VOLATILE TERPENE FINGERPRINTING FOR PINE NUT AUTHENTICATION

CÈLIA ASENSIO MANZANO

GRAU EN CIÈNCIA I TECNOLOGIA DELS ALIMENTS

CAMPUS DE L'ALIMENTACIÓ

UNIVERSITAT DE BARCELONA

TUTORA: STEFANIA VICHI

DATA DE PRESENTACIÓ: 27/06/2024

foods

MDPI

Research Article

Volatile terpene fingerprinting for pine nut authentication

Cèlia Asensio Manzano[1]

[1] Department of Nutrition, Food Sciences and Gastronomy, University of Barcelona, Torribera Food Campus, Prat de la Riba, 171, 08912, Sta. Coloma de Gramanet, Spain.
* Correspondence: celiaasensiomanzano16@gmail.com; Tel.: +34 633 42 49 15

Abstract: Pine nuts are highly valued products on the market, with 20 species being the most commercially significant both globally and locally. Among these, Mediterranean pine nuts (*Pinus pinea*) are the most highly valuated, reaching prices up to 100€/Kg, in contrast with other species sold at much lower prices (Chinese and Russian pine nuts). The high prices added to the lack of fast and low-cost analytical methods to assess the authenticity in routine analysis make pine nuts highly vulnerable to fraudulent practices. This study proposes a reliable method for pine nuts geographical and botanical origin authentication. The volatile and semi-volatile terpene hydrocarbon fingerprint of a set of 245 pine nuts from different origins (Spain, China, and Russia) and different species was analysed by HS-SPME-GC-MS. PLS-DA models were built to differentiate between Iberian and non-Iberian samples and between production regions on *Pinus pinea* samples, with satisfactory clustering on all categories based on their respective PLS-DA score plots. Both models were internally and externally validated, achieving correct classification values of 100% and over 96% respectively, ensuring that model predictions are reliable. Hence, this method has proved to be a suitable option for pine nut authentication on industry routine analysis supporting official controls.

Keywords: Pine nut; Authenticity; Food fraud; Fingerprinting; Sesquiterpene hydrocarbons; HS-SPME-GC-MS; PLS-DA

1. Introduction

Pine nuts are among the most expensive products on the market used in many culinary preparations worldwide. These nuts are considered gourmet healthy products, due to their nutritional values. They are rich in proteins (35%) and fats (50%) predominantly unsaturated fatty acids such as omega-6 and omega-9, and contain a great number of micronutrients, liposoluble bioactive and other compounds of interest (1). According to the Food and Agriculture Organization (FAO), only 29 of the 636 scientifically recognized species of the genus *Pinus* produce edible pine nuts, and only 20 of these are significantly commercialized both globally and locally (2). One of the most significant species worldwide is the southern European species *Pinus pinea*, which has been consumed for more than 20 centuries (1). Its production covers an area of 903,723 ha distributed mainly in Spain (490.000 ha, especially in Catalonia and Castile-Leon), Portugal (130.000 ha), Italy (40.000 ha) and Turkey (183,128 ha) (3). Besides *Pinus pinea*, the most commercially important species of pine nuts worldwide are *Pinus koraiensis* (Chinese and Korean pine), *Pinus gerardiana* (Pakistani pine) and *Pinus sibirica* (Russian pine) (4).

Mediterranean pine nut (*Pinus pinea*) is the most highly valued species on the market, with an exceptional flavour and nutritional composition. Compared to Chinese and Pakistani pine nuts, it contains double protein amount and less carbohydrates and fats. *Pinus pinea* kernels are thin and with a homogeneous soft colour, can be distinguished physically and organoleptically from other important species such as *Pinus koraiensis* kernels,

Citation: To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Received: date
Revised: date
Accepted: date
Published: date

which are thicker, with triangular shape and a characteristic brown hat on the tip, or *Pinus gerardiana* kernels, which are cylindrical and of a darker tone. However, these differences may not be detected by the consumer who are unaware of these variations, or they may not be useful if this product is in flour or another processed forms (5). Although more appreciated, there are no recognized geographical designation to differentiate Mediterranean *Pinus pinea* from other species or origins (1). These are mainly consumed locally, but their production is insufficient to satisfy the current high demand. In contrast, production and particularly exportation of pine nuts from other countries such as China or Russia have increased considerably in the last years, with China leading the market over the past decade, accounting for 61% of global exports in 2021 (mainly from *Pinus koraiensis*) (6). The pine nut production and commercialization lacks great commercial structures and involves numerous intermediaries between the producer and the consumer which makes traceability harder (1). Regarding market value, *Pinus pinea* kernels can reach prices of up to 100 €/Kg, while other species sold indistinctly under the generic name "pine nuts" are available at much lower prices in the market and compete with Mediterranean production (7).

