (2023) World olive oil and table olive figures. https://www.internationaloliveoil.org/what-we-do/economic-affairs-promotion-unit/ (Accessed 5 July 2024).
37. Bimbo, F., Roselli, L., Carlucci, C., de Gennaro, B.C. (2020). Consumer Misuse of Country-of-Origin Label: Insights from the Italian Extra-Virgin Olive Oil Market. *Nutrients, 12,* 2150. https://doi.org/10.3390/nu12072150.
38. Dias, L. G., Fernandes, A., Veloso, A. C., Machado, A. A., Pereira, J. A., Peres, A. M. (2014). Single-cultivar extra virgin olive oil classification using a potentiometric electronic tongue. *Food Chemistry*, *160*, 321–329. https://doi.org/10.1016/j.foodchem.2014.03.072.

39. Montealegre, C., Marina Alegre, M. L., García-Ruiz, C. (2010). Traceability markers to the botanical origin in olive oils. *Journal of Agricultural and Food Chemistry*, *58*(1), 28–38. https://doi.org/10.1021/jf902619z.
40. Japón-Luján, R., Ruiz-Jiménez, J., Luque de Castro, M. D. (2006). Discrimination and classification of olive tree varieties and cultivation zones by biophenol contents. *Journal of agricultural and food chemistry*, *54*(26), 9706-9712. https://doi.org/10.1021/jf062546w.
41. Sanz-Cortés, F., Parfitt, D. E., Romero, C., Struss, D., Llácer, G., Badenes, M. L. (2003). Intraspecific olive diversity assessed with AFLP. *Plant breeding, 122*(2), 173-177. https://doi.org/10.1046/j.1439-0523.2003.00808.x.
42. Tous, J. (2017). The influence of growing region and cultivar on olives and olive oil characteristics and on their functional constituents. In A. Kiritsakis, & F. Shahidi (Eds.), *Olives and olive oil as functional food. Bioactivity, chemistry and processing* (pp. 45–80). New York, NY: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119135340.ch4.
43. Food and Agriculture Organization of the United Nations, FAOSTAT, food and agricultural data 2022. Available online: https://www.fao.org/faostat/en/#home (accessed on 22 June 2024).
44. Amaral, J. S., Casal, S., Citová, I., Santos, A., Seabra, R. M., Oliveira, B. P. (2006). Characterization of several hazelnut (Corylus avellana L.) cultivars based in chemical, fatty acid and sterol composition. *European Food Research and Technology*, *222*, 274-280. https://doi.org/10.1007/s00217-005-0068-0.
45. Król, K., Gantner, M. (2020). Morphological traits and chemical composition of hazelnut from different geographical origins: A review. *Agriculture*, *10*(9), 375. https://doi.org/10.3390/agriculture10090375.
46. International Nut and Dried Fruit Council (INC). (2023). *NUTS & DRIED FRUITS STATISTICAL YEARBOOK 2022/2023*. Available online: https://inc.nutfruit.org/wp-content/uploads/2023/05/Statistical-Yearbook-2022-2023.pdf (accessed on 22 June 2024).
47. Bozoğlu, M., Başer, U., Kilic Topuz, B., Alhas Eroglu, N. (2019). An overview of hazelnut markets and policy in Turkey. *Journal of Agriculture and Nature*, *22*(5), 733–743. https://doi.org/10.18016/ksutarimdoga.v22i45606.532645.
48. Baki, R. (2023). The evaluation of target markets for hazelnut exports with the classification approach of potential market alternatives. *British Food Journal*, *125*(10), 3540–3552. https://doi.org/10.1108/BFJ-02-2023-0100.

49. Yıldırım, Ç., Bozoğlu, M., Urago, G. G. (2024). Türkiye's Competitive Power in the World Hazelnut Market. *Applied Fruit Science*, 1–8. https://doi.org/10.1007/s10341-024-01053-4.
50. Maestri, E., Imperiale, D., Marmiroli, N. (2018). Nuts, nut products and other seeds. *In FoodIntegrity Handbook–A guide to food authenticity issues and analytical solutions* (pp. 127-135). Eurofins Analytics France. https://hdl.handle.net/11381/2860866.
51. Silvestri, C., Bacchetta, L., Bellincontro, A., Cristofori, V. (2021). Advances in cultivar choice, hazelnut orchard management, and nut storage to enhance product quality and safety: an overview. *Journal of the Science of Food and Agriculture*, *101*(1), 27–43. https://doi.org/10.1002/jsfa.10557.
52. Awan, H. U. M., Pettenella, D. (2017). Pine nuts: a review of recent sanitary conditions and market development. *Forests*, *8*, 367. https://doi.org/10.3390/f8100367.
53. Evaristo I., Batista D., Correia I., Correia P., Costa R. (2013). Chemical profiling of Portuguese *Pinus pinea* L. nuts and comparative analysis with *Pinus koraiensis* Sieb. & Zucc. commercial kernels. *Options Méditerranéennes. Séries A: Mediterranean Seminars*, *105*, 99–104. Available online: http://om.ciheam.org/om/pdf/a105/00006787.pdf (accessed on 11 April 2024).
54. Mutke, S. (2022). Producción y comercio de piñones comestibles en el mundo. *Montes*, *147*, 40–43. Available online: https://www.revistamontes.net/Buscador.aspx?id=15180 (accessed on 14 April 2024).
55. Moscetti, R., Berhe, D. H., Agrimi, M., Haff, R. P., Liang, P., Ferri, S., Monarca, D., Massantini, R. (2021). Pine nut species recognition using NIR spectroscopy and image analysis. *Journal of Food Engineering*, *292*, 110357. https://doi.org/10.1016/j.jfoodeng.2020.110357.
56. Destaillats, F., Cruz-Hernandez, C., Giuffrida, F., Dionisi, F., Mostin, M., Verstegen, G. (2011). Identification of the botanical origin of commercial pine nuts responsible for dysgeusia by gas-liquid chromatography analysis of fatty acid profile. *Journal of Toxicology*, *2011*(1), 316789. https://doi.org/10.1155/2011/316789.
57. Mutke, S., Pastor, A., Picardo, A. (2013). Toward a traceability of European pine nuts "from forest to fork". *Options Méditerranéennes. Séries A: Mediterranean Seminars*, *105*, 105–109. Available online : http://om.ciheam.org/om/pdf/a105/00006788.pdf (accessed on 14 April 2024).
58. Conte, L., Bendini, A., Valli, E., Lucci, P., Moret, S., Maquet, A., Lacoste, F., Brereton, P., García-González, D. L., Moreda, W., Toschi, T. G. (2020). Olive oil quality and

authenticity: A review of current EU legislation, standards, relevant methods of analyses, their drawbacks and recommendations for the future. *Trends in Food Science & Technology*, *105*, 483–493. https://doi.org/10.1016/j.tifs.2019.02.025.

59. Official Journal of the European Union. (2023). European Commission *Regulation No 2023/2429 as regards marketing standards for the fruit and vegetables sector, certain processed fruit and vegetable products and the bananas sector*, pp. 1–77.
60. Official Journal of the European Union. (2022). European Commission *Regulation No 2022/2104 as regards marketing standards for olive oil*, pp. 1–22.
61. Official Journal of the European Union. (2024). European Commission *Regulation No 2024/1143 on geographical indications for wine, spirit drinks and agricultural products, as well as traditional specialities guaranteed and optional quality terms for agricultural products*, pp. 1–76.
62. De la Guardia, M., Illueca, A. G. (2013). *Food protected designation of origin: methodologies and applications* (1st ed.). Elsevier, Amsterdam, Netherlands. ISBN: 978-0-444-59562-1.
63. European Commission, eAmbrosia – the EU geographical indication register. https://ec.europa.eu/agriculture/eambrosia/geographical-indications-register/ (accessed on 7 August 2024).
64. UNECE (2010), *UNECE Standard DDP-04 concerning the marketing and commercial quality control of Hazelnut Kernels*. 2010 Editions. United Nations, New York and Geneva. Available online: https://unece.org/sites/default/files/2023-08/04_HazelnutKernels_e.pdf (accessed on 28 June 2024).
65. UNECE (2013), *UNECE Standard DDP-12 concerning the marketing and commercial quality control of Pine Nut Kernels*. 2013 Editions. United Nations, New York and Geneva. Available online: https://unece.org/sites/default/files/2023-08/12_PineNutKernels_e.pdf (accessed on 28 June 2024).
66. Yan, J., Erasmus, S. W., Toro, M. A., Huang, H., van Ruth, S. M. (2020). Food fraud: Assessing fraud vulnerability in the extra virgin olive oil supply chain. *Food Control*, *111*, 107081. https://doi.org/10.1016/j.foodcont.2019.107081.
67. European Commission. Knowledge Centre for Food Fraud and Quality Monthly Food Fraud Summary Reports. Available online: https://knowledge4policy.ec.europa.eu/food-fraud-quality/monthly-food-fraud-summary-reports_en (accessed on 10 June 2024).
68. Casadei, E., Valli, E., Panni, F., Donarski, J., Farrús Gubern, J., Lucci, P., Conte, L., Lacoste, F., Maquet, A., Brereton, P., Bendini, A., Toschi, T. G. (2021). Emerging Trends in Olive

Oil Fraud and Possible Countermeasures. *Food Control*, *124*, 107902. https://doi.org/10.1016/j.foodcont.2021.107902.

69. Areté Research & Consulting in Economics. (2020). Study on the implementation of conformity checks in the olive oil sector throughout the European Union. http://doi.org/10.2762/274137.
70. Suman, M., Cavanna, D., Sammarco, G., Lambertini, F., Loffi, C. (2021). Fighting food frauds exploiting chromatography-mass spectrometry technologies: Scenario comparison between solutions in scientific literature and real approaches in place in industrial facilities. *TrAC Trends in Analytical Chemistry*, *142*, 116305. https://doi.org/10.1016/j.trac.2021.116305.
71. Campmajó, G., Navarro, G. J., Núñez, N., Puignou, L., Saurina, J., Núñez, O. (2019). Non-targeted HPLC-UV fingerprinting as chemical descriptors for the classification and authentication of nuts by multivariate chemometric methods. *Sensors*, *19*(6), 1388. https://doi.org/10.3390/s19061388.
72. Marvin, H. J., Bouzembrak, Y., Janssen, E. M., van der Fels-Klerx, H. V., van Asselt, E. D., Kleter, G. A. (2016). A holistic approach to food safety risks: Food fraud as an example. *Food research international*, *89*, 463-470. https://doi.org/10.1016/j.foodres.2016.08.028.
73. European Union. (2017). *2016 Annual Report. The EU Food Fraud Network and the System for Administrative Assistance - Food Fraud*, Luxembourg. Available online: https://food.ec.europa.eu/document/download/a3bc6102-1dbf-4a34-b1c4-25d3ac0af997_en?filename=ff_ffn_annual-report_2016.pdf.
74. European Union. (2021). *2020 Annual Report. RASFF The Rapid Alert System for Food and Feed*, Luxembourg. Available online: https://food.ec.europa.eu/document/download/4b178b4f-c40c-4405-9c01-db558aa1392a_en?filename=rasff_pub_annual-report_2020.pdf (accessed on 4 June 2024).
75. International Nut and Dried Fruit Council (INC). (2020). *2020 IMPORT BORDER REJECTIONS EDIBLE NUTS & DRIED FRUITS*. Available online: https://inc.nutfruit.org/wp-content/uploads/2022/09/2020_EU-USA-JAP-AUS_Border_Rejections_INC-1.pdf (accessed on 4 June 2024).
76. Tähkäpää, S., Maijala, R., Korkeala, H., Nevas, M. (2015). Patterns of food frauds and adulterations reported in the EU rapid alert system for food and feed and in Finland. *Food Control*, *47*, 175–184. https://doi.org/10.1016/j.foodcont.2014.07.007.

77. Ciarmiello, L. F., Mazzeo, M.F., Minasi, P., Peluso, A., De Luca, A., Piccirillo, P., Siciliano, R. A., Carbone, V. (2014). Analysis of different european hazelnut (Corylus avellana L.) cultivars: authentication, phenotypic features, and phenolic profiles. *Journal of Agricultural and Food Chemistry*, *62*, 6236–6246. https://doi.org/10.1021/jf5018324.
78. Fardin-Kia, A. R., Handy, S. M., Rader, J. I. (2012). Characterization of pine nuts in the U.S. market, including those associated with "pine mouth", by GC-FID. *Journal of Agricultural and Food Chemistry*, *60*(10), 2701–2711. https://doi.org/10.1021/jf205188m.
79. Loewe-Muñoz, V., Álvarez, A., Navarro-Cerrillo, R. M. (2018). Morphometric and chemical fruit variability of selected stone pine trees (*Pinus pinea* L.) grown in non-native environments. *Plant Biosystems - An International Journal Dealing with All Aspects of Plant Biology*, *152*(3), 547–555. https://doi.org/10.1080/11263504.2018.1435587.
80. Mikkelsen, A. Æ., Jessen, F., Ballin, N. Z. (2014). Species determination of pine nuts in commercial samples causing pine nut syndrome. *Food Control*, *40*, 19–25. https://doi.org/10.1016/j.foodcont.2013.11.030.
81. Bayen, S., Elliott, C., Arlorio, M., Ballin, N. Z., Birse, N., Brockmeyer, J., Chahal, S., Corradini, M. G., Hanner, R., Hann, S., Laursen, K. H., Mihailova, A., Steininger-Mairinger, T., Suman, M., Tian, L., van Ruth, S., Xia, J. (2024). Towards a harmonized approach for food authenticity marker validation and accreditation. *Trends in Food Science & Technology*, *149*, 104550. https://doi.org/10.1016/j.tifs.2024.104550.
82. Creydt, M., Fischer, M. (2018). Omics approaches for food authentication. *Electrophoresis*, *39*(13), 1569–1581. https://doi.org/10.1002/elps.201800004.
83. Bertoldi D., Cossignani L., Blasi F., Perini M., Barbero A., Pianezze S., Montesano D. (2019). Characterization and geographical traceability of Italian goji berries. *Food Chemistry*, *275*, 585–593. https://doi.org/10.1016/j.foodchem.2018.09.098.
84. Camin, F., Boner, M., Bontempo, L., Fauhl-Hassek, C., Kelly, S. D., Riedl, J., Rossmann, A. (2017). Stable isotope techniques for verifying the declared geographical origin of food in legal cases. *Trends in Food Science & Technology*, *61*, 176–187. https://doi.org/10.1016/j.tifs.2016.12.007.
85. De Rijke, E., Schoorl, J. C., Cerli, C., Vonhof, H. B., Verdegaal, S.J.A., Vivó-Truyols, G., Lopatka, M., Dekter R., Bakker, D., Sjerps, M. J., Ebskamp M., De Koster C. G. (2016). The use of $\delta^{2}H$ and $\delta^{18}O$ isotopic analyses combined with chemometrics as a

traceability tool for the geographical origin of bell peppers. *Food Chemistry*, *204*, 122–128. https://doi.org/10.1016/j.foodchem.2016.01.134.

86. Kelly, S., Heaton, K., Hoogewerff, J. (2005). Tracing the geographical origin of food: the application of multi-element and multi-isotope analysis. *Trends in Food Science & Technology*, *16*, 555–567. https://doi.org/10.1016/j.tifs.2005.08.008.
87. Podio, N. S., Baroni, M. V., Badini, R. G., Inga, M., Ostera, H. A., Cagnoni, M., Gautier, E. A., García, P. P., Hoogewerff, J., Wunderlin, D.A. (2013). Elemental and isotopic fingerprint of Argentinean wheat. Matching soil, water, and crop composition to differentiate provenance. *Journal of Agricultural and Food Chemistry*, *61*(16), 3763–3773. https://doi.org/10.1021/jf305258r.
88. Laursen, K.H., Bontempo, L., Camin, F., Roßmann, A. (2016). Advances in isotopic analysis for food authenticity testing. In G. Downey (Ed), *Advances in Food Authenticity Testing* (pp. 227–252). Woodhead Publishing. https://doi.org/10.1016/B978-0-08-100220-9.00009-6.
89. Mazarakioti, E. C., Zotos, A., Thomatou, A. A., Kontogeorgos, A., Patakas, A., Ladavos, A. (2022). Inductively coupled plasma-mass spectrometry (ICP-MS), a useful tool in authenticity of agricultural products' and foods' origin. *Foods*, *11*(22), 3705. https://doi.org/10.3390/foods11223705.
90. Nasr, E. G., Epova, E. N., Sebilo, M., Larivière, D., Hammami, M., Souissi, R., Abderrazak, H., Donard, O. F. (2022). Olive oil traceability studies using inorganic and isotopic signatures: a review. *Molecules*, *27*(6), 2014. https://doi.org/10.3390/molecules27062014.
91. Zhao, Y., Zhang, B., Chen, G., Chen, A., Yang, S., Ye, Z. (2014). Recent developments in application of stable isotope analysis on agro-product authenticity and traceability. *Food chemistry*, *145*, 300–305. https://doi.org/10.1016/j.foodchem.2013.08.062.
92. Capo, R. C., Stewart, B. W., Chadwick, O. A. (1998). Strontium isotopes as tracers of ecosystem processes: theory and methods. *Geoderma*, *82*(1-3), 197–225. https://doi.org/10.1016/S0016-7061(97)00102-X.
93. Coelho, I., Castanheira, I., Bordado, J. M., Donard, O., Silva, J. A. L. (2017). Recent developments and trends in the application of strontium and its isotopes in biological related fields. *TrAC Trends in Analytical Chemistry*, *90*, 45–61. https://doi.org/10.1016/j.trac.2017.02.005.

94. Stewart, B. W., Capo, R. C., Chadwick, O. A. (1998). Quantitative strontium isotope models for weathering, pedogenesis and biogeochemical cycling. *Geoderma*, *82*(1-3), 173–195. https://doi.org/10.1016/S0016-7061(97)00101-8.
95. Elflein, L., Raezke, K. P. (2008). Improved detection of honey adulteration by measuring differences between $^{13}C/^{12}C$ stable carbon isotope ratios of protein and sugar compounds with a combination of elemental analyzer-isotope ratio mass spectrometry and liquid chromatography-isotope ratio mass spectrometry (δ^{13}C-EA/LC-IRMS). *Apidologie*, *39*(5), 574–587. https://doi.org/10.1051/apido:2008042.
96. Bontempo, L.., Camin, F.., Larcher, R., Nicolini, G., Perini, M., Rossmann, A. (2009). Coast and year effect on H, O and C stable isotope ratios of Tyrrhenian and Adriatic italian olive oils. *Rapid Communications in Mass Spectrometry: An International Journal Devoted to the Rapid Dissemination of Up-to-the-Minute Research in Mass Spectrometry*, *23*(7), 1043–1048. https://doi.org/10.1002/rcm.3968.
97. Camin, F., Larcher, R., Perini, M., Bontempo, L., Bertoldi, D., Gagliano, G., Nicolini, G., Versini, G. (2010). Characterisation of authentic Italian extra-virgin olive oils by stable isotope ratios of C, O and H and mineral composition. *Food chemistry*, *118*(4), 901–909. https://doi.org/10.1016/j.foodchem.2008.04.059.
98. Martin, G. J., Martin, M. L. (2003). Climatic significance of isotope ratios. *Phytochemistry Reviews*, *2*(1), 179–190. https://doi.org/10.1023/B:PHYT.0000004187.23624.dd.
99. Scartazza, A., Mata, C., Matteucci, G., Yakir, D., Moscatello, S., Brugnoli, E. (2004). Comparisons of δ^{13}C of photosynthetic products and ecosystem respiratory CO_2 and their responses to seasonal climate variability. *Oecologia*, *140*, 340–351. https://doi.org/10.1007/s00442-004-1588-1.
100. Portarena, S., Farinelli, D., Lauteri, M., Famiani, F., Esti, M., Brugnoli, E. (2015). Stable isotope and fatty acid compositions of monovarietal olive oils: Implications of ripening stage and climate effects as determinants in traceability studies. *Food Control*, *57*, 129–135. https://doi.org/10.1016/j.foodcont.2015.03.052.
101. Bontempo, L., Paolini, M., Franceschi, P., Ziller, L., García-González, D. L., Camin, F. (2019). Characterisation and attempted differentiation of European and extra-European olive oils using stable isotope ratio analysis. *Food Chemistry*, *276*, 782–789. https://doi.org/10.1016/j.foodchem.2018.10.077.
102. Paolini, M., Bontempo, L., Camin, F. (2017). Compound-specific δ^{13}C and δ^{2}H analysis of olive oil fatty acids. *Talanta*, *174*, 38–43. https://doi.org/10.1016/j.talanta.2017.05.080.

103. Chung, I. M., Kim, Y. J., Moon, H. S., Han, J. G., Kong, W. S., Yarnes, C. T., Kim, S. H. (2022). Improved accuracy of geographical origin identification of shiitake grown in sawdust medium: A compound-specific isotope model-based pilot study. *Food Chemistry*, *369*, 130955. https://doi.org/10.1016/j.foodchem.2021.130955.
104. Faberi A., Marianella R. M., Fuselli F., La Mantia A., Ciardiello F., Montesano C., Mascini M., Sergi M., Compagnone D. (2014). Fatty acid composition and $\delta^{13}C$ of bulk and individual fatty acids as marker for authenticating Italian PDO/PGI extra virgin olive oils by means of isotopic ratio mass spectrometry. *Journal of Mass Spectrometry*, *49*, 840–849. https://doi.org/10.1002/jms.3399.
105. Schellenberg, A., Chmielus, S., Schlicht, C., Camin, F., Perini, M., Bontempo, L., Heinrich, K., Kelly, S. D., Rossmann, A., Thomas, F., Jamin, E., Horacek, M. (2010). Multielement stable isotope ratios (H, C, N, S) of honey from different European regions. *Food Chemistry*, *121*(3), 770–777. https://doi.org/10.1016/j.foodchem.2009.12.082.
106. Yoshida, M., Tabata, A., Niino, T., Chiku, K., Nakashita, R., Suzuki, Y. (2022). Potential application of light element stable isotope ratio in crude fiber for geographical origin verification of raw and cooked kabocha pumpkin (Cucurbita maxima). *Food Chemistry*, *373*, 131462. https://doi.org/10.1016/j.foodchem.2021.131462.
107. Kelly, S., Brereton, P., Guillou, C., Broll, H., Laube, I., Downey, G., Rossman, A., Hözl, S., Van der Veer, G. (2011). New approaches to determining the origin of food. In *Food Chain Integrity* (pp. 238-258). Woodhead Publishing. https://doi.org/10.1533/9780857092519.3.238.
108. Drivelos, S. A., Georgiou, C. A. (2012). Multi-element and multi-isotope-ratio analysis to determine the geographical origin of foods in the European Union. *TrAC Trends in Analytical Chemistry*, *40*, 38-51. http://dx.doi.org/10.1016/j.trac.2012.08.003.
109. Irrgeher, J., Prohaska, T. (2016). Application of non-traditional stable isotopes in analytical ecogeochemistry assessed by MC ICP-MS-A critical review. *Analytical and Bioanalytical Chemistry*, *408*, 369-385. https://doi.org/10.1007/s00216-015-9025-3.
110. Luykx, D. M. A. M., van Ruth, S. M. (2008). An overview of analytical methods for determining the geographical origin of food products. *Food Chemistry*, *107*, 897–911. https://doi.org/10.1016/j.foodchem.2007.09.038.
111. Böhme, K., Calo-Mata, P., Barros-Velázquez, J., Ortea, I. (2019). Review of recent DNA-based methods for main food-authentication topics. *Journal of agricultural and food chemistry*, *67*(14), 3854-3864. https://doi.org/10.1021/acs.jafc.8b07016.
112. Scarano, D., Rao, R. (2014). DNA markers for food products authentication. *Diversity*, *6*(3), 579-596. https://doi.org/10.3390/d6030579.

113. Wadood, S. A., Boli, G., Xiaowen, Z., Hussain, I., Yimin, W. (2020). Recent development in the application of analytical techniques for the traceability and authenticity of food of plant origin. *Microchemical Journal*, *152*, 104295. https://doi.org/10.1016/j.microc.2019.104295.
114. Fanelli, V., Mascio, I., Miazzi, M. M., Savoia, M. A., De Giovanni, C., Montemurro, C. (2021). Molecular approaches to agri-food traceability and authentication: An updated review. *Foods*, *10*(07), 1644. https://doi.org/10.3390/foods10071644.
115. Khalil, I., Hashem, A., Nath, A. R., Julkapli, N. M., Yehye, W. A., Basirun, W. J. (2021). DNA/Nano based advanced genetic detection tools for authentication of species: Strategies, prospects and limitations. *Molecular and Cellular Probes*, *59*, 101758. https://doi.org/10.1016/j.mcp.2021.101758.
116. Kumar, P., Rani, A., Singh, S., Kumar, A. (2022). Recent advances on DNA and omics-based technology in Food testing and authentication: A review. *Journal of Food Safety*, *42*(4), e12986. https://doi.org/10.1111/jfs.12986.
117. Madesis, P., Ganopoulos, I., Sakaridis, I., Argiriou, A., Tsaftaris, A. (2014). Advances of DNA-based methods for tracing the botanical origin of food products. *Food Research International*, *60*, 163-172. https://doi.org/10.1016/j.foodres.2013.10.042.
118. Johanningsmeier, S. D., Harris, G. K., Klevorn, C. M. (2016). Metabolomic technologies for improving the quality of food: practice and promise. *Annual review of food science and technology*, *7*(1), 413-438. https://doi.org/10.1146/annurev-food-022814-015721.
119. Pagare, S., Bhatia, M., Tripathi, N., Pagare, S., Bansal, Y. K. (2015). Secondary metabolites of plants and their role: Overview. *Current trends in biotechnology and pharmacy*, *9*(3), 293-304. https://www.indianjournals.com/ijor.aspx?target=ijor:ctbp&volume=9&issue=3&article=011&type=pdf.
120. Pedrosa, M. C., Lima, L., Heleno, S., Carocho, M., Ferreira, I. C., & Barros, L. (2021). Food metabolites as tools for authentication, processing, and nutritive value assessment. *Foods*, *10*(9), 2213. https://doi.org/10.3390/foods10092213.
121. Pichersky, E., Gang, D. R. (2000). Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. *Trends in plant science*, *5*(10), 439-445. https://doi.org/10.1016/S1360-1385(00)01741-6.
122. Böhme, K., Calo-Mata, P., Barros-Velázquez, J., Ortea, I. (2019). Recent applications of omics-based technologies to main topics in food authentication. *TrAC Trends in Analytical Chemistry*, *110*, 221-232. https://doi.org/10.1016/j.trac.2018.11.005.

123. Cubero-Leon, E., Peñalver, R., Maquet, A. (2014). Review on metabolomics for food authentication. *Food Research International*, *60*, 95-107. https://doi.org/10.1016/j.foodres.2013.11.041.
124. Selamat, J., Rozani, N. A. A., & Murugesu, S. (2021). Application of the metabolomics approach in food authentication. *Molecules*, *26*(24), 7565. https://doi.org/10.3390/molecules26247565.
125. Afzaal, M., Saeed, F., Hussain, M., Shahid, F., Siddeeg, A., Al-Farga, A. (2022). Proteomics as a promising biomarker in food authentication, quality and safety: A review. *Food Science & Nutrition*, *10*(7), 2333-2346. https://doi.org/10.1002/fsn3.2842.
126. Bajoub, A., Medina-Rodríguez, S., Gómez-Romero, M., Bagur-González, M. G., Fernández-Gutiérrez, A., Carrasco-Pancorbo, A. (2017). Assessing the varietal origin of extra-virgin olive oil using liquid chromatography fingerprints of phenolic compound, data fusion and chemometrics. *Food Chemistry*, *215*, 245-255. https://doi.org/10.1016/j.foodchem.2016.07.140.
127. Blasi, F., Ianni, F., Cossignani, L. (2024). Phenolic profiling for geographical and varietal authentication of extra virgin olive oil. *Trends in Food Science & Technology*, 104444. https://doi.org/10.1016/j.tifs.2024.104444.
128. Chien, H. J., Zheng, Y. F., Wang, W. C., Kuo, C. Y., Hsu, Y. M., Lai, C. C. (2023). Determination of adulteration, geographical origins, and species of food by mass spectrometry. *Mass Spectrometry Reviews*, *42*(6), 2273-2323. https://doi.org/10.1002/mas.21780.
129. Ortea, I., O'Connor, G., Maquet, A. (2016). Review on proteomics for food authentication. *Journal of proteomics*, *147*, 212-225. https://doi.org/10.1016/j.jprot.2016.06.033.
130. García-González, D. L., Aparicio, R. (2010). Research in olive oil: challenges for the near future. *Journal of agricultural and food chemistry*, *58*(24), 12569-12577. https://doi.org/10.1021/jf102735n.
131. Servili, M., Selvaggini, R., Esposto, S., Taticchi, A., Montedoro, G., Morozzi, G. (2004). Health and sensory properties of virgin olive oil hydrophilic phenols: agronomic and technological aspects of production that affect their occurrence in the oil. *Journal of Chromatography A*, *1054*(1-2), 113-127. https://doi.org/10.1016/j.chroma.2004.08.070.
132. Parcerisa, J., Boatella, J., Codony, R., Farrà, A., Garcia, J., Lopez, A., Rafecas, M., Romero, A. (1993). Influence of variety and geographical origin on the lipid fraction of hazelnuts

(Corylus avellana L.) from Spain: I. Fatty acid composition. *Food Chemistry*, *48*(4), 411-414. https://doi.org/10.1016/0308-8146(93)90326-B.

133. Parcerisa, J., Rafecas, M., Castellote, A. I., Codony, R., Farran, A., Garcia, J., López, A., Romero, A., Boatella, J. (1994). Influence of variety and geographical origin on the lipid fraction of hazelnuts (Coryllus avellana L.) from Spain:(II). Triglyceride composition. *Food chemistry*, *50*(3), 245–249. https://doi.org/10.1016/0308-8146(94)90128-7.
134. Beltrán Sanahuja, A., Grané Teruel, N., Martín Carratalá, M. L., Garrigós Selva, M. C. (2011). Characterization of almond cultivars by the use of thermal analysis techniques. Application to cultivar authenticity. *Journal of the American Oil Chemists' Society*, *88*, 1687-1693. https://doi.org/10.1007/s11746-011-1847-3.
135. Verdú, A., Vazquez-Araujo, L., Miquel, A., Martinez-Sanchez, F., Carbonell-Barrachina, A. A. (2010). Discriminant analysis of almond cultivars used in turrón. *Italian Journal of Food Science*, *22*(1), 76-82. https://www.proquest.com/scholarly-journals/discriminant-analysis-almond-cultivars-used/docview/876961321/se-2.
136. Avula, B., Smillie, T. J., Wang, Y. H., Zweigenbaum, J., Khan, I. A. (2015). Authentication of true cinnamon (Cinnamon verum) utilising direct analysis in real time (DART)-QToF-MS. *Food Additives & Contaminants: Part A*, *32*(1), 1–8. https://doi.org/10.1080/19440049.2014.981763.
137. Marti, G., Boccard, J., Mehl, F., Debrus, B., Marcourt, L., Merle, P., Delort, E., Baroux, L., Sommer, H., Rudaz, S., Wolfender, J. L. (2014). Comprehensive profiling and marker identification in non-volatile citrus oil residues by mass spectrometry and nuclear magnetic resonance. *Food chemistry*, *150*, 235–245. https://doi.org/10.1016/j.foodchem.2013.10.103.
138. Matsushita, T., Zhao, J. J., Igura, N., Shimoda, M. (2018). Authentication of commercial spices based on the similarities between gas chromatographic fingerprints. *Journal of the Science of Food and Agriculture*, *98*(8), 2989–3000. https://doi.org/10.1002/jsfa.8797.
139. Quintanilla-Casas, B., Torres-Cobos, B., Guardiola, F., Romero, A., Tres, A., Vichi, S. (2022). Geographical authentication of virgin olive oil by GC-MS sesquiterpene hydrocarbon fingerprint: Scaling down to the verification of PDO compliance. *Food Control*, *139*, 109055. https://doi.org/10.1016/j.foodcont.2022.109055.
140. Quintanilla-Casas, B., Torres-Cobos, B., Guardiola, F., Servili, M., Alonso-Salces, R. M., Valli, E., Bendini, A., Toschi, T. G., Vichi, S., Tres, A. (2022). Geographical authentication of virgin olive oil by GC–MS sesquiterpene hydrocarbon fingerprint: Verifying EU and

single country label-declaration. *Food Chemistry*, *378*, 132104. https://doi.org/10.1016/j.foodchem.2022.132104 .

141. Vichi, S., Riu-Aumatell, M., Mora-Pons, M., Buxaderas, S., López-Tamames, E. (2005). Characterization of volatiles in different dry gins. *Journal of agricultural and food chemistry*, *53*(26), 10154–10160. https://doi.org/10.1021/jf058121b.
142. Degenhardt, J., Köllner, T. G., Gershenzon, J. (2009). Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry*, *70*(15-16), 1621–1637. https://doi.org/10.1016/j.phytochem.2009.07.030.
143. Vichi, S., Tres, A., Quintanilla-Casas, B., Bustamante, J., Guardiola, F. (2018). Sesquiterpene hydrocarbons, a promising tool for virgin olive oil geographical authentication. In M. Kontominas (Ed.), *Authentication and detection of Adulteration of olive oil*. New York, NY: Nova Science Publishers, Inc., ISBN 978-1-53614-596-0.
144. Sales, C., Portolés, T., Johnsen, L. G., Danielsen, M., Beltran, J. (2019). Olive oil quality classification and measurement of its organoleptic attributes by untargeted GC–MS and multivariate statistical-based approach. *Food chemistry*, *271*, 488-496. https://doi.org/10.1016/j.foodchem.2018.07.200.
145. Valli, E., Panni, F., Casadei, E., Barbieri, S., Cevoli, C., Bendini, A., García-González, D. L., Gallina Toschi, T. (2020). An HS-GC-IMS method for the quality classification of virgin olive oils as screening support for the panel test. *Foods*, *9*(5), 657. https://doi.org/10.3390/foods9050657.
146. Vichi, S., Guadayol, J. M., Caixach, J., López-Tamames, E., Buxaderas, S. (2007). Comparative study of different extraction techniques for the analysis of virgin olive oil aroma. *Food Chemistry*, *105*(3), 1171-1178. https://doi.org/10.1016/j.foodchem.2007.02.018.
147. López-Feria, S., Cárdenas, S., García-Mesa, J. A., Valcárcel, M. (2007). Usefulness of the direct coupling headspace–mass spectrometry for sensory quality characterization of virgin olive oil samples. *Analytica chimica acta*, *583*(2), 411-417. https://doi.org/10.1016/j.aca.2006.10.027.
148. Zunin, P., Boggia, R., Lanteri, S., Leardi, R., De Andreis, R., Evangelisti, F. (2004). Direct thermal extraction and gas chromatographic–mass spectrometric determination of volatile compounds of extra-virgin olive oils. *Journal of Chromatography A*, *1023*(2), 271-276. https://doi.org/10.1016/j.chroma.2003.10.035.
149. Alonso-Salces, R. M., Héberger, K., Holland, M. V., Moreno-Rojas, J. M., Mariani, C., Bellan, G., Reniero, F., Guillou, C. (2010). Multivariate analysis of NMR fingerprint of

the unsaponifiable fraction of virgin olive oils for authentication purposes. *Food Chemistry*, *118*(4), 956-965. https://doi.org/10.1016/j.foodchem.2008.09.061.

150. Benitez-Sanchez, P. L., León-Camacho, M. L., Aparicio, R. (2003). A comprehensive study of hazelnut oil composition with comparisons to other vegetable oils, particularly olive oil. *European Food Research and Technology*, *218*, 13–19. https://doi.org/10.1007/s00217-003-0766-4.
151. Goriainov, S. V., Esparza, C. A., Borisova, A. R., Orlova, S. V., Vandyshev, V. V., Hajjar, F., Platonov, E. A., Chromchenkova, E. P., Novikov, O. O., Borisov, R. S., Kalabin, G. A. (2021). Phytochemical study of the composition of the unsaponifiable fraction of various vegetable oils by gas chromatography-mass spectrometry. *Journal of Analytical Chemistry*, *76*, 1635–1644. https://doi.org/10.1134/S1061934821140045.
152. Lercker, G., Rodriguez-Estrada, M. T. (2000). Chromatographic analysis of unsaponifiable compounds of olive oils and fat-containing foods. *Journal of Chromatography A*, *881*(1-2), 105-129. https://doi.org/10.1016/S0021-9673(00)00455-6.
153. Ghisoni, S., Lucini, L., Rocchetti, G., Chiodelli, G., Farinelli, D., Tombesi, S., Trevisan, M. (2020). Untargeted metabolomics with multivariate analysis to discriminate hazelnut (Corylus avellana L.) cultivars and their geographical origin. *Journal of the Science of Food and Agriculture*, *100*, 500–508. https://doi.org/10.1002/jsfa.9998.
154. Matthäus, B., Özcan, M. M. (2012). The comparison of properties of the oil and kernels of various hazelnuts from Germany and Turkey. *European Journal of Lipid Science and Technology*, *114*, 801–806. https://doi.org/10.1002/ejlt.201100299.
155. Parcerisa, J., Richardson, D. G., Rafecas, M., Codony, R., Boatella, J. (1998). Fatty acid, tocopherol and sterol content of some hazelnut varieties (Corylus avellana L.) harvested in Oregon (USA). *Journal of Chromatography A*, *805*, 259–268. https://doi.org/10.1016/S0021-9673(98)00049-1.
156. AOAC Official Method 996.06 (2000). *Fat (total, saturated and unsaturated) in foods. Hydrolytic extraction gas chromatography method*. In: Official Methods of Analysis of AOAC International, (17th ed.). AOAC International, Maryland.
157. Petrović, M., Kezić, N., Bolanča, V. (2010). Optimization of the GC method for routine analysis of the fatty acid profile in several food samples. *Food Chemistry*, *122*(1), 285–291. https://doi.org/10.1016/j.foodchem.2010.02.018.
158. Ruiz-Samblás, C., González-Casado, A., Cuadros-Rodríguez, L. (2015). Triacylglycerols determination by high-temperature gas chromatography in the analysis of vegetable

oils and foods: a review of the past 10 years. *Critical Reviews in Food Science and Nutrition*, *55*(11), 1618–1631. https://doi.org/10.1080/10408398.2012.713045.