Therefore, the high variance of prices, lack of geographical and botanical origin traceability and high competitiveness from emerging markets, make pine nuts highly vulnerable to fraudulent practices. According to the European Commission, during the first three months of this year (2024), 3,5% of food fraud suspicions have been reported in Europe in the category "Nuts, nuts products and seeds", which include pine nuts (8–10).

In addition to economic repercussions, food safety concerns are a significant consequence of fraud. Any non-compliance with label specifications means that the composition, including absence of allergens and other undesirable compounds, cannot be guaranteed, thereby raising important safety issues. In the case of pine nuts, it is notable to mention Pine Mouth Syndrome (PMS), also called pine nut syndrome (PNS), a taste disturbance also known as cacogeusia, characterized by a metallic and bitter flavour that emerges after 1-3 days of pine nuts consumption (11). This alteration was first described in the European medical conference 2001 and several hundred cases have been described in the literature after that. It has been exclusively associated with the consumption of a non-edible species of pine nuts, *Pinus armandii,* which is sometimes sold mixed with Chinese pine nuts (*Pinus koraiensis*), or as other edible pine nut species (12,13).

Research on analytical methods for food authentication has grown greatly over the past 20 years, resulting in numerous research articles. Even so, the food industry, particularly the nut industry, still lacks fast and low-cost analytical methods to assess the authenticity of products in routine analysis (14). Therefore, the development of efficient and affordable tools to determine the botanical and geographical origin of pine nuts is crucial to prevent food fraud and increase consumer confidence.

In recent years, several studies have been carried to find suitable methods for the authentication of nuts. The most significant analytical methods for high-fat content foods include spectroscopy, stable-isotope analysis, DNA-based methods and chromatography methods that have high selectivity, sensitivity and accuracy (14). Spectroscopic techniques are low-cost, non-destructive and easy to implement. Near infrared spectroscopy has been used for geographical identification of samples of *Pinus pinea* kernels grown in different areas of Chile (15), and for the authentication of an Italian Hazelnut PDO (Nocciola Romana) (16). Specific isotopic markers have shown satisfactory results for the geographical authentication of hazelnuts (17), but they are not suitable for verifying their botanical origin, as these markers are primarily influenced by soil and climatic factors. In this regard, DNA methods are reliable to identify the botanical origin of nuts. Although they tend to be overly complex and expensive, various studies are currently demonstrating the successful application of rapid and cost-effective molecular methods, such as RAPD-PCR for differentiation and identification of hazelnut cultivars (18) or the study of polymorphic

sites of the chloroplast genome for varietal determination of hazelnuts (19). Nevertheless, although genetic approaches are suitable to assess the botanical origin, they cannot determine the geographical provenance. Likewise, the study of the fatty acid profile was proposed for the botanical identification of pine nuts (20), but their efficiency as geographical markers has not been demonstrated. In contrast, methods based on gas chromatography coupled to mass spectrometry (21) allow both botanical and geographical authentication of several hazelnuts with classification rates higher than 90%, proving to be effective methods and applicable to routine analysis. Most of the current methods are preliminary methods, highlighting the necessity to develop methods that include an external validation.

Some of the above-mentioned studies are based on a targeted approach, which focuses on the detection of specific analytes or a group of them. These methods are useful for food authentication when the molecules to be detected are known a priori, and they are usually robust, reproducible and easy transferable among different laboratories. However, they often provide limited information for detecting fraud and insufficient protection for consumers. Additionally, when working with complex matrices such as food products, the quantification of compounds using a target approach can be challenging and may provide insufficient information when dealing with complex issues like origin and species authentication. In these cases, non-targeted methods that enable the acquisition of multiple non-target parameters to obtain a comprehensive view of the sample composition could be a better option. Fingerprinting methods are non-targeted analytical approaches based on the use of raw analytical signals, such as a chromatogram, and are currently a major focus of research for food authentication (22,23).