159. Andrikopoulos, N. K. (2002). Chromatographic and spectroscopic methods in the analysis of triacylglycerol species and regiospecific isomers of oils and fats. *Critical reviews in food science and nutrition*, *42*(5), 473-505. https://doi.org/10.1080/20024091054229.
160. Aparicio, R., Morales, M. T., Aparicio-Ruiz, R., Tena, N., García-González, D. L. (2013). Authenticity of olive oil: Mapping and comparing official methods and promising alternatives. *Food Research International*, *54*(2), 2025-2038. https://doi.org/10.1016/j.foodres.2013.07.039.
161. Buchgraber, M., Ulberth, F., Emons, H., Anklam, E. (2004). Triacylglycerol profiling by using chromatographic techniques. *European Journal of Lipid Science and Technology*, *106*(9), 621-648. https://doi.org/10.1002/ejlt.200400986.
162. Firestone, D. (1994). Liquid chromatographic method for determination of triglycerides in vegetable oils in terms of their partition numbers: summary of collaborative study. *Journal of AOAC International*, *77*(4), 954-957. https://doi.org/10.1093/jaoac/77.4.954.
163. Official Journal of the European Union. (2022). European Commission *Regulation No 2022/2105 laying down rules on conformity checks of marketing standards for olive oil and methods of analysis of the characteristics of olive oil*, pp. 23–48.
164. Wolfe, W. L., Zissis, G. J. (Eds.). (1985). *The infrared handbook*. General Dynamics. ISBN: 978-0960359011.
165. Ballin, N. Z., Laursen, K. H. (2019). To target or not to target? Definitions and nomenclature for targeted versus non-targeted analytical food authentication. *Trends in Food Science & Technology*, *86*, 537−543. https://doi.org/10.1016/j.tifs.2018.09.025.
166. Pieke, E. N., Granby, K., Trier, X., Smedsgaard, J. (2017). A framework to estimate concentrations of potentially unknown substances by semi-quantification in liquid chromatography electrospray ionization mass spectrometry. *Analytica Chimica Acta*, *975*, 30−41. https://doi.org/10.1016/j.aca.2017.03.054.
167. International Organization for Standardization (ISO). (2005). ISO/IEC 17025 General requirements for the competence of testing and calibration laboratories.
168. Quintanilla-Casas, B., Bertin, S., Leik, K., Bustamante, J., Guardiola, F., Valli, E., Bendini, A., Toschi, T. G., Tres, A., Vichi, S. (2020). Profiling versus fingerprinting analysis of sesquiterpene hydrocarbons for the geographical authentication of extra virgin olive oils. *Food Chemistry*, *307*, 125556. https://doi.org/10.1016/j.foodchem.2019.125556.

169. Cuadros-Rodríguez, L., Ruiz-Samblás, C., Valverde-Som, L., Pérez-Castaño, E., González-Casado, A. (2016). Chromatographic fingerprinting: An innovative approach for food 'identitation' and food authentication–A tutorial. *Analytica Chimica Acta*, *909*, 9−23. https://doi.org/10.1016/j.aca.2015.12.042.
170. Baccolo, G., Quintanilla-Casas, B., Vichi, S., Augustijn, D., Bro, R. (2021). From untargeted chemical profiling to peak tables–A fully automated AI driven approach to untargeted GC-MS. *TrAC Trends in Analytical Chemistry*, *145*, 116451. https://doi.org/10.1016/j.trac.2021.116451.
171. Rinnan, Å., Amigo, J. M., Skov, T. (2014). Multiway methods in food science. In D. Granato, G. Ares (Eds.) *Mathematical and statistical methods in food science and technology* (pp. 143−174). John Wiley & Sons. https://doi.org/10.1002/9781118434635.ch9.
172. Creydt, M., Fischer, M. (2020). Food authentication in real life: How to link nontargeted approaches with routine analytics?. *Electrophoresis*, *41*(20), 1665−1679. https://doi.org/10.1002/elps.202000030.
173. Esslinger, S., Riedl, J., Fauhl-Hassek, C. (2014). Potential and limitations of non-targeted fingerprinting for authentication of food in official control. *Food Research International*, *60*, 189−204. https://doi.org/10.1016/j.foodres.2013.10.015.
174. McGrath, T. F., Haughey, S. A., Patterson, J., Fauhl-Hassek, C., Donarski, J., Alewijn, M., van Ruth, S., Elliott, C. T. (2018). What are the scientific challenges in moving from targeted to non-targeted methods for food fraud testing and how can they be addressed?–Spectroscopy case study. *Trends in Food Science & Technology*, *76*, 38−55. https://doi.org/10.1016/j.tifs.2018.04.001.
175. Eilers, P. H. (2004). Parametric time warping. *Analytical chemistry*, *76*(2), 404−411. https://doi.org/10.1021/ac034800e.
176. Eilers, P. H., Boelens, H. F. (2005). Baseline correction with asymmetric least squares smoothing. *Leiden University Medical Centre Report*, *1*(1), 5.
177. Hilario, M., Kalousis, A., Pellegrini, C., Mueller, M. (2006). Processing and classification of protein mass spectra. *Mass spectrometry reviews*, *25*(3), 409−449. https://doi.org/10.1002/mas.20072.
178. Niezen, L. E., Schoenmakers, P. J., Pirok, B. W. (2022). Critical comparison of background correction algorithms used in chromatography. *Analytica Chimica Acta*, *1201*, 339605. https://doi.org/10.1016/j.aca.2022.339605.

179. Xu, Z., Sun, X., Harrington, P. D. B. (2011). Baseline correction method using an orthogonal basis for gas chromatography/mass spectrometry data. *Analytical chemistry*, *83*(19), 7464−7471. https://doi.org/10.1021/ac2016745.
180. Crawford, L. R., Morrison, J. D. (1968). Computer methods in analytical mass spectrometry. Identification of an unknown compound in a catalog. *Analytical chemistry*, *40*(10), 1464−1469. https://doi.org/10.1021/ac60266a027.
181. Nam, S. L., de la Mata, A. P., Dias, R. P., Harynuk, J. J. (2020). Towards standardization of data normalization strategies to improve urinary metabolomics studies by GC× GC-TOFMS. *Metabolites*, *10*(9), 376. https://doi.org/10.3390/metabo10090376.
182. Amigo, J. M., Skov, T., Bro, R. (2010). ChroMATHography: solving chromatographic issues with mathematical models and intuitive graphics. *Chemical reviews*, *110*(8), 4582−4605. https://doi.org/10.1021/cr900394n.
183. Nielsen, N. P. V., Carstensen, J. M., Smedsgaard, J. (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of chromatography A*, *805*(1-2), 17−35. https://doi.org/10.1016/S0021-9673(98)00021-1.
184. Tomasi, G., Van Den Berg, F., Andersson, C. (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *18*(5), 231−241. https://doi.org/10.1002/cem.859.
185. Bloemberg, T. G., Gerretzen, J., Lunshof, A., Wehrens, R., Buydens, L. M. (2013). Warping methods for spectroscopic and chromatographic signal alignment: a tutorial. *Analytica chimica acta*, *781*, 14−32. https://doi.org/10.1016/j.aca.2013.03.048.
186. Saccenti, E., Hoefsloot, H. C., Smilde, A. K., Westerhuis, J. A., Hendriks, M. M. (2014). Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics*, *10*, 361−374. https://doi.org/10.1007/s11306-013-0598-6.
187. Berrueta, L. A., Alonso-Salces, R. M., Héberger, K. (2007). Supervised pattern recognition in food analysis. *Journal of chromatography A*, *1158*(1-2), 196−214. https://doi.org/10.1016/j.chroma.2007.05.024.
188. Medina, S., Perestrelo, R., Silva, P., Pereira, J. A., Câmara, J. S. (2019). Current trends and recent advances on food authenticity technologies and chemometric approaches. *Trends in Food Science & Technology*, *85*, 163−176. https://doi.org/10.1016/j.tifs.2019.01.017.

189. González-Domínguez, R., Sayago, A., Fernández-Recamales, Á. (2022). An overview on the application of chemometrics tools in food authenticity and traceability. *Foods*, *11*(23), 3940. https://doi.org/10.3390/foods11233940.
190. Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, *24*(6), 417. https://psycnet.apa.org/doi/10.1037/h0071325.
191. Jolliffe, I. T., Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202.
192. Wold, S., Esbensen, K., Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), 37−52. https://doi.org/10.1016/0169-7439(87)80084-9.
193. Noshad, M., Behbahani, B. A., Karabagias, I. K. (2023). Volatilomic with chemometrics: a toward authentication approach for food authenticity control. *European Food Research and Technology*, *249*(9), 2215−2226. https://doi.org/10.1007/s00217-023-04291-0.
194. Oliveri, P. (2017). Class-modelling in food analytical chemistry: development, sampling, optimisation and validation issues–a tutorial. *Analytica Chimica Acta*, *982*, 9−19. https://doi.org/10.1016/j.aca.2017.05.013.
195. Tres, A., Ruiz-Samblas, C., Van Der Veer, G., Van Ruth, S. M. (2013). Geographical provenance of palm oil by fatty acid and volatile compound fingerprinting techniques. *Food Chemistry*, *137*(1-4), 142−150. https://doi.org/10.1016/j.foodchem.2012.09.094.
196. Christoph, N., Rossmann, A., Voerkelius, S. (2003). Possibilities and limitations of wine authentication using stable isotope and meteorological data, data banks and statistical tests. Part 1: Wines from Franconia and Lake Constance 1992 to 2001. *Mitteilungen Klosterneuburg*, *53*, 23−40. Available online: https://www.weinobst.at/dam/jcr:3fb45e81-d0ea-47cc-9486-b574c0abe3c7/23-2003.pdf (accessed on July 10 2024).
197. Camin, F., Pavone, A., Bontempo, L., Wehrens, R., Paolini, M., Faberi, A., Marianella, R. M., Capitani, D., Vista, S., Mannina, L. (2016). The use of IRMS, 1H NMR and chemical analysis to characterise Italian and imported Tunisian olive oils. *Food Chemistry*, *196*, 98–105. https://doi.org/10.1016/j.foodchem.2015.08.132.

198. Nie, J., Yang, J., Liu, C., Li, C., Shao, S., Yao, C., Chen, B., Tao, Y., Wang, F., Zhang, Y., Rogers, K. M., Wang, P., Yuan, Y. (2023). Stable isotope and elemental profiles determine geographical origin of saffron from China and Iran. *Food Chemistry*, *405*, 134733. https://doi.org/10.1016/j.foodchem.2022.134733.
199. Medina, S., Pereira, J. A., Silva, P., Perestrelo, R., Câmara, J. S. (2019). Food fingerprints–A valuable tool to monitor food authenticity and safety. *Food Chemistry*, *278*, 144−162. https://doi.org/10.1016/j.foodchem.2018.11.046.
200. Riedl, J., Esslinger, S., Fauhl-Hassek, C. (2015). Review of validation and reporting of non-targeted fingerprinting approaches for food authentication. *Analytica Chimica Acta*, *885*, 17–32. https://doi.org/10.1016/j.aca.2015.06.003.
201. Bajoub, A., Bendini, A., Fernández-Gutiérrez, A., Carrasco-Pancorbo, A. (2018). Olive oil authentication: A comparative analysis of regulatory frameworks with especial emphasis on quality and authenticity indices, and recent analytical techniques developed for their assessment. A review. *Critical Reviews in Food Science and Nutrition*, *58*(5), 832–857. https://doi.org/10.1080/10408398.2016.1225666.
202. Tahir, H. E., Arslan, M., Mahunu, G. K., Mariod, A. A., Hashim, S. B., Xiaobo, Z., Jiyong, S., El-Seedi, H. R., Musa, T. H. (2022). The use of analytical techniques coupled with chemometrics for tracing the geographical origin of oils: A systematic review (2013–2020). *Food Chemistry*, *366*, 130633. https://doi.org/10.1016/j.foodchem.2021.130633.
203. Valli, E., Bendini, A., Berardinelli, A., Ragni, L., Riccò, B., Grossi, M., Toschi, T. G. (2016). Rapid and innovative instrumental approaches for quality and authenticity of olive oils. *European Journal of Lipid Science and Technology*, *118*(11), 1601–1619. https://doi.org/10.1002/ejlt.201600065.
204. Calò, F., Girelli, C. R., Wang, S. C., Fanizzi, F. P. (2022). Geographical origin assessment of extra virgin olive oil via NMR and MS combined with chemometrics as analytical approaches. *Foods*, *11*(1), 113. https://doi.org/10.3390/foods11010113.
205. Camin, F., Larcher, R., Nicolini, G., Bontempo, L., Bertoldi, D., Perini, M., Schlicht, C., Schellenberg, A., Thomas,F., Heinrich, K., Voerkelius, S., Horacek, M., Ueckermann, H., Froeschl, H., Wimmer, B., Heiss, G., Baxter, M., Rossmann, A., Hoogewerff, J. (2010). Isotopic and Elemental Data for Tracing the Origin of European Olive Oils. *Journal of Agricultural and Food Chemistry*, *58*, 570-577. https://doi.org/10.1021/jf902814s.
206. Lukić, I., Da Ros, A., Guella, G., Camin, F., Masuero, D., Mulinacci, N., Vrhovsek, U., Mattivi, F. (2020). Lipid Profiling and Stable Isotopic Data Analysis for Differentiation of

Extra Virgin Olive Oils Based on Their Origin. *Molecules*, *25*, 4. https://doi.org/10.3390/molecules25010004.

207. Agrimonti, C., Vietina, M., Pafundo, S., Marmiroli, N. (2011). The use of food genomics to ensure the traceability of olive oil. *Trends in Food Science & Technology*, *22*(5), 237-244. https://doi.org/10.1016/j.tifs.2011.02.002.
208. Ben-Ayed, R., Grati-Kamoun, N., Sans-Grout, C., Moreau, F., Rebai, A. (2012). Characterization and authenticity of virgin olive oil (Olea europaea L.) cultivars by microsatellite markers. *European Food Research and Technology*, *234*, 263-271. https://doi.org/10.1007/s00217-011-1631-5.
209. Kalaitzis, P., El-Zein, Z. (2016). Olive oil authentication, traceability and adulteration detection using DNA-based approaches. *Lipid Technology*, *28*(10-11), 173-176. https://doi.org/10.1002/lite.201600048.
210. Muzzalupo, I., Pisani, F., Greco, F., & Chiappetta, A. (2015). Direct DNA amplification from virgin olive oil for traceability and authenticity. *European Food Research and Technology*, *241*, 151-155. https://doi.org/10.1007/s00217-015-2455-5.
211. Rabiei, Z., & Enferadi, S. T. (2012). *Traceability of origin and authenticity of olive oil*. INTECH Open Access Publisher. https://www.researchgate.net/profile/Zohreh-Rabiei-2/publication/221923415_Traceability_of_Origin_and_Authenticity_of_Olive_Oil/links/02e7e52288f7fd56c7000000/Traceability-of-Origin-and-Authenticity-of-Olive-Oil.pdf
212. Guinda, Á., Lanzón, A., Albi, T. (1996). Differences in hydrocarbons of virgin olive oils obtained from several olive varieties. *Journal of Agricultural and Food Chemistry, 44*, 1723–1726. https://doi.org/10.1021/jf950571o.
213. Osorio-Bueno, E., Sánchez-Casas, J., Montaño-García, A., Gallardo-González, L. (2005). Discriminating power of the hydrocarbon content from virgin olive oil of Extremadura cultivars. *Journal of the American Oil Chemists' Society, 82*, 1–6. https:// doi.org/10.1007/s11746-005-1034-0.
214. Vichi, S., Lazzez, A., Grati-Kamoun, N., Lopez Tamames, E., Buxaderas, S. (2010). Evolution of sesquiterpene hydrocarbons in virgin olive oil during fruit ripening. *Journal of Agricultural and Food Chemistry, 58*, 6972–6976. https://doi.org/10.1021/jf100497c.
215. Bachmann, R., Klockmann, S., Haerdter, J., Fischer, M., Hackl, T. (2018). 1H NMR spectroscopy for determination of the geographical origin of hazelnuts. *Journal of Agricultural and Food Chemistry*, *66*, 11873−11879. https://doi.org/10.1021/acs.jafc.8b03724.
216. Biancolillo, A., De Luca, S., Bassi, S., Roudier, L., Bucci, R., Magrì, A. D., Marini, F. (2018). Authentication of an Italian PDO hazelnut ("Nocciola Romana") by NIR spectroscopy.

Environmental Science and Pollution Research, *25*, 28780–28786. https://doi.org/10.1007/s11356-018-1755-2.

217. Inaudi, P., Giacomino, A., Malandrino, M., La Gioia, C., Conca, E., Karak, T., Abollino, O. (2020). The inorganic component as a possible marker for quality and for authentication of the hazelnut's origin. *International Journal of Environmental Research and Public Health*, *17*, 447. https://doi.org/10.3390/ijerph17020447.
218. Klockmann, S., Reiner, E., Bachmann, R., Hackl, T., Fischer, M. (2016). Food fingerprinting: metabolomic approaches for geographical origin discrimination of hazelnuts (Corylus avellana) by UPLC-QTOF-MS. *Journal of Agricultural and Food Chemistry*, *64*, 9253–9262. https://doi.org/10.1021/acs.jafc.6b04433.
219. Krauß, S., Vieweg, A., Vetter, W. (2019). Stable isotope signatures ($\delta^{2}H$-, $\delta^{13}C$-, $\delta^{15}N$-values) of walnuts (Juglans regia L.) from different regions in Germany. *Journal of the Science of Food and Agriculture*, *100*, 1625–1634. https://doi.org/10.1002/jsfa.10174.
220. Lang, C., Weber, N., Möller, M., Schramm, L., Schelm, S., Kohlbacher, O., Fischer M. (2021). Genetic authentication: Differentiation of hazelnut cultivars using polymorphic sites of the chloroplast genome. *Food Control*, *130*, 108344. https://doi.org/10.1016/j.foodcont.2021.108344.
221. Moscetti, R., Radicetti, E., Monarca, D., Cecchini, M., Massantini, R. (2015). Near infrared spectroscopy is suitable for the classification of hazelnuts according to Protected Designation of Origin. *Journal of the Science of Food and Agriculture*, *95*(13), 2619–2625. https://doi.org/10.1002/jsfa.6992.
222. Sammarco, G., Bardin, D., Quaini, F., Dall'Asta, C., Christmann, J., Weller, P., Suman, M. (2023). A geographical origin assessment of Italian hazelnuts: Gas chromatography-ion mobility spectrometry coupled with multivariate statistical analysis and data fusion approach. *Food Research International*, *171*, 113085. https://doi.org/10.1016/j.foodres.2023.113085.
223. Shakiba, N., Gerdes, A., Holz, N., Wenck, S., Bachmann, R., Schneider, T., Seifert, S., Fischer, M., Hackl, T. (2022). Determination of the geographical origin of hazelnuts (Corylus avellana L.) by Near-Infrared spectroscopy (NIR) and a Low-Level Fusion with nuclear magnetic resonance (NMR). *Microchemical Journal*, *174*, 107066. https://doi.org/10.1016/j.microc.2021.107066.
224. Anderson, K. A., Smith, B. W. (2006). Effect of season and variety on the differentiation of geographic growing origin of pistachios by stable isotope profiling. *Journal of Agricultural and Food Chemistry*, *54*, 1747-1752. https://doi.org/10.1021/jf052928m.

225. Wadood, S.A., Nie, J., Li, C., Rogers, K.M., Zhang, Y., Yuan, Y. (2022). Geographical origin classification of peanuts and processed fractions using stable isotopes, *Food Chemistry: X*, *16*, Article 100456. https://doi.org/10.1016/j.fochx.2022.100456.
226. Sammarco, G., Rossi, M., Summan, M., Cavanna, D., Viotto, L., Pettenà, P., Dall'Asta, Iacumin, P. (2023). Hazelnut products traceability through combined isotope ratio mass spectrometry and multi-elemental analysis, *Journal of the Science of Food and Agriculture Reports*, *3*, 633-645. https://doi.org/10.1002/jsf2.171.
227. Oliveira, I., Sousa, A., Morais, J. S., Ferreira, I. C., Bento, A., Estevinho, L., Pereira, J. A. (2008). Chemical composition, and antioxidant and antimicrobial activities of three hazelnut (Corylus avellana L.) cultivars. *Food and Chemical Toxicology*, *46*(5), 1801–1807. https://doi.org/10.1016/j.fct.2008.01.026.
228. Kıralan, S., Yorulmaz, A., Şimşek, A., Tekin, A. (2015). Classification of Turkish hazelnut oils based on their triacylglycerol structures by chemometric analysis. *European Food Research and Technology*, *240*, 679-688. https://doi.org/10.1007/s00217-014-2371-0.
229. Manfredi, M., Robotti, E., Quasso, F., Mazzucco, E., Calabrese, G., Marengo, E. (2018). Fast classification of hazelnut cultivars through portable infrared spectroscopy and chemometrics. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *189*, 427–435. https://doi.org/10.1016/j.saa.2017.08.050.
230. Sammarco, G., Dall'Asta, C., Suman, M. (2023). Near infrared spectroscopy and multivariate statistical analysis as rapid tools for the geographical origin assessment of Italian hazelnuts. *Vibrational Spectroscopy*, *126*, 103531. https://doi.org/10.1016/j.vibspec.2023.103531.
231. Stella, E., Moscetti, R., Carletti, L., Menghini, G., Fabrizi, F., Cristofori, V., Monarca, D., Cecchini, M., Massantini, R. (2013). Near-infrared spectroscopy is feasible to discriminate hazelnut cultivars. *Journal of Agricultural Engineering*, *44*(s2). https://doi.org/10.4081/jae.2013.297.
232. Torres-Cobos, B., Quintanilla-Casas, B., Romero, A., Ninot, A., Alonso-Salces, R. M., Toschi, T. G., Bendini, A., Guardiola, F., Tres, A., Vichi, S. (2021). Varietal authentication of virgin olive oil: Proving the efficiency of sesquiterpene fingerprinting for Mediterranean Arbequina oils. *Food Control*, *128*, 108200. https://doi.org/10.1016/j.foodcont.2021.108200.
233. Alasalvar, C., Shahidi, F., Cadwallader, K. R. (2003). Comparison of natural and roasted Turkish tombul hazelnut (Corylus avellana L.) volatiles and flavor by DHA/GC/MS and descriptive sensory analysis. *Journal of Agricultural and Food Chemistry*, *51*(17), 5067−5072. https://doi.org/10.1021/jf0300846.

234. Burdack-Freitag, A.,Schieberle, P. (2010). Changes in the key odorants of Italian hazelnuts (Coryllus avellana L. Var. Tonda Romana) induced by roasting. *Journal of Agricultural and Food Chemistry*, *58*(10), 6351−6359. https://doi.org/10.1021/jf100692k.
235. Kesen, S., Sonmezdag, A. S., Kelebek, H., Selli, S. (2022). Comparison of aroma, aroma-active, and phenolic compounds of crude and refined hazelnut oils. *Journal of the American Oil Chemists' Society*, *99*(3), 265−275. https://doi.org/10.1002/aocs.12568.
236. Squara, S., Stilo, F., Cialiè Rosso, M., Liberto, E., Spigolon, N., Genova, G., Castello, G., Bicchi, C., Cordero, C. (2022). Corylus avellana L. aroma blueprint: Potent odorants signatures in the volatilome of high quality hazelnuts. *Frontiers in Plant Science*, *13*, 840028. https://doi.org/10.3389/fpls.2022.840028.
237. Cozzolino, D. (2012). Recent trends on the use of infrared spectroscopy to trace and authenticate natural and agricultural food products. *Applied Spectroscopy Reviews*, *47*(7), 518-530. https://doi.org/10.1080/05704928.2012.667858.
238. Arndt, M., Rurik, M., Drees, A., Ahlers, C., Feldmann, S., Kohlbacher, O., Fischer, M. (2021). Food authentication: Determination of the geographical origin of almonds (Prunus dulcis MILL.) via near-infrared spectroscopy. *Microchemical Journal*, *160*, 105702. https://doi.org/10.1016/j.microc.2020.105702.
239. Firmani, P., Bucci, R., Marini, F., Biancolillo, A. (2019). Authentication of “Avola almonds” by near infrared (NIR) spectroscopy and chemometrics. *Journal of Food Composition and Analysis*, *82*, 103235. https://doi.org/10.1016/j.jfca.2019.103235.
240. Teixeira, A. M., Sousa, C. (2019). A review on the application of vibrational spectroscopy to the chemistry of nuts. *Food chemistry*, *277*, 713–724. https://doi.org/10.1016/j.foodchem.2018.11.030.
241. Vega-Castellote, M., Sánchez, M. T., Torres-Rodríguez, I., Entrenas, J. A., Pérez-Marín, D. (2024). NIR Sensing Technologies for the Detection of Fraud in Nuts and Nut Products: A Review. *Foods*, *13*(11), 1612. https://doi.org/10.3390/foods13111612.
242. Vitale, R., Bevilacqua, M., Bucci, R., Magrì, A. D., Magrì, A. L., Marini, F. (2013). A rapid and non-invasive method for authenticating the origin of pistachio samples by NIR spectroscopy and chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *121*, 90–99. https://doi.org/10.1016/j.chemolab.2012.11.019.
243. Destaillats, F., Cruz-Hernandez, C., Giuffrida, F., Dionisi, F. (2010). Identification of the botanical origin of pine nuts found in food products by gas−liquid chromatography

analysis of fatty acid profile. *Journal of Agricultural and Food Chemistry*, *58*(4), 2082-2087. https://doi.org/10.1021/jf9041722.

244. Loewe, V., Navarro-Cerrillo, R. M., García-Olmo, J., Riccioli, C., Sánchez-Cuesta, R. (2017). Discriminant analysis of Mediterranean pine nuts (Pinus pinea L.) from Chilean plantations by near infrared spectroscopy (NIRS). *Food Control*, *73*, 634–643. https://doi.org/10.1016/j.foodcont.2016.09.012.
245. Ríos-Reina, R., Callejón, R. M., Amigo, J. M. (2021). Feasibility of a rapid and non-destructive methodology for the study and discrimination of pine nuts using near-infrared hyperspectral analysis and chemometrics, *Food Control*, *130*, 108365. https://doi.org/10.1016/j.foodcont.2021.108365.
246. Mahalovich, M. F., Kimsey, M. J., Fortin-Noreus, J. K., Robbins, C. T. (2016). Isotopic heterogeneity in whitebark pine (Pinus albicaulis Engelm.) nuts across geographic, edaphic and climatic gradients in the Northern Rockies (USA). *Forest Ecology and Management*, *359*, 174−189. https://doi.org/10.1016/j.foreco.2015.09.047.
247. Baffi, C., Trincherini, P. R. (2016). Food traceability using the $^{87}Sr/^{86}Sr$ isotopic ratio mass spectrometry. *European Food Research and Technology*, *242*, 1411−1439. https://doi.org/10.1007/s00217-016-2712-2.
248. Ballin, N. Z., Mikkelsen, K. (2016). Polymerase chain reaction and chemometrics detected several Pinus species including Pinus armandii involved in pine nut syndrome. *Food Control*, *64*, 234-239. https://doi.org/10.1016/j.foodcont.2015.12.036.
249. Handy, S. M., Parks, M. B., Deeds, J. R., Liston, A., De Jager, L. S., Luccioli, S., Kwegyir-Afful, E., Fardin-Kia, A. R., Begley, T. H., Rader, J. I., Diachenko, G. W. (2011). Use of the chloroplast gene ycf1 for the genetic differentiation of pine nuts obtained from consumers experiencing dysgeusia. *Journal of agricultural and food chemistry*, *59*(20), 10995-11002. https://doi.org/10.1021/jf203215v.
250. Handy, S. M., Timme, R. E., Jacob, S. M., Deeds, J. R. (2013). Development of a locked nucleic acid real-time polymerase chain reaction assay for the detection of Pinus armandii in mixed species pine nut samples associated with dysgeusia. *Journal of agricultural and food chemistry*, *61*(5), 1060-1066. https://doi.org/10.1021/jf304223a.
251. Nader, W., Brendel, T., Schubbert, R. (2013). DNA-analysis: enhancing the control of food authenticity through emerging technologies. *Agro Food Industry Hi Tech*, *24*, 42-46. https://cdnmedia.eurofins.com/eurofins-germany/media/2853908/1302_dna-analysis_food-authenticity_agro-food-industry-hi-tech.pdf.
252. Wolff, R. L., Pédrono, F., Pasquier, E., Marpeau, A. M. (2000). General characteristics of Pinus spp. Seed fatty acid compositions, and importance of Δ5-olefinic acids in the

taxonomy and phylogeny of the genus. *Lipids*, *35*(1), 1-22. https://doi.org/10.1007/s11745-000-0489-y.

253. Mara Adelina, M., Wang, H., Zhang, L., Zhao, Y. (2021). Comparative analysis of volatile profiles in two grafted pine nuts by headspace-SPME/GC–MS and electronic nose as responses to different roasting conditions. *Food Research International*, *140*, 110026. https://doi.org/10.1016/j.foodres.2020.110026.
254. Rogachev, A. D., Salakhutdinov, N. F. (2015). Chemical composition of *Pinus sibirica* (Pinaceae). *Chemistry & biodiversity*, *12*(1), 1–53. https://doi.org/10.1002/cbdv.201300195.
255. Nikolic, M. V., Jakovljevic, V. L., Bradic, J. V., Tomovic, M. T., Petrovic, B. P., Petrovic, A. M. (2022). Korean and siberian pine: review of chemical composition and pharmacological profile. *Acta Poloniae Pharmaceutica*, *79*(6), 785–797. https://doi.org/10.32383/appdr/161040.
256. Vichi, S., Guadayol, J. M., Caixach, J., López-Tamames, E., Buxaderas, S. (2006). Monoterpene and sesquiterpene hydrocarbons of virgin olive oil by headspace solid-phase microextraction coupled to gas chromatography/mass spectrometry. *Journal of Chromatography A*, *1125*(1), 117−123. https://doi.org/10.1016/j.chroma.2006.05.029.
257. Quintanilla-Casas, B., Strocchi, G., Bustamante, J., Torres-Cobos, B., Guardiola, F., Moreda, W., Martínez-Rivas, J. M., Valli, E., Bendini, A., Toschi, T. G., Tres, A., Vichi, S. (2021). Large-scale evaluation of shotgun triacylglycerol profiling for the fast detection of olive oil adulteration. Food Control, 123, 107851. https://doi.org/10.1016/j.foodcont.2020.107851.
258. Aguzzoni, A., Bassi, M., Robatscher, P., Tagliavini, M., Tirler, W., Scandellari, F. (2018). Plant Sr isotope ratios as affected by the Sr isotope ratio of the soil and of the external Sr inputs. *Journal of Agricultural and Food Chemistry*, *66*(40), 10513-10521. https://doi.org/10.1021/acs.jafc.8b02604.
259. Vitòria, L., Otero, N., Soler, A., Canals, A. (2004). Fertilizer Characterization: Isotopic data (N, S, O, C, and Sr). *Environmental Science & Technology*, *38*, 3254–3262. https://doi.org/10.1021/es0348187.
260. Quintanilla-Casas, B., Rinnan, Å., Romero, A., Guardiola, F., Tres, A., Vichi, S., Bro, R. (2022). Using fluorescence excitation-emission matrices to predict bitterness and pungency of virgin olive oil: A feasibility study. *Food Chemistry*, *395*, 133602. https://doi.org/10.1016/j.foodchem.2022.133602.

261. Cebi, N., Bekiroglu, H., Erarslan, A. (2023). Nondestructive Metabolomic Fingerprinting: FTIR, NIR and Raman Spectroscopy in Food Screening. *Molecules*, *28*(23), 7933. https://doi.org/10.3390/molecules28237933.
262. Ellis, D. I., Brewster, V. L., Dunn, W. B., Allwood, J. W., Golovanov, A. P., Goodacre, R. (2012). Fingerprinting food: current technologies for the detection of food adulteration and contamination. *Chemical Society Reviews*, *41*(17), 5706-5727. https://doi.org/10.1039/C2CS35138B.
263. Liu, H., Nie, J., Liu, Y., Wadood, S. A., Rogers, K. M., Yuan, Y., Gan, R. Y. (2023). A review of recent compound-specific isotope analysis studies applied to food authentication. *Food Chemistry*, *415*, 135791. https://doi.org/10.1016/j.foodchem.2023.135791.
264. Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern Recognition Letters*, *27*, 861-874. https://doi.org/10.1016/j.patrec.2005.10.010.
265. Barbieri, S., Cevoli, C., Bendini, A., Quintanilla-Casas, B., García-González, D. L., Toschi, T. G. (2020). Flash Gas Chromatography in Tandem with Chemometrics: A Rapid Screening Tool for Quality Grades of Virgin Olive Oils. *Foods*, *9*(7), 862. https://doi.org/10.3390/foods9070862.
266. Quintanilla-Casas, B., Marin, M., Guardiola, F., García-González, D. L., Barbieri, S., Bendini, A., Toschi, T. G., Vichi, S., Tres, A. (2020). Supporting the sensory panel to grade virgin olive oils: an in-house-validated screening tool by volatile fingerprinting and chemometrics. *Foods*, *9*(10), 1509. https://doi.org/10.3390/foods9101509.

CHAPTER 10.

ANNEXES

Annex 1. Supplementary material of Publication 2

Table S1. Sample information, composition of the training and validation (test) sets, and test samples misclassified by each isotopic and SH fingerprinting PLS-DA model. Due to its big size a part of the table is shown as a way of example.

					Training and validation (test) sets					
Primary ID	Country of origin	Year	ITA	Region	ITA1	ITA2	ITA3	REG1	REG2	REG3
Sample 1	ESP	2018/19	No_ITA	Not informed	training	training	training			
Sample 53	GRC	2018/19	No_ITA	Not informed	training	training	**test[2]**			
Sample 96	ITA	2018/19	ITA	Apulia	training	**test[1,2]**	training	training	test	training
Sample 336	POR	2018/19	No_ITA	Not informed	test	training	training			
Sample 370	TUN	2019/20	No_ITA	Not informed	**test[1]**	training	training			
Sample 378	TUR	2016/17	No_ITA	Not informed	training	training	test			

test[1]: samples misclassified by isotopic models; test[2]: samples misclassified by SH fingerprinting models; test[1,2]: samples misclassified by both isotopic and SH fingerprinting models.

Table S2. Thresholds of the PLSDA models (3 training sets per each type of model) for each class optimized with ROC curves. Thresholds of the PLSDA models build with a) isotopic data; b) sesquiterpene fingerprinting data.

a) Isotopic data

Set	**ITA/non-ITA**	**CAL**	**SIC**	**APU**
Set 1	0.364	0.270	0.295	0.310
Set 2	0.340	0.289	0.468	0.428
Set 3	0.356	0.489	0.396	0.285
Mean	0.353	0.349	0.386	0.341
SD	0.012	0.121	0.087	0.076

ITA: Italy; non-ITA: the other countries of origin. CAL: Calabria; SIC: Sicily; APU: Apulia.

b) Sesquiterpene fingerprinting data

Set	**ITA/non-ITA**	**CAL**	**SIC**	**APU**
Set 1	0.487	0.478	0.309	0.625
Set 2	0.422	0.482	0.334	0.600
Set 3	0.418	0.457	0.373	0.648
Mean	0.442	0.472	0.339	0.624
SD	0.038	0.013	0.032	0.024

ITA: Italy; non-ITA: the other countries of origin. CAL: Calabria; SIC: Sicily; APU: Apulia.