In fingerprinting approaches, once all data are acquired through one or more analytical techniques such as spectroscopy or chromatography, multivariate qualitative chemometric methods are applied to extract relevant information and discriminate the data based on their metabolic profile (24). In multivariate methods several steps are followed: exploratory techniques, classification or discriminant analysis and a validation step. Exploratory techniques are unsupervised methods that provide information on the relationship between samples, variables and the interaction between samples and variables, revealing trends among them (25). The most popular exploratory techniques are principal components analysis (PCA) and hierarchical cluster analysis (HCA). The PCA is based on the generation of new variables (main component or PCs) as a combination of the original variables that retain the maximum possible information of the original data. HCA is useful for identifying patrons and underlying structures as it organizes information in hierarchical groups based on similarity and represented in a denogram. (26). Classification or discriminant techniques are supervised techniques that associate analytical data of samples with their membership in predefined classes. They classify unknown samples in the class whose characteristics they most closely match. The main discriminant techniques are partial least squares discriminant analysis (PLS-DA), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and k nearest neighbours (KNN), among others. PLS-DA is one of the most commonly used discriminant techniques. It involves defining multiple classes, after which samples are classified into one of these classes based on maximizing the correlation between the data and each category. In doing so, PLS-DA identifies the features that exhibit the greatest differences between categories while reducing the impact of variables not relevant to a specific category. However, this method tends to overfit the data, making an external validation necessary (25). External validation is performed by predicting the class of samples that had not been used to construct the model, aiming to verify that the results are statistically valid and that can accurately classify new samples (26).

Regarding marker selection, studies on virgin olive oils (27,28) revealed that sesquiterpene hydrocarbons are highly effective for varietal and geographical authentication.

Sesquiterpenes are a group of C-15 (29) semi-volatile secondary metabolites that play an important role in defence against herbivores and plant pathogens (27). The production of these compounds is influenced by pedoclimatic and genetic factors, making it closely linked to the cultivar and geographical area, while being minimally influenced by storage and processing conditions (30). These compounds can be easily extracted by headspace solid-phase microextraction (HS-SPME) of the sample headspace followed by GC-MS (31), a simple, solvent-free and automatable technique. Sesquiterpene fingerprinting performed by HS-SPME-GC-MS followed by PLS-DA has proven to be a good choice for botanical and geographical authentication of virgin olive oil in routine analyses. This methodology could be useful for the authentication of other food matrices. While other nut species typically lack appreciable amounts of sesquiterpenes in their kernels, conifers are known for their abundant production of volatile and semi-volatile terpene (VST) metabolites (32). Some VST hydrocarbons have also been identified in pine nut kernels, indicating that this fraction could serve as potential authentication markers (33).

Hence, the objective of this study is to verify whether VST fingerprinting combined to PLS-DA, which proven successful in other food matrices, could serve as an effective tool for the routine authentication of both the geographical and botanical origin of pine nuts. For this purpose, the VST fingerprints of 245 pine nuts samples from different origins (Spain, China, and Russia) and different species were analysed by HS-SPME-GC-MS. PLS-DA models were built to differentiate between (i) different species of pine nuts based on their country of origin, and (ii) Iberian *Pinus pinea* samples from Catalonia and Castile and Leon regions. Both internal (cross-validation) and external validation were performed. Finally, regression coefficients of PLS-DA model were evaluated in order to tentatively identify the compounds characterizing each class of pine nut and discriminate then from others.

2. Materials and Methods

2.1 Sampling

The sample set consisted of 245 traceable pine nuts samples from different geographical origins (Table 1). Of these samples, 170 were Iberian production pine nuts (*Pinus pinea* cultivars) and 75 were non-Iberian production pine nuts (other species). Among the Iberian samples, 74 were cultivated in Catalonia (CAT), directly obtained from the Institut de Recerca i Tecnologia Agroalimentària (IRTA) and 96 were cultivated in Castile and Leon (CL), directly obtained from the Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA) and the Centro de Servicios y Promoción Forestal y de su Industria de Castilla y León (CESEFOR). The remaining non-Iberian pine nuts were cultivated in different non-European countries (China, Russia and others) and obtained from commercial suppliers. Samples were collected over four consecutive harvest years (2020, 2021, 2022 and 2023). The entire set was preserved at 4ºC and analysed in January 2024. Random selection was employed during the sample analysis to prevent any selection bias.