Table S3. Median, interquartile range (Q1 - Q3), maximum and minimum values (max - min), and non-parametric independent samples median test p-values to assess the differences in isotopic ratios across a) ITA/non-ITA; b) countries of origin; c) regions of origin.

a) ITA/non-ITA

	ITA (n= 242)			**non-ITA (n= 151)**		
	Median (‰)	Q_1 - Q_3 (‰)	max - min (‰)	Median (‰)	Q_1 - Q_3 (‰)	max - min (‰)
$\delta^{13}C$ (‰)	-29.7[a]	-30.1 – (-29.1)	-28.2 – (-31.6)	-29.6[a]	-29.9 –(-29.1)	-27.54 – (-30.73)
$\delta^{2}H$ (‰)	-144.0[a]	-148.8 – (-140.0)	-126.7 – (-170.2)	-143.2[a]	-147.7 – (-137.6)	-118.4 – (-157.6)
$\delta^{18}O$ (‰)	24.0[b]	23.3 -25.1	31.0 – 21.0	25.8[a]	24.9 – 27.1	32.5 - 22.9

	p-value
$\delta^{13}C$ (‰)	0.411[1]
$\delta^{2}H$ (‰)	0.649[1]
$\delta^{18}O$ (‰)	< 0.001[*1]

b) Countries of origin

	ITA (n= 242)			**ESP (n= 51)**		
	Median (‰)	Q_1 - Q_3 (‰)	max - min (‰)	Median (‰)	Q_1 - Q_3 (‰)	max - min (‰)
$\delta^{13}C$ (‰)	-29.7[b]	-30.1 – (-29.1)	-28.2 – (-31.6)	-29.7[b]	-29.9 – (-29.3)	-28.0 – (-30.7)
$\delta^{2}H$ (‰)	-144.0[d]	-148.8 – (-140.0)	-126.7 – (-170.2)	-145.8[de]	-148.4 – (140.8)	-118.4 – (-157.6)
$\delta^{18}O$ (‰)	24.0[d]	23.3 -25.1	31.0 – 21.0	26.3[b]	24.2 – 27.0	30.7 – 22.9
	GRC (n= 39)			**POR (n= 23)**		
	Median (‰)	Q_1 - Q_3 (‰)	max - min (‰)	Median (‰)	Q_1 - Q_3 (‰)	max - min (‰)
$\delta^{13}C$ (‰)	-29.2[a]	-29.7 – (-28.8)	-27.5 –(-30.6)	-29.9[abc]	-30.0 – (-29.2)	-28.48 – (-30.41)
$\delta^{2}H$ (‰)	-147.2[e]	-149.5 – (-145.0)	-130.8 – (-155.1)	-132.8[a]	-134.2 – (-128.7)	-126.5 – (-152.6)
$\delta^{18}O$ (‰)	25.1[c]	24.6 – 25.5	27.8 – 23.8	29.5[a]	28.9 – 30.5	32.5 – 25.9
	TUN (n= 17)			**TUR (n= 21)**		
	Median (‰)	Q_1 - Q_3 (‰)	max - min (‰)	Median (‰)	Q_1 - Q_3 (‰)	max - min (‰)
$\delta^{13}C$ (‰)	-30.1[c]	-30.4 – (-29.8)	-29.2 – (-30.5)	-29.2[a]	-29.7 – (-28.5)	-28.17 – (-30.5)
$\delta^{2}H$ (‰)	-142.5[c]	-143.3 – (-141.4)	-136.1 – (-145.6)	-137.6[b]	-144.6 – (-131-2)	-126.4 – (-149.2)
$\delta^{18}O$ (‰)	25.6[b]	25.3 – 26.4	27.1 – 23.4	26.0[b]	25.3 – 26.7	27.8 -23.7

	p-value
$\delta^{13}C$ (‰)	<0.001[*1]
$\delta^{2}H$ (‰)	<0.001[*1]
$\delta^{18}O$ (‰)	<0.001[*1]

c) Regions of origin

	CAL (n= 58)			SIC (n= 40)		
	Median (‰)	Q_1 - Q_3 (‰)	max - min (‰)	Median (‰)	Q_1 - Q_3 (‰)	max - min (‰)
$\delta^{13}C$ (‰)	-29.5[b]	-29.9 – (-29.0)	-28.5 – (-30.9)	-28.9[a]	-29.1 – (-28.9)	-28.4 – (-30.0)
$\delta^{2}H$ (‰)	-141.8[a]	-145.0 – (135.8)	-127.3 – (-149.2)	-146.5[b]	-150.8 – (-143.7)	-135.8 – (-156.7)
$\delta^{18}O$ (‰)	24.5[ab]	23.5 – 25.9	27.6 -22.0	25.1[a]	24.4 – 25.7	26.4 – 23.6

	APU (n= 73)			
	Median (‰)	Q_1 - Q_3 (‰)	max - min (‰)	p-value
$\delta^{13}C$ (‰)	-29.9[c]	-30.2 – (-29.5)	-28.2 – (-30.7)	< 0.001[*1]
$\delta^{2}H$ (‰)	-142.0[a]	-146.4 – (-139.0)	-128.4 – (-156.4)	< 0.001[*1]
$\delta^{18}O$ (‰)	23.7[b]	23.0 – 24.8	28.7 – 21.9	< 0.001[*1]

* p-value < 0.05 means are significantly different. [1] p-value obtained from independent samples median test. Differences between regions were noted in the same row as a > b > c > d > e. ITA: Italy; non-ITA: other countries of origin. ESP: Spain; GRC: Greece; POR: Portugal; TUN: Tunisia; TUR: Turkey; CAL: Calabria; SIC: Sicily; APU: Apulia.

Table S4. Results of the leave-10%-out cross-validation of the PLS-DA training models developed on the a) isotopic data; b) sesquiterpene fingerprinting data, to discriminate samples in Italy/non-Italy. Results are mean values (± standard deviation) obtained from three iterations.

a) Isotopic data

	n	non-ITA (n)	ITA (n)	Correct classification (%)	Sensitivity	Specificity
non-ITA	119	96 ± 1.7	23 ± 1.7	80.7 ± 1.5	0.81 ± 0.02	
ITA	194	56.7 ± 3.2	137.3 ± 3.2	70.8 ± 1.7		0.71 ± 0.02
Total	313			74.5 ± 1.0		

Model parameters: mean values obtained with the training sets from 3 iterations. 2 LVs, Q2 = 0.299, RMSEcv = 0.406. For all models, ANOVA p-value < 0.05. ITA: Italy; non-ITA: the other countries of origin.

b) Sesquiterpene fingerprinting data

	n	non-ITA (n)	ITA (n)	Correct classification (%)	Sensitivity	Specificity
non-ITA	119	118.7 ± 0.6	0.3 ± 0.6	99.7 ± 0.5	1.00 ± 0.01	
ITA	194	0.0 ± 0.0	194.0 ± 0.0	100.0 ± 0.0		1.00 ± 0.00
Total	313			99.9 ± 0.2		

Model parameters: mean values obtained with the training sets from 3 iterations. 8-9 LVs, Q2 = 0.721, RMSEcv = 0.228. For all models, ANOVA p-value < 0.05. ITA: Italy; non-ITA: the other countries of origin.

Table S5. Results of the leave-10%-out cross-validation of the PLS-DA training models developed on the a) isotopic data; b) sesquiterpene fingerprinting data, to discriminate samples in three Italian regions or origin (Calabria, Sicilia and Apulia). Results are mean values (± standard deviation) obtained from three iterations.

a) Isotopic data

	n	CAL (n)	SIC (n)	APU (n)	Not assigned (n)	Correct classification (%)
Apulia	58	6.3 ± 3.1	6 ± 3.6	45 ± 3.0	0.7 ± 1.2	77.6 ± 5.2
Calabria	46	16.3 ± 3.5	4.7 ± 2.1	24 ± 2.6	1.0 ± 1.0	35.5 ± 7.6
Sicily	32	2.3 ± 1.2	27.3 ± 1.2	2.3 ± 1.2	0.0 ± 0.0	85.4 ± 3.6
Total	136					65.2 ± 4.9

Model parameters: mean values obtained with the training sets from 3 iterations. 2-3 LVs, Q2 = 0.225, RMSEcv = 0.460. For all models, ANOVA p-value < 0.05.

b) sesquiterpene fingerprinting data

	n	CAL (n)	SIC (n)	APU (n)	Not assigned (n)	Correct classification (%)
Apulia	58	0.0 ± 0.0	0.0 ± 0.0	58.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0
Calabria	46	46.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0
Sicily	32	0.0 ± 0.0	32.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0
Total	136				0.0 ± 0.0	100.0 ± 0.0

Model parameters: mean values obtained with the training sets from 3 iterations. 13-14 LVs, Q2 = 0.627, RMSEcv = 0.292. For all models, ANOVA p-value < 0.05.

Table S6. Average temperature and rainfall during 2016-2019 across the 6 countries involved in the study.

Average temperature (°C)						
	2016	2017	2018	2019	Mean	sd
ESP	14.42	14.68	14.29	14.48	14.47	0.16
GRC	14.84	14.51	15.20	15.16	14.93	0.32
ITA	13.58	13.46	14.10	13.94	13.77	0.30
POR	16.12	16.52	15.82	16.13	16.15	0.29
TUN	21.05	20.71	20.71	20.72	20.80	0.17
TUR	12.08	11.90	13.11	12.46	12.39	0.54
Precipitation (mm)						
	2016	2017	2018	2019	Mean	sd
ESP	653.70	475.55	788.38	575.70	623.33	132
GRC	675.78	696.50	730.94	799.71	725.73	54
ITA	780.69	673.37	743.45	894.17	772.92	92
POR	984.55	598.82	995.66	699.88	819.73	201
TUN	252.83	268.55	286.35	319.82	281.89	29
TUR	649.54	536.78	700.11	602.38	622.20	70

Source: https://climateknowledgeportal.worldbank.org/country

Annex 2. Supplementary material of Publication 3

Table S1. Virgin olive oil samples set (n=404) according to their cultivar, harvest year, country, geographical region and commercial class. Due to its big size a part of the table is shown as a way of example.

Samples (n)	Cultivar	Harvest Year	Country	Region	Commercial class
4	Arauco	2016-2017	Argentina	Mendoza	EVOO
61	Arbequina	2016-2017	Spain	Catalonia	EVOO
10	Argudell	2017-2018	Spain	Catalonia	
5	Empeltre	2017-2018	Spain	Catalonia	EVOO
5	Frantoio	2016-2017	Argentina	Mendoza	EVOO
4	Hojiblanca	2017-2018	Spain	Andalousia	EVOO
10	Moroccan Picholine	2017-2018	Morocco	Fès-Mèknes	EVOO
4	Coupage	2017-2018	Spain	Catalonia	VOO

EVOO: Extra virgin olive oil commercial class; VOO: Virgin olive oil commercial class.

Table S2. Full sample set, training set and validation set analysed for the 'Arbequina' *vs* non-'Arbequina' models.

	Full set	Training set	Validation set
'Arbequina'			
Argentina	4	3	1
Chile	1	1	0
Italy	1	1	0
Morocco	9	7	2
Portugal	3	2	1
Spain	160	128	32
Other			
Argentina	37	30	7
Chile	1	1	0
Italy	35	28	7
Morocco	29	23	6
Portugal	13	10	3
Spain	111	89	22
Total samples			
'Arbequina'	178	142	36
non-'Arbequina'	226	181	45
Total samples	404	323	81

Annex 3. Supplementary material of Publication 4

Table S1. Lyophilization results for the same hazelnut sample grinded by two methods and lyophilised in triplicate: cryogenic mill © (n=3) and domestic grinder (G) (n=3). Samples subjected to grinding and subsequent cryogenic milling (G+C) did not reach constant weight within 127h of lyophilisation (data not reported).

	W_s[a] (g)	Water loss[b] (g)	Water (%)	Time[c] (h)
C-1	5.01	0.141	2.8[d]	127
C-2	4.10	0.129	3.2	127
C-3	5.09	0.160	3.1	127
G-1	5.10	0.163	3.2	72
G-2	4.99	0.158	3.2	72
G-3	5.09	0.157	3.1	72

[a]: sample weight; [b]: water loss after lyophilisation, calculated at constant sample weight; [c]: lyophilisation time necessary to reach constant sample weight; [d]: complete lyophilization was not achieved within 127h.

Annex 4. Supplementary material of Publication 5

Table S1. Sample information and composition of the seven training (train) and validation (test) sets. Due to its big size a part of the table is shown as a way of example.

				Training and validation sets						
Primary ID	Harvest year	Country of origin	Cultivar	1	2	3	4	5	6	7
Isotopes Sample 1	2019	Spain	Tonda di Giffoni	train	train	train	train	train	train	train
Isotopes Sample 2	2019	Spain	Tonda di Giffoni	train	train	train	test	train	train	train
Isotopes Sample 3	2019	Spain	Tonda di Giffoni	train	train	train	train	train	train	test
Isotopes Sample 4	2019	Spain	Tonda di Giffoni	train	test	train	test	test	train	train
Isotopes Sample 5	2019	Spain	Tonda di Giffoni	train	test	train	train	train	train	test

Table S2. Mean ± standard deviation, and statistical tests (Normality test of Shapiro-Wilk, Independent-samples t-test , one-way analysis of variance (ANOVA), U Mann Whitney and Kruskal-Wallis) p-values to assess the interannual variability of the isotopic ratios for each geographical region.

		Mean ± standard deviation (‰)				**p-value**			
Variables	Year	CHL	ESP	GEO	ITA	CHL	ESP	GEO	ITA
$\delta^{18}O$	2019	24.1 ± 0.7	23.7 ± 0.7	-	24.6 ± 0.6	< 0.001*[1]	< 0.001*[4]	-	0.003*[2]
	2020	23.3 ± 0.5	24.1 ± 0.6	-	25.3 ± 0.8	< 0.001*[1]	< 0.001*[4]	-	0.003*[2]
	2021	-	23.7 ± 0.5	20.0 ± 0.4	24.4 ± 0.5	-	< 0.001*[4]	0.038*[1]	0.003*[2]
	2022	-	22.9 ± 0.7	19.7 ± 0.5	-	-	< 0.001*[4]	0.038*[1]	-
$\delta^{13}C_{palmitic}$	2019	-29.9 ± 0.7	-29.7 ± 0.4	-	-30.2 ± 0.3	0.443[1]	0.071[2]	-	< 0.001*[2]
	2020	-30.0 ± 0.6	-30.1 ± 0.7	-	-29.3 ± 0.6	0.443[1]	0.071[2]	-	< 0.001*[2]
	2021	-	-29.9 ± 0.4	-30.7 ± 0.5	-30.1 ± 0.6	-	0.071[2]	< 0.001*[1]	< 0.001*[2]
	2022	-	-29.8 ± 0.4	-32.8 ± 0.6	-	-	0.071[2]	< 0.001*[1]	-
$\delta^{13}C_{Stearic}$	2019	-30.8 ± 1.4	-31.1 ± 1.1	-	-31.7 ± 1.3	0.319[1]	< 0.001*[4]	-	0.229[2]
	2020	-31.2 ± 1.0	-30.8 ± 1.0	-	-30.9 ± 0.5	0.319[1]	< 0.001*[4]	-	0.229[2]
	2021	-	-31.1 ± 1.2	-32.6 ± 1.3	-31.4 ± 1.6	-	< 0.001*[4]	0.277[1]	0.229[2]
	2022	-	-30.0 ± 0.4	-33.0 ± 0.7	-	-	< 0.001*[4]	0.277[1]	-
$\delta^{13}C_{Oleic}$	2019	-28.2 ± 0.6	-28.2 ± 0.4	-	-28.5 ± 0.5	0.214[1]	0.182[2]	-	0.007*[4]
	2020	-28.5 ± 0.6	-28.5 ± 0.7	-	-27.8 ± 0.7	0.214[1]	0.182[2]	-	0.007*[4]
	2021	-	-28.4 ± 0.5	-29.1 ± 0.4	-28.5 ± 0.5	-	0.182[2]	< 0.001*[3]	0.007*[4]
	2022	-	-28.5 ± 0.4	-31.5 ± 0.6	-	-	0.182[2]	< 0.001*[3]	-
$\delta^{13}C_{Linoleic}$	2019	-30.4 ± 0.7	-30.1 ± 0.5	-	-30.7 ± 1.0	0.263[1]	0.003*[2]	-	0.061[2]
	2020	-30.6 ± 0.6	-30.6 ± 0.8	-	-30.0 ± 0.8	0.263[1]	0.003*[2]	-	0.061[2]
	2021	-	-30.6 ± 0.8	-31.2 ± 0.5	-30.8 ± 0.7	-	0.003*[2]	< 0.001*[1]	0.061[2]
	2022	-	-30.8 ± 0.5	-34.1 ± 0.7	-	-	0.003*[2]	< 0.001*[1]	-
		Mean ± standard deviation (‰)				**p-value**			
Variables	Year	CHL	ESP	GEO	ITA	CHL	ESP	GEO	ITA
$\delta^{2}H_{Palmitic}$	2019	-177.8 ± 7.6	-162.7 ± 7.2	-	-162.5 ± 8.2	0.248[1]	< 0.001*[4]	-	0.019*[2]
	2020	-180.3 ± 5.7	-159.1 ± 7.7	-	-165.4 ± 7.5	0.248[1]	< 0.001*[4]	-	0.019*[2]
	2021	-	-158.8 ± 4.3	-182.6 ± 7.3	-171.5 ± 6.7	-	< 0.001*[4]	< 0.001*[1]	0.019*[2]
	2022	-	-171.8 ± 5.6	-190.6 ± 3.4	-	-	< 0.001*[4]	< 0.001*[1]	-
$\delta^{2}H_{Oleic}$	2019	-189.5 ± 5.6	-173.2 ± 6.3	-	-176.8 ± 6.6	0.024*[1]	< 0.001*[4]	-	0.023*[2]
	2020	-193.2 ± 4.4	-173.4 ± 4.8	-	-177.6 ± 6.7	0.024*[1]	< 0.001*[4]	-	0.023*[2]
	2021	-	-172.8 ± 3.8	-196.0 ± 6.4	-183.7 ± 5.6	-	< 0.001*[4]	0.066[1]	0.023*[2]
	2022	-	-179.9 ± 3.0	-199.0 ± 3.3	-	-	< 0.001*[4]	0.066[1]	-
$\delta^{2}H_{Linoleic}$	2019	-214.5 ± 6.7	-200.3 ± 5.8	-	-198.1 ± 8.2	0.376[1]	< 0.001*[4]	-	0.151[2]
	2020	-216.1 ± 4.9	-196.7 ± 4.3	-	-200.0 ± 8.1	0.376[1]	< 0.001*[4]	-	0.151[2]
	2021	-	-196.4 ± 7.6	-219.5 ± 6.0	-204.1 ± 6.0	-	< 0.001*[4]	0.941[1]	0.151[2]
	2022	-	-204.7 ± 5.2	-219.4 ± 4.0	-	-	< 0.001*[4]	0.941[1]	-

CHL: Chile; ESP: Spain; GEO: Georgia; ITA: Italy. * p-value < 0.05 means are significantly different. [1] p-value obtained from independent samples t-test. [2] p-value obtained from one-way analysis of variance (ANOVA). [3] p-value obtained from U Mann Whitney test. [4] p-value obtained from Kruskal-Wallis test.

Table S3. Results of the leave-10%-out cross-validation of the PLS-DA training models developed on the $\delta^{18}O$, and $\delta^{2}H$, $\delta^{13}C$ of the main FAMEs data to discriminate samples in three regions of origin. Results are mean values (± standard deviation) obtained from seven iterations.

	n	CHL (n)	ESP (n)	GEO (n)	ITA (n)	Not assigned (n)	Correct classification (%)
CHL	32	21.1 ± 1.6	1.3 ± 0.5	0.0 ± 0.0	0.0 ± 0.0	9.6 ± 1.6	66.1 ± 4.9
ESP	73	0.3 ± 0.5	63.4 ± 1.7	0.0 ± 0.0	0.0 ± 0.0	9.3 ± 1.6	86.9 ± 2.4
GEO	32	0.0 ± 0.0	0.0 ± 0.0	32.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0
ITA	29	2.1 ± 1.1	12.6 ± 1.0	0.0 ± 0.0	5.7 ± 2.8	8.6 ± 1.8	19.7 ± 9.7
Total	166						73.7 ± 2.4

N = 166, 5 LVs, Q2 = 0.495, RMSEcv = 0.353. For all models, ANOVA p-value < 0.05. CHL: Chile; ESP: Spain; GEO: Georgia; ITA: Italy.

Table S4. Thresholds of the PLSDA models (7 training sets) for each class optimized with ROC curves.

Set	CHL	ESP + ITA	GEO
Set 1	0.323	0.555	0.534
Set 2	0.385	0.520	0.524
Set 3	0.365	0.541	0.542
Set 4	0.382	0.567	0.516
Set 5	0.379	0.525	0.546
Set 6	0.353	0.518	0.544
Set 7	0.350	0.501	0.512
Mean	0.362	0.532	0.531
SD	0.020	0.021	0.013

Table S5. Results of the leave-10%-out cross-validation of the PLS-DA training models developed on the $\delta^{18}O$, and $\delta^{2}H$, $\delta^{13}C$ of the main FAMEs data to discriminate samples in three regions of origin. Results are mean values (± standard deviation) obtained from seven iterations.

	n	CHL (n)	ESP + ITA (n)	GEO (n)	Not assigned (n)	Correct classification (%)
CHL	32	28.0 ± 1.2	3.4 ± 1.3	0.0 ± 0.0	0.6 ± 0.5	87.5 ± 3.6
ESP + ITA	102	1.3 ± 1.4	99.1 ± 1.6	0.0 ± 0.0	1.6 ± 1.4	97.2 ± 1.5
GEO	32	0.0 ± 0.0	0.0 ± 0.0	32.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0
Total	166					95.9 ± 0.9

N = 166, 6 LVs, Q2 = 0.648, RMSEcv = 0.296. For all models, ANOVA p-value < 0.05. CHL: Chile; ESP: Spain; GEO: Georgia; ITA: Italy.

Table S6. Misclassified and not assigned samples information of each test set.

Set	Misclassified samples	Harvest year	Country of origin
Set 1	AV 86 2020	2020	Italy
Set 2	AV 58 2019	2019	Chile
	AV 69 2020	2020	Chile
	AV 81 2021	2021	Italy
Set 3	AV 65 2019	2019	Chile
	AV 86 2021	2021	Italy
Set 4	AV 56 2019	2019	Chile
	AV 68 2019	2019	Chile
	AV 61 2020	2020	Chile
	AV 76 2021	2021	Italy
Set 5	AV 60 2019	2019	Chile
	AV 70 2020	2020	Chile
	AV 46 2019	2019	Spain
	AV 75 2020	2020	Italy
	AV 34 2022	2022	Spain
Set 6	-	-	-
Set 7	AV 56 2019	2019	Chile
	AV 70 2020	2020	Chile
	AV 83 2020	2020	Italy

Annex 5. Supplementary material of Publication 6

Table S1. Sample set data: sample name, genuine sample, type of adulterant, percentage of adulteration and category of sample. Due to its big size a part of the table is shown as a way of example.

N	Sample name	Genuine	Adulterant	% Adulterant	Category
1	AUT16_001	AUT001	-	0	GENUINE
2	001_2F6	AUT001	SFO6	2	HL
3	001_5S5	AUT001	SBO5	5	HL
4	001_5O5	AUT001	HOSFO5	5	HO
5	001_10H5	AUT001	HZO5	10	HO
6	AUT16_012	AUT012	-	0	GENUINE

HL: High linoleic; HO: High oleic; HZO: Hazelnut oil; SFO: Sunflower oil; HOSFO: High oleic sunflower oil; SBO: Soybean oil.

Table S2. The starting COW parameters applied in HPLC-RID and HT-GC-MS for data alignment in Matlab are: The optim_space (1 x 4) [segment minimum, maximum and slack minimum, maximum]. The options (1 x 4) [trigger plot and progress text, number of optimizations from grid maxima, maximum number of optimization steps, fraction of maximal deviation from centre in COW alignment]. The options [1 3 50 0.15] (plot; 3 starts from 3 maxima in grid search; maximum 50 steps; 15%) were the same for all the alignments.

HT-GC-MS	optim_space	options
237	[5 70 2 20]	[1 3 50 0.15]
239	[5 90 2 20]	[1 3 50 0.15]
260	[5 90 2 20]	[1 3 50 0.15]
262	[5 70 2 20]	[1 3 50 0.15]
264	[5 90 2 20]	[1 3 50 0.15]
265	[5 90 2 20]	[1 3 50 0.15]
311	[5 90 2 20]	[1 3 50 0.15]
313	[5 90 2 20]	[1 3 50 0.15]
335	[5 110 2 20]	[1 3 50 0.15]
337	[5 110 2 20]	[1 3 50 0.15]
339	[5 70 2 20]	[1 3 50 0.15]
HPLC-RID	[5 90 2 20]	[1 3 50 0.15]

Table S3. External validation results of PLS-DA models obtained by a) FIA-HESI-HRMS, b) HTGC–MS, c) HPLC-RID.

a) FIA-HESI-HRMS

FIA-HESI-HRMS						
SET1	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	3	3	50		0.5
Adulterant	23	0	23	100	1	
HL	12	0	12	100	1	
HO	11	0	11	100	1	
SET2	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	4	2	66.67		0.67
Adulterant	23	4	19	82.61	0.83	
HL	12	2	10	83.33	0.83	
HO	11	2	9	81.82	0.82	
SET3	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	3	3	50		0.5
Adulterant	23	1	22	95.65	0.96	
HL	12	1	11	91.67	0.92	
HO	11	0	11	100	1	
SET4	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	3	3	50		0.5
Adulterant	23	5	18	78.26	0.78	
HL	12	1	11	91.67	0.92	
HO	11	4	7	63.64	0.64	
SET5	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	2	4	33.33		0.33
Adulterant	23	2	21	91.3	0.91	
HL	12	1	11	91.67	0.92	
HO	11	1	10	90.91	0.91	
SET6	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	4	2	66.67		0.67
Adulterant	23	3	20	86.96	0.87	
HL	12	2	10	83.33	0.83	
HO	11	1	10	90.91	0.91	
SET7	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	2	4	33.33		0.33
Adulterant	23	2	21	91.3	0.91	
HL	12	1	11	91.67	0.92	
HO	11	1	10	90.91	0.91	

b) HTGC–MS

HTGC–MS						
SET1	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	5	1	83.33		0.83
Adulterant	23	2	21	91.3	0.91	
HL	12	0	12	100	1	
HO	11	2	9	81.82	0.82	
SET2	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	6	0	100		1
Adulterant	23	1	22	95.65	0.96	
HL	12	0	12	100	1	
HO	11	1	10	90.91	0.91	
SET3	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	4	2	66.67		0.67
Adulterant	23	2	21	91.3	0.91	
HL	12	0	12	100	1	
HO	11	2	9	81.82	0.82	
SET4	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	4	2	66.67		0.67
Adulterant	23	1	22	95.65	0.96	
HL	12	0	12	100	1	
HO	11	1	10	90.91	0.91	
SET5	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	4	2	66.67		0.67
Adulterant	23	1	22	95.65	0.96	
HL	12	0	12	100	1	
HO	11	1	10	90.91	0.91	
SET6	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	4	2	66.67		0.67
Adulterant	23	1	22	95.65	0.96	
HL	12	0	12	100	1	
HO	11	1	10	90.91	0.91	
SET7	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	5	1	83.33		0.83
Adulterant	23	2	21	91.3	0.91	
HL	12	1	11	91.67	0.92	
HO	11	1	10	90.91	0.91	

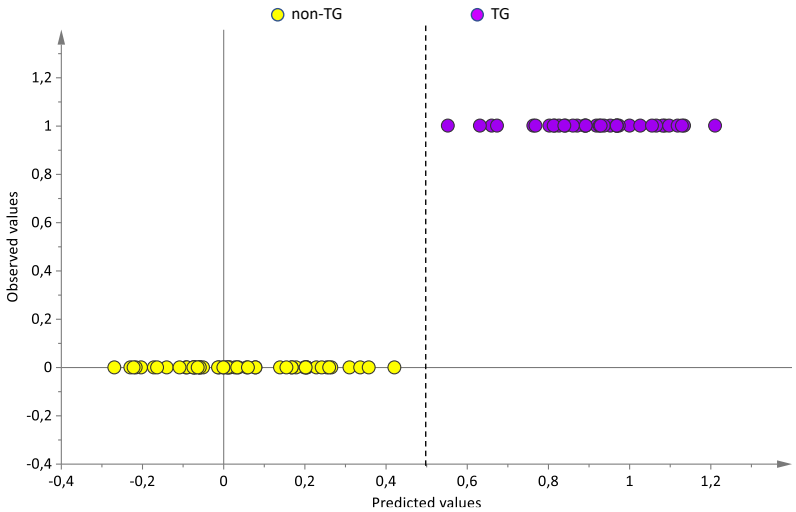
c) HPLC-RID

HPLC-RID						
SET1	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	4	2	66.67		0.67
Adulterant	23	2	21	91.30	0.91	
HL	12	0	12	100.00	1	
HO	11	2	9	81.82	0.82	
SET2	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	4	2	66.67		0.67
Adulterant	23	2	21	91.30	0.91	
HL	12	1	11	91.67	0.92	
HO	11	1	10	90.91	0.91	
SET3	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	3	3	50.00		0.5
Adulterant	23	2	21	91.30	0.91	
HL	12	0	12	100.00	1	
HO	11	2	9	81.82	0.82	
SET4	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	4	2	66.67		0.67
Adulterant	23	2	21	91.30	0.91	
HL	12	0	12	100.00	1	
HO	11	2	9	81.82	0.82	
SET5	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	6	3	3	50.00	
Adulterant	23	23	4	19	82.61	0.83
HL	12	12	1	11	91.67	0.92
HO	11	11	3	8	72.73	0.73
SET6	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	6	4	2	66.67	
Adulterant	23	23	3	20	86.96	0.87
HL	12	12	0	12	100.00	1
HO	11	11	3	8	72.73	0.73
SET7	n	Genuine	Adulterant	Correct classification (%)	Sensitivity	Specificity
Genuine	6	6	3	3	50.00	
Adulterant	23	23	2	21	91.30	0.91
HL	12	12	1	11	91.67	0.92
HO	11	11	1	10	90.91	0.91

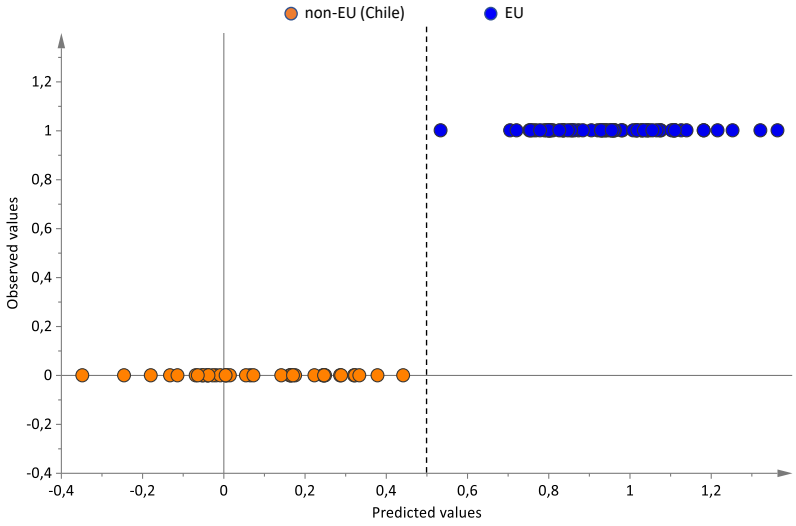
HL: High linoleic; HO: High oleic.

Annex 6. Supplementary material of Publication 5

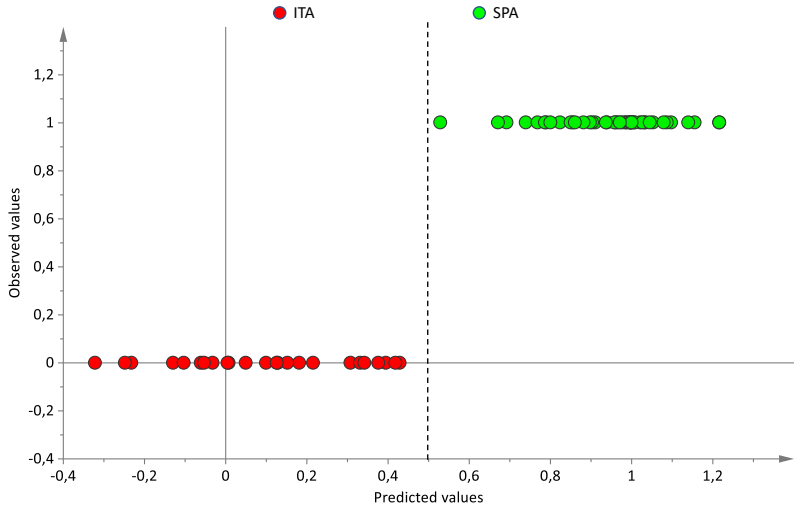
a)



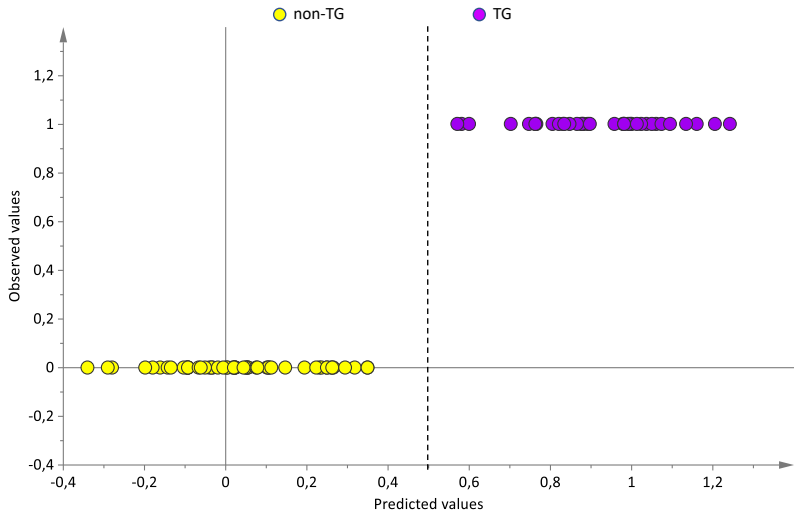
b)



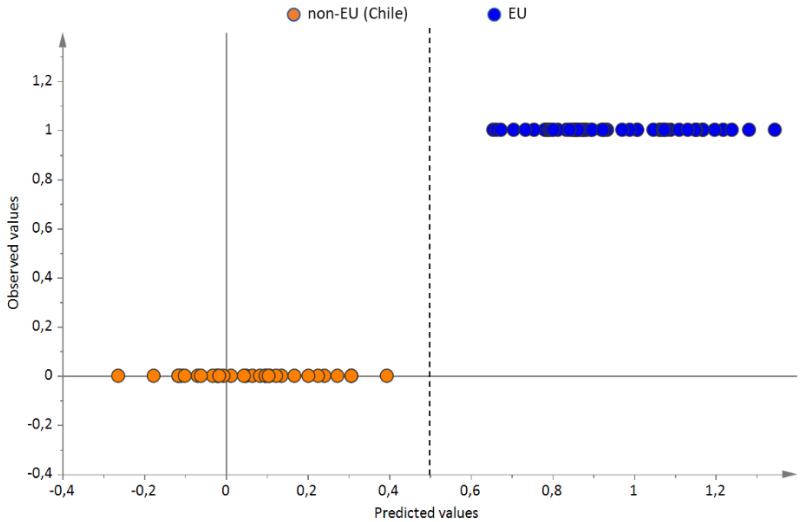
c)


ITA
SPA
Observed values
Predicted values
1,2
1
0,8
0,6
0,4
0,2
0
-0,2
-0,4
-0,4
-0,2
0
0,2
0,4
0,6
0,8
1
1,2

d)


non-TG
TG
Observed values
Predicted values
1,2
1
0,8
0,6
0,4
0,2
0
-0,2
-0,4
-0,4
-0,2
0
0,2
0,4
0,6
0,8
1
1,2

e)


non-EU (Chile)
EU
Observed values
Predicted values
1,2
1
0,8
0,6
0,4
0,2
0
-0,2
-0,4
-0,4
-0,2
0
0,2
0,4
0,6
0,8
1
1,2

f)

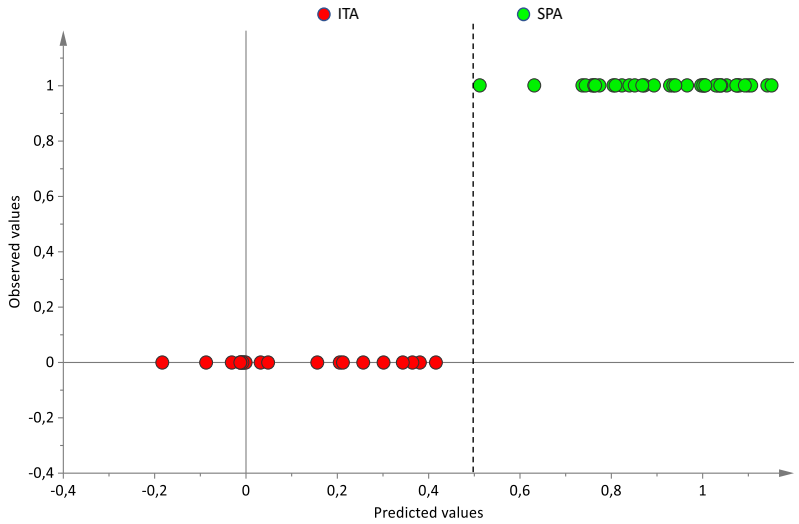


Fig. S1. Observed vs. predicted values of global calibration sets with a classification threshold of 0.5: a) fingerprinting cultivar model: "Tonda di Giffoni (TG) vs. other cultivars (non-TG)"; b) fingerprinting European model: "European (EU) vs. non-European (non-EU)"; c) fingerprinting Spain-Italy model: "Spain (SPA) vs. Italy (ITA)"; d) untargeted profiling cultivar model: "Tonda di Giffoni (TG) vs. other cultivars (non-TG)"; e) untargeted profiling European model: "European (EU) vs. non-European (non-EU)"; f) untargeted profiling Spain-Italy model: "Spain (SPA) vs. Italy (ITA)".

Table S1. Sample information: sample ID (Primary ID), harvest year (Year), country of origin (Country) and hazelnut cultivar (Cultivar). Due to its big size a part of the table is shown as a way of example.

Primary ID	Year	Country	Cultivar
AV_43	2019	Spain	Tonda di Giffoni
AV_7	2019	Spain	Castanyera
AV_12	2019	Spain	Gironell
AV_2	2019	Spain	Negret
AV_79	2019	Italy	Tonda di Giffoni
AV_66	2019	Chile	Tonda di Giffoni

Table S2. Peak areas table of the untargeted profiling approach using PARADISe: Compound ID, Interval ID, Model order (number of optimal components of each PARAFAC2 model), Component number (component from the PARAFAC2 model that correspond to a chemical compound), estimated retention time (min) of each peak. Due to its big size a part of the table is shown as a way of example.

Compound ID	Interval ID	Model Order	Component number	Est. Retention Time (min)	AV_43	AV_7	AV_35
1	1	2	2	14.96551667	57065.42	35530.64	48933.72
2	1	2	1	14.97588333	1491436	898377.9	1500205
3	2	4	1	15.04333333	69209.73	54299.31	74104.91
4	3	3	1	15.37021667	344210.6	255338.9	351617.7
5	4	3	2	15.44286667	0	6506.022	7489.115

Table S3. Intervals information of the untargeted profiling approach using PARADISe: interval ID, number of components of each PARAFAC2 model for every interval (Model Order), number of components of each PARAFAC2 model that correspond to actual chemical species (Chemical Components), initial and final retention times of each interval (Start and End RT). Due to its big size a part of the table is shown as a way of example.

Interval ID	Model Order	Chemical Components	Start RT (min)	End RT (min)
1	2	2	14.93438333	15.01221667
2	4	1	15.0174	15.07446667
3	3	1	15.3391	15.40135
4	3	2	15.39616667	15.44286667
5	4	2	15.43768333	15.5155

Table S4. Model parameters obtained with the training sets from 7 iterations. TG: 'Tonda di Giffoni'; non-TG: other cultivars; EU: European (Spanish and Italian); non-EU: non-European (Chilean); ESP: Spanish hazelnuts; ITA: Italian hazelnuts.

FINGERPRINTING TG/non-TG	LV	RY^2	Q^2	Q^2 permutation	RMSEcv
1	7	0.906	0.674	-0.21	0.28775
2	7	0.895	0.729	-0.289	0.25274
3	7	0.884	0.593	-0.17	0.33305
4	7	0.906	0.57	-0.208	0.27045
5	7	0.883	0.662	-0.176	0.27152
6	7	0.895	0.61	-0.179	0.32273
7	7	0.887	0.651	-0.223	0.36844
mean	7.0	0.89	0.641	-0.21	0.30
standard deviation	0.0	0.01	0.05	0.04	0.04

PARADISe TG/non-TG	LV	RY^2	Q^2	Q^2 permutation	RMSEcv
1	7	0.889	0.402	-0.688	0.35072
2	7	0.886	0.427	-0.597	0.32155
3	5	0.806	0.405	-0.445	0.37438
4	6	0.839	0.454	-0.464	0.342
5	7	0.887	0.306	-0.628	0.37952
6	5	0.821	0.482	-0.36	0.35183
7	7	0.879	0.35	-0.755	0.36762
mean	6.3	0.86	0.404	-0.56	0.36
standard deviation	1.0	0.04	0.06	0.14	0.02
FINGERPRINTING EU/non-EU		RY^2	Q^2	Q^2 permutation	RMSEcv
1	7	0.915	0.731	-0.302	0.25657
2	6	0.89	0.693	-0.271	0.2816
3	7	0.913	0.722	-0.268	0.2478
4	7	0.923	0.734	-0.225	0.25605
5	7	0.931	0.73	-0.317	0.31507
6	7	0.914	0.718	-0.287	0.26705
7	6	0.891	0.687	-0.155	0.37815
mean	6.7	0.91	0.716	-0.26	0.29
standard deviation	0.5	0.02	0.02	0.06	0.05
PARADISe EU/non-EU		RY^2	Q^2	Q^2 permutation	RMSEcv
1	4	0.743	0.568	-0.293	0.30439
2	5	0.847	0.531	-0.365	0.31103
3	6	0.874	0.62	-0.57	0.26892
4	6	0.864	0.617	-0.528	0.26808
5	6	0.881	0.64	-0.538	0.25679
6	5	0.826	0.638	-0.393	0.28838
7	5	0.832	0.589	-0.395	0.30469
mean	5.3	0.84	0.600	-0.44	0.29
standard deviation	0.8	0.05	0.04	0.10	0.02
FINGERPRINTING Spanish/Italian		RY^2	Q^2	Q^2 permutation	RMSEcv
1	7	0.959	0.756	-0.33	0.20029
2	5	0.853	0.578	-0.2	0.34064
3	6	0.908	0.518	-0.153	0.32182
4	7	0.94	0.648	-0.22	0.24936
5	5	0.9	0.588	-0.0791	0.32918
6	3	0.735	0.559	-0.165	0.33397
7	5	0.88	0.689	-0.155	0.36187
mean	5.4	0.88	0.619	-0.19	0.31
standard deviation	1.4	0.07	0.08	0.08	0.06
PARADISe Spanish/Italian		RY^2	Q^2	Q^2 permutation	RMSEcv
1	5	0.911	0.726	-0.404	0.239
2	4	0.855	0.665	-0.325	0.31245
3	5	0.886	0.655	-0.445	0.29926
4	4	0.857	0.691	-0.368	0.25699
5	6	0.931	0.533	-0.617	0.32702
6	4	0.866	0.703	-0.329	0.26037
7	5	0.916	0.683	-0.425	0.26103
mean	4.7	0.89	0.665	-0.42	0.28
standard deviation	0.8	0.03	0.06	0.10	0.03

Annex 7. Supplementary material of Publication 8

Table S1. Sample information and composition of the seven training and validation sets (train1 and train12 are used to build the cultivar training models which are validated with the test1 and test12 samples; train2 and train12 are used to build the origin training models which are validated with the test2 and test12 samples). Due to its big size a part of the table is shown as a way of example.

				Training and validation sets						
Sample ID	Harvest year	Country of origin	Cultivar	1	2	3	4	5	6	7
Sample 1	2019	Spain	non-TG1	train1	train1	train1	train1	train1	train1	test1
Sample 2	2019	Spain	non-TG1	train1	train1	train1	train1	test1	test1	train1
Sample 3	2019	Spain	non-TG1	train1	train1	train1	train1	train1	train1	train1
Sample 4	2019	Spain	non-TG1	train1	test1	train1	train1	test1	train1	train1
Sample 5	2019	Spain	non-TG1	train1	train1	test1	train1	train1	train1	test1

Table S2. PLS-DA cultivar models developed on each training and validation set (7 iterations) of unsaponifiable fraction (UF) and triacylglycerol (TAG) fingerprinting data.

Cultivar model: TG/non-TG								
Unsaponifiable fraction fingerprinting								
	Calibration and cross-validation							**External Validation**
	LV	R2X(%)[b]	R2Y(%)[c]	Q^2	RMSEcv	ANOVA cv residuals	% Correct classification	% Correct classification
Iteration 1	8	73.5	94.3	0.796	0.224	<0.05	100	94.9
Iteration 2	8	73.9	94.4	0.818	0.223	<0.05	100	97.4
Iteration 3	8	72.9	94.2	0.843	0.201	<0.05	100	94.9
Iteration 4	8	73.1	94.0	0.811	0.205	<0.05	100	100
Iteration 5	8	74.1	94.2	0.824	0.208	<0.05	100	87.2
Iteration 6	8	72.0	94.3	0.803	0.221	<0.05	100	92.3
Iteration 7	8	73.5	94.7	0.834	0.202	<0.05	100	92.3
Average[a]	8	73.3	94.3	0.82	0.21	<0.05	100.0 ± 0.0	94.1 ± 4.1
TAG fingerprinting								
	Calibration and cross-validation							**External Validation**
	LV	R2X(%)[b]	R2Y(%)[c]	Q^2	RMSEcv	ANOVA cv residuals	% Correct classification	% Correct classification
Iteration 1	10	99.6	90.6	0.445	0.370	<0.05	100	82.1
Iteration 2	10	99.6	90.1	0.354	0.390	<0.05	100	92.3
Iteration 3	9	99.6	89.7	0.549	0.335	<0.05	100	76.9
Iteration 4	10	99.6	90.0	0.327	0.380	<0.05	100	82.1
Iteration 5	9	99.6	89.4	0.568	0.335	<0.05	100	82.1
Iteration 6	9	99.6	86.3	0.346	0.4	<0.05	100	84.6
Iteration 7	9	99.6	86.9	0.447	0.375	<0.05	100	94.9
Average[a]	9	99.6	89.0	0.43	0.37	<0.05	100.0 ± 0.0	85.0 ± 6.4

[a] Average: mean values obtained with the training sets from 7 iterations. [b] Fraction of X variation modelled by the selected number of latent variables. [c] Cumulative prediction fraction, according to cross-validation, of the variation of Y by the selected number of latent variables. TG: 'Tonda di Giffoni'; non-TG: other cultivars.

Table S3. PLS-DA origin models developed on each training and validation set (7 iterations) of unsaponifiable fraction (UF) and triacylglycerol (TAG) fingerprinting data.

Cultivar model: TG/non-TG								
Unsaponifiable fraction fingerprinting								
	Calibration and cross-validation							**External Validation**
	LV	R2X(%)[b]	R2Y(%)[c]	Q^2	RMSEcv	ANOVA cv residuals	% Correct classification	% Correct classification
Iteration 1	11	77.6	90.8	0.790	0.262	<0.05	100	95.1
Iteration 2	12	77.9	92.6	0.792	0.261	<0.05	100	100.0
Iteration 3	13	78.3	93.5	0.771	0.336	<0.05	100	95.1
Iteration 4	12	77.3	92.7	0.785	0.286	<0.05	100	95.1
Iteration 5	12	77.9	92.5	0.761	0.268	<0.05	100	100.0
Iteration 6	13	78.1	93.6	0.766	0.256	<0.05	100	100.0
Iteration 7	12	77.6	92.0	0.791	0.279	<0.05	100	95.1
Average[a]	12	77.8	92.5	0.78	0.28	<0.05	100.0 ± 0.0	97.2 ± 2.6
TAG fingerprinting								
	Calibration and cross-validation							**External Validation**
	LV	R2X(%)[b]	R2Y(%)[c]	Q^2	RMSEcv	ANOVA cv residuals	% Correct classification	% Correct classification
Iteration 1	11	99.6	79.6	0.508	0.293	<0.05	98.8	65.9
Iteration 2	11	99.6	78.3	0.348	0.343	<0.05	99.4	90.2
Iteration 3	11	99.6	77.9	0.406	0.329	<0.05	99.4	82.9
Iteration 4	12	99.6	80.1	0.425	0.325	<0.05	99.4	75.6
Iteration 5	11	99.6	79.4	0.399	0.330	<0.05	100	90.2
Iteration 6	12	99.6	81.1	0.430	0.317	<0.05	100	87.8
Iteration 7	12	99.6	80.3	0.361	0.340	<0.05	99.4	82.9
Average[a]	11	99.6	79.50	0.41	0.33	<0.05	99.5 ± 0.4	82.2 ± 8.9

[a] Average: mean values obtained with the training sets from 7 iterations. [b] Fraction of X variation modelled by the selected number of latent variables. [c] Cumulative prediction fraction, according to cross-validation, of the variation of Y by the selected number of latent variables. CHL: Chile; ESP: Spain; GEO: Georgia; ITA: Italy.

Table S4. Uncertainty ranges and optimal thresholds obtained from receiver operating characteristic (ROC) curves for each method and subset.

Unsaponifiable fraction fingerprinting							
Set	Threshold	Cultivar	CHL	ESP	GEO	ITA	Model
1	Optimal[a]	0.452	0.519	0.513	0.507	0.426	One analysis approach
2	Optimal[a]	0.465	0.458	0.399	0.462	0.585	One analysis approach
3	Optimal[a]	0.438	0.533	0.461	0.479	0.603	One analysis approach
4	Optimal[a]	0.439	0.492	0.458	0.434	0.529	One analysis approach
5	Optimal[a]	0.507	0.495	0.436	0.513	0.554	One analysis approach
6	Optimal[a]	0.396	0.522	0.423	0.495	0.533	One analysis approach
7	Optimal[a]	0.485	0.500	0.429	0.424	0.580	One analysis approach
Unsaponifiable fraction fingerprinting							
Set	Threshold	Cultivar	CHL	ESP	GEO	ITA	Model
1	Optimal[a]	0.489	0.496	0.566	0.454	0.430	One analysis approach
	Lower	0.339	0.346	0.356[b]	0.304	0.343[b]	Uncertainty range in combined approach
	Upper	0.639	0.646	0.566[c]	0.604	0.520[c]	
2	Optimal[a]	0.517	0.441	0.476	0.497	0.404	One analysis approach
	Lower	0.367	0.291	0.326	0.347	0.311[b]	Uncertainty range in combined approach
	Upper	0.667	0.591	0.626	0.647	0.464[c]	
3	Optimal[a]	0.527	0.456	0.463	0.497	0.397	One analysis approach
	Lower	0.377	0.306	0.343[b]	0.347	0.397[b]	Uncertainty range in combined approach
	Upper	0.677	0.606	0.554[c]	0.647	0.460[c]	
4	Optimal[a]	0.568	0.546	0.504	0.470	0.387	One analysis approach
	Lower	0.418	0.396	0.504[b]	0.320	0.387[b]	Uncertainty range in combined approach
	Upper	0.718	0.696	0.647[c]	0.620	0.537[c]	
5	Optimal[a]	0.463	0.499	0.509	0.410	0.410	One analysis approach
	Lower	0.313	0.349	0.359	0.410[b]	0.410[b]	Uncertainty range in combined approach
	Upper	0.613	0.649	0.659	0.544[c]	0.567[c]	
6	Optimal[a]	0.432	0.540	0.508	0.494	0.375	One analysis approach
	Lower	0.282	0.390	0.358	0.344	0.375[b]	Uncertainty range in combined approach
	Upper	0.582	0.690	0.658	0.644	0.586[c]	
7	Optimal[a]	0.429	0.487	0.482	0.495	0.461	One analysis approach
	Lower	0.279	0.337	0.332	0.345	0.326[b]	Uncertainty range in combined approach
	Upper	0.579	0.637	0.632	0.645	0.461[c]	

[a] Optimal thresholds obtained from the ROC curves maximizing the value of (sensitivity + specificity). [b] Lower thresholds obtained from the ROC curves leading to maximum sensitivity. [c] Higher thresholds obtained from the ROC curves leading to maximum specificity. The upper and lower thresholds were set to ± 0.15 of the optimal thresholds (Quintanilla-Casas et al., 2020a). CHL: Chile; ESP: Spain; GEO: Georgia; ITA: Italy.

Table S5. Leave 10% out cross validation of PLS-DA cultivar models developed on unsaponifiable fraction (UF), triacylglycerol (TAG) fingerprinting data and the combined approach. Results are mean values (± standard deviation) obtained from seven iterations.

Cultivar model: TG/non-TG						
Unsaponifiable fraction fingerprinting (LVs = 8, Q^2 = 0.82, RMSEcv = 0.21)[a]						
	n	non-TG (n)	TG (n)	Correct classification (%)	Sensitivity	Specificity
non-TG	81	81	0	100 ± 0	1 ± 0	
TG	73	0	73	100 ± 0		1 ± 0
Total	154			100 ± 0		
TAG fingerprinting (LVs = 9-10, Q^2 = 0.43, RMSEcv = 0.37)[a]						
	n	non-TG (n)	TG (n)	Correct classification (%)	Sensitivity	Specificity
non-TG	81	81	0	100 ± 0	1 ± 0	
TG	73	0	73	100 ± 0		1 ± 0
Total	154			100 ± 0		
Combined strategy (TAG + UF fingerprinting)						
	n	non-TG (n)	TG (n)	Correct classification (%)	Sensitivity	Specificity
non-TG	81	81	0	100 ± 0	1 ± 0	
TG	73	0	73	100 ± 0		1 ± 0
Total	154			100 ± 0		

For all models, ANOVA p-value < 0.05. [a] Model parameters: mean values obtained with the training sets from 7 iterations. TG: 'Tonda di Giffoni'; non-TG: other cultivars

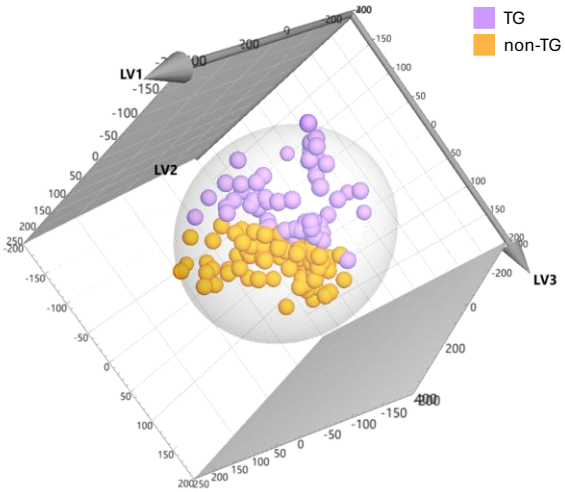
Table S6. Leave 10% out cross validation of PLS-DA origin models developed on unsaponifiable fraction (UF), triacylglycerol (TAG) fingerprinting data, and the combined approach. Results are mean values (± standard deviation) obtained from seven iterations.

Geographical origin model: CHL/ESP/GEO/ITA							
Unsaponifiable fraction fingerprinting (LVs = 11-13, Q^2 = 0.78, RMSEcv = 0.28)[a]							
	n	CHL (n)	ESP (n)	GEO (n)	ITA (n)	Not assigned	Correct classification (%)
CHL	32	32	0	0	0	0	100.0 ± 0.0
ESP	73	0	73	0	0	0	100.0 ± 0.0
GEO	32	0	0	32	0	0	100.0 ± 0.0
ITA	29	0	0	0	29	0	100.0 ± 0.0
Total	166						100.0 ± 0.0
TAG fingerprinting (LVs = 11-12, Q^2 = 0.41, RMSEcv = 0.33)[a]							
	n	CHL (n)	ESP (n)	GEO (n)	ITA (n)	Not assigned	Correct classification (%)
CHL	32	32	0	0	0	0	100.0 ± 0.0
ESP	73	0	72.7 ± 0.5	0	0	0.3 ± 0.5	99.6 ± 0.7
GEO	32	0	0	32	0	0	100.0 ± 0.0
ITA	29	0	0	0	28.4 ± 0.5	0.4 ± 0.5	98.0 ± 1.8
Total	166						99.5 ± 0.4

Combined strategy (TAG +UF fingerprinting)							
	n	CHL (n)	ESP (n)	GEO (n)	ITA (n)	Not assigned	Correct classification (%)
CHL	32	32	0	0	0	0	100.0 ± 0.0
ESP	73	0	73	0	0	0	100.0 ± 0.0
GEO	32	0	0	32	0	0	100.0 ± 0.0
ITA	29	0	0	0	29	0	100.0 ± 0.0
Total	166						100.0 ± 0.0

For all models, ANOVA p-value < 0.05. [a] Model parameters: mean values obtained with the training sets from 7 iterations. CHL: Chile; ESP: Spain; GEO: Georgia; ITA: Italy.

a)



b)

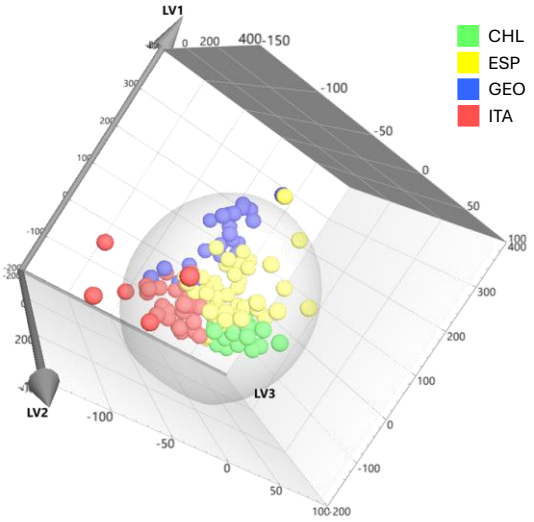
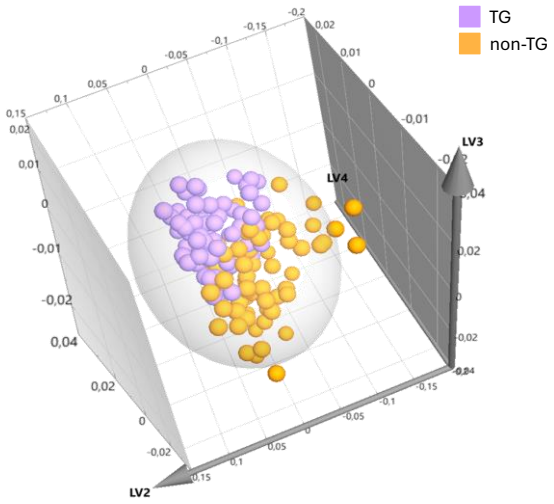


Fig. S1. Score plots of the PLS-DA models developed using the UF fingerprinting data: a) cultivar model plotting the first, second and third latent variables (LV1, R^2Y= 14.3%; LV2, R^2Y= 15.8%; LV3, R^2Y= 25.3%); b) origin model plotting the first, second and third latent variables (LV1, R^2Y= 9.5%; LV2, R^2Y= 7.9%; LV3, R^2Y= 18.9%). TG: 'Tonda di Giffoni'; non-TG: other cultivars; CHL: Chile; ESP: Spain; GEO: Georgia; ITA: Italy.

a)



b)

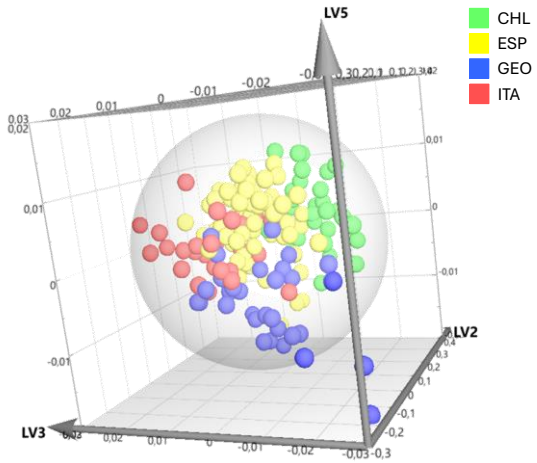


Fig. S2. Score plots of the PLS-DA models developed using the TAG fingerprinting data: a) cultivar model plotting the second, third and fourth latent variables (LV2, R^2Y= 12.6%; LV3, R^2Y= 25.0%; LV4, R^2Y= 13.9%); b) origin model plotting the second, third and fifth latent variables (LV2, R^2Y= 10.5%; LV3, R^2Y= 10.5%; LV5, R^2Y= 11.4%). TG: 'Tonda di Giffoni'; non-TG: other cultivars; CHL: Chile; ESP: Spain; GEO: Georgia; ITA: Italy.

Annex 8. Supplementary material of Publication 9

Table S1. Sample information and composition of the seven training and validation sets (train1 and train12 are used to build the cultivar training models which are validated with the test1 and test12 samples; train2 and train12 are used to build the origin training models which are validated with the test2 and test12 samples). Due to its big size a part of the table is shown as a way of example.

				Training and validation sets						
Primary ID	Harvest Year	Country of Origin	Cultivar	1	2	3	4	5	6	7
Sample 1	2019	Spain	non-TG1	train1	train1	train1	train1	train1	train1	test1
Sample 2	2019	Spain	non-TG2	train1	train1	train1	train1	train1	train1	test1
Sample 3	2019	Spain	non-TG2	train1	train1	train1	train1	test1	train1	train1
Sample 4	2019	Spain	non-TG3	train1	train1	train1	test1	train1	train1	train1
Sample 5	2019	Spain	non-TG3	train1	train1	test1	train1	train1	train1	test1

Table S2. Leave 10% out cross validation of PLS-DA cultivar models developed on whole kernels analysed by benchtop NIR (NIR) and handheld NIR (hNIR) spectrometers. Results are mean values (± standard deviation) obtained from seven iterations.

Whole kernel - Cultivar model: TG/non-TG						
NIR spectrometer (LVs = 9, Q^2 = 0.47, RMSEcv = 0.38)[a]						
	n	non-TG (n)	TG (n)	Correct classification (%)	Sensitivity	Specificity
non-TG	81	78 ± 2	3 ± 2	96.5 ± 1.9	0.96 ± 0.02	
TG	73	3 ± 1	70 ± 1	95.5 ± 0.7		0.95 ± 0.01
Total	154			96.0 ± 1.1		
hNIR spectrometer (LVs = 6, Q^2 = 0.35, RMSEcv = 0.41)[a]						
	n	non-TG (n)	TG (n)	Correct classification (%)	Sensitivity	Specificity
non-TG	81	72 ± 3	9 ± 3	88.4 ± 4.1	0.88 ± 0.04	
TG	73	10 ± 3	63 ± 3	85.9 ± 3.9		0.86 ± 0.04
Total	154			87.2 ± 3.6		

For all models, ANOVA p-value < 0.05. [a] Model parameters: mean values obtained with the training sets from 7 iterations. LVs: latent variables of the training model; Q^2: Cumulative fraction of Y variation predicted by the X training model up to the specified latent variable, according to cross-validation; RMSEcv: root mean square error of the cross validation of the training model. TG: 'Tonda di Giffoni'; non-TG: other cultivars.

Table S3. Leave 10% out cross validation of PLS-DA origin models developed on whole kernels analysed by benchtop NIR (NIR) and handheld NIR (hNIR) spectrometers. Results are mean values (± standard deviation) obtained from seven iterations.

Whole kernel - Geographical origin model: CHL/ESP/GEO/ITA							
NIR spectrometer (LVs = 14, Q^2 = 0.27, RMSEcv = 0.36)[a]							
	n	CHL (n)	ESP (n)	GEO (n)	ITA (n)	Not assigned (n)	Correct classification (%)
CHL	32	27.1 ± 1.6	0.4 ± 0.5	0.0 ± 0.0	0.0 ± 0.0	4.4 ± 1.4	84.8 ± 4.9
ESP	73	0.3 ± 0.5	69.7 ± 1.4	0.0 ± 0.0	0.0 ± 0.0	3.0 ± 1.4	95.5 ± 1.9
GEO	32	0.3 ± 0.5	0.9 ± 0.5	21.1 ± 3.7	0.0 ± 0.0	4.7 ± 3.7	78.3 ± 13.6
ITA	29	0.0 ± 0.0	3.4 ± 1.0	0.0 ± 0.0	20.3 ± 2.7	4.3 ± 1.9	72.4 ± 9.6
Total	166					16.4 ± 7.8	86.4 ± 5.3
hNIR spectrometer (LVs = 9, Q^2 = 0.23, RMSEcv = 0.36)[a]							
	n	CHL (n)	ESP (n)	GEO (n)	ITA (n)	Not assigned (n)	Correct classification (%)
CHL	32	20.6 ± 3.0	0.7 ± 1.0	0.0 ± 0.0	0.9 ± 0.9	9.9 ± 2.5	64.3 ± 9.4
ESP	73	0.4 ± 0.8	62.3 ± 2.3	0.7 ± 0.5	0.0 ± 0.0	9.6 ± 2.6	85.3 ± 3.1
GEO	32	0.1 ± 0.4	4.4 ± 1.8	9.9 ± 2.0	0.6 ± 0.5	12.0 ± 1.2	36.5 ± 7.2
ITA	29	0.9 ± 0.7	3.1 ± 1.7	0.0 ± 0.0	13.3 ± 2.4	10.7 ± 2.3	47.4 ± 8.7
Total	166					42.1 ± 4.7	66.3 ± 3.9

For all models, ANOVA p-value < 0.05. [a] Model parameters: mean values obtained with the training sets from 7 iterations. LVs: latent variables of the training model; Q^2: Cumulative fraction of Y variation predicted by the X training model up to the specified latent variable, according to cross-validation; RMSEcv: root mean square error of the cross validation of the training model. CHL: Chile; ESP: Spain; GEO: Georgia; ITA: Italy.

Table S4. Leave 10% out cross validation of PLS-DA cultivar models developed on ground samples analysed by benchtop NIR (NIR), handheld NIR (hNIR) and MIR spectrometers. Results are mean values (± standard deviation) obtained from seven iterations.

Ground samples - Cultivar model: TG/non-TG						
NIR spectrometer (LVs = 10, Q^2 = 0.53, RMSEcv = 0.34)[a]						
	n	non-TG (n)	TG (n)	Correct classification (%)	Sensitivity	Specificity
non-TG	81	79.1 ± 0.9	1.9 ± 0.9	97.7 ± 1.1	0.98 ± 0.01	
TG	73	0.1 ± 0.4	72.9 ± 0.4	99.8 ± 0.5		1.0 ± 0.01
Total	154			98.7 ± 0.7		
hNIR spectrometer (LVs = 9, Q^2 = 0.39, RMSEcv = 0.39)[a]						
	n	non-TG (n)	TG (n)	Correct classification (%)	Sensitivity	Specificity
non-TG	81	69.1 ± 1.9	11.9 ± 1.9	85.4 ± 2.3	0.85 ± 0.02	
TG	73	4.6 ± 1.7	68.4 ± 1.7	93.7 ± 2.4		0.94 ± 0.02
Total	154			89.3 ± 2.0		
MIR spectrometer (LVs = 10, Q^2 = 0.63, RMSEcv = 0.30)[a]						
	n	non-TG (n)	TG (n)	Correct classification (%)	Sensitivity	Specificity
non-TG	81	80.7 ± 0.8	0.3 ± 0.8	99.6 ± 0.9	1.00 ± 0.01	
TG	73	0.0 ± 0.0	73.0 ± 0.0	100.0 ± 0.0		1.0 ± 0.0
Total	154			99.8 ± 0.5		

For all models, ANOVA p-value < 0.05. [a] Model parameters: mean values obtained with the training sets from 7 iterations. LVs: latent variables of the training model; Q^2: Cumulative fraction of Y variation predicted by the X training model up to the specified latent variable, according to cross-validation; RMSEcv: root mean square error of the cross validation of the training model. TG: 'Tonda di Giffoni'; non-TG: other cultivars.

Table S5. Leave 10% out cross validation of PLS-DA origin models developed on ground samples analysed by benchtop NIR (NIR), handheld NIR (hNIR) and MIR spectrometers. Results are mean values (± standard deviation) obtained from seven iterations.

Ground samples - Geographical origin model: CHL/ESP/GEO/ITA							
NIR spectrometer (LVs = 14, Q^2 = 0.66, RMSEcv = 0.25)[a]							
	n	CHL (n)	ESP (n)	GEO (n)	ITA (n)	Not assigned (n)	Correct classification (%)
CHL	32	27.1 ± 1.6	0.4 ± 0.5	0.0 ± 0.0	0.0 ± 0.0	4.4 ± 1.4	84.8 ± 4.9
ESP	73	0.3 ± 0.5	69.7 ± 1.4	0.0 ± 0.0	0.0 ± 0.0	3.0 ± 1.4	95.5 ± 1.9
GEO	32	0.3 ± 0.5	0.9 ± 0.5	21.1 ± 3.7	0.0 ± 0.0	4.7 ± 3.7	78.3 ± 13.6
ITA	29	0.0 ± 0.0	3.4 ± 1.0	0.0 ± 0.0	20.3 ± 2.7	4.3 ± 1.9	72.4 ± 9.6
Total	166					16.4 ± 7.8	86.4 ± 5.3
hNIR spectrometer (LVs = 12, Q^2 = 0.50, RMSEcv = 0.29)[a]							
	n	CHL (n)	ESP (n)	GEO (n)	ITA (n)	Not assigned (n)	Correct classification (%)
CHL	32	20.6 ± 3.0	0.7 ± 1.0	0.0 ± 0.0	0.9 ± 0.9	9.9 ± 2.5	64.3 ± 9.4
ESP	73	0.4 ± 0.8	62.3 ± 2.3	0.7 ± 0.5	0.0 ± 0.0	9.6 ± 2.6	85.3 ± 3.1
GEO	32	0.1 ± 0.4	4.4 ± 1.8	9.9 ± 2.0	0.6 ± 0.5	12.0 ± 1.2	36.5 ± 7.2
ITA	29	0.9 ± 0.7	3.1 ± 1.7	0.0 ± 0.0	13.3 ± 2.4	10.7 ± 2.3	47.4 ± 8.7
Total	166					42.1 ± 4.7	66.3 ± 3.9
MIR spectrometer (LVs = 12, Q^2 = 0.61, RMSEcv = 0.26)[a]							
	n	CHL (n)	ESP (n)	GEO (n)	ITA (n)	Not assigned (n)	Correct classification (%)
CHL	32	20.6 ± 3.0	0.7 ± 1.0	0.0 ± 0.0	0.9 ± 0.9	9.9 ± 2.5	64.3 ± 9.4
ESP	73	0.4 ± 0.8	62.3 ± 2.3	0.7 ± 0.5	0.0 ± 0.0	9.6 ± 2.6	85.3 ± 3.1
GEO	32	0.1 ± 0.4	4.4 ± 1.8	9.9 ± 2.0	0.6 ± 0.5	12.0 ± 1.2	36.5 ± 7.2
ITA	29	0.9 ± 0.7	3.1 ± 1.7	0.0 ± 0.0	13.3 ± 2.4	10.7 ± 2.3	47.4 ± 8.7
Total	166					42.1 ± 4.7	66.3 ± 3.9

For all models, ANOVA p-value < 0.05. [a] Model parameters: mean values obtained with the training sets from 7 iterations. LVs: latent variables of the training model; Q^2: Cumulative fraction of Y variation predicted by the X training model up to the specified latent variable, according to cross-validation; RMSEcv: root mean square error of the cross validation of the training model. CHL: Chile; ESP: Spain; GEO: Georgia; ITA: Italy.

Annex 9. Supplementary material of Publication 11

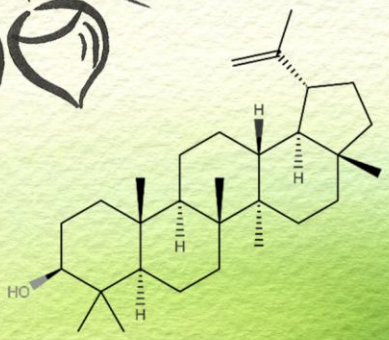
Table S1. Sample information: sample ID (Primary ID), harvest year (Year), region of origin (Region) and the seven training and validation sets. Due to its big size a part of the table is shown as a way of example.

			Training and validation sets						
Primary ID	Harvest year	Country of origin	1	2	3	4	5	6	7
Sample 1	2022	China	Validation	training	Validation	training	training	training	Validation
Sample 2	2022	China	training	training	training	training	Validation	training	training
Sample 3	2022	Turkey							
Sample 4	2022	Russia	training	Validation	Validation	training	training	training	training
Sample 5	2022	China	training	training	training	training	Validation	training	Validation

Table S2. Permutation test results (n = 20 random models) of the preliminary multi-class geographical model.

	Q^2	Q^2 permutation
CHN	0.763	-0.257
TUR	0.44	-0.164
RUS	0.543	-0.194
ESP	0.896	-0.257

N = 253, 9 LVs, Q2 = 0.659, RMSEcv = 0.216, ANOVA p-value < 0.05. CHN: China; TUR: Turkey; RUS: Russia; ESP: Spain.


H
H
H
HO