Table 1. Number and geographical origin of all samples from the four harvest years: 2020,2021, 2022 and 2023.

		Harvest years				
	Origin	**2020**	**2021**	**2022**	**2023**	Total
Iberian	Castile and Leon (CL)	11	10	75	0	96
	Catalunya (CAT)	25	24	25	0	74
Non-Iberian	China	0	20	29	0	49
	Russia	0	5	12	1	18
	Non-EU (others)	0	5	3	0	8
	Total	36	64	144	1	245

2.2 Headspace-solid phase microextraction (HS-SPME)

Around 1 g of pine nuts (5-8 pine nuts) was introduced into vials of 10 mL closed with headspace caps. The headspace solid-phase microextraction (HS-SPME) was performed with the help of an autosampler Combi-pal (CTC Analytics, Zwinger, Switzerland) at the conditions reported by Vichi S. et al. (31)**.** Briefly, sample was conditioned under agitation (250rpm) for 10 minutes at 70 °C. After that, a divinylbenzene/carboxen/polydimethylsiloxane (50/30 μm DVB/CAR/PDMS, 2cm length) fiber provided by Supelco (Bellefonte, PA) was inserted through the PTFE/silicone septum and exposed to the sample headspace, at 70 °C for 60 minutes. Subsequently, the fiber was removed into the protective needle and exposed into the gas chromatography injection port at 260 °C for 10 min to allow the desorption of analytes. In this step, the injector was maintained in splitless mode for 5 min.

2.3 Gas chromatography-mass spectrometry (GC-MS)

The sample set was analysed by an Agilent 6890 N Network GC system coupled to a quadrupolar mass selective analyser Agilent 5975C Inert MSD (Agilent Technologies, Santa Clara, California, USA). The carrier gas used was helium 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) from Supelco (Bellefonte, PA). Column temperature was initially held at 40 °C for 3 min, then increased to 100 °C at a rate of 4 °C/min, after that increased to 200 °C at 5 °C/min and finally increased to 260 °C at 15 °C/min, holding the last temperature for 5 min. Other temperatures were 230 °C for ion source and 280 °C for transfer line. Mass spectra were acquired at 2.3 scan/s with an electron energy of 70 eV. Data was acquired using the selected ion monitoring (SIM) mode, obtaining the Extracted Ion Chromatogram (EIC) of 7 specific ions: *m*/*z* 93, 95, 119, 159, 161, 189, 204, which had been reported to be specific for VST (34)**.**

2.4 Fingerprinting approach

The seven EICs were acquired from 0,094 min to 47,192 min obtaining 6621 scans per ion and therefore 46347 variables per sample (6621 scans x 7 ions). After acquiring data for all samples, a data matrix was constructed for each ion, with scan intensities of each Extracted Ion Chromatogram (EIC) represented along the columns and individual samples along the rows. Then, EICs of each ion were normalized and aligned among them using the algorithm Correlation Optimized Warping (COW) on Matlab®. Finally, the seven aligned matrices were concatenated conforming a two-way unfolded matrix (245 samples x 46347 variables).

2.5 Chemometrics

2.5.1 Data pre-processing and exploration

Before performing partial least squares discriminant analysis (PLS-DA) a pre-processing and exploration step was performed using SIMCA software v13.0 © (Umetrics AB, Sweden). For pre-processing, two different treatments were tested (mean centering and scaling), where scaling proved to be the optimal one. For exploration, a Principal Component Analysis (PCA) was performed in order to identify potential outliers according to Hotelling's T^2 range and model residuals.

2.5.2 Partial least squares discriminant analysis (PLS-DA)

Two different types of binary PLS-DA models were built using SIMCA software v13.0 © (Umetrics AB, Sweden): one to classify all samples from distinct species (n=245) between "Iberian" and "non-Iberian", and one to classify only *Pinus pinea* Iberian samples (n=170) between "CAT" and "CL". Hotelling's T^2 and range and model residuals were evaluated to identify potential outliers.

The full data set (n=245) was divided randomly into training set and validation set using the Matlab® program, always maintaining the original proportions of the sample classes. 80% of the data set was used for the training set (n= 196) and 20% of this was used for the validation set (n= 49). This process was effectuated three times obtaining three different validation sets (three iterations).

The training sets were used to construct PLS-DA training models. On each model, the number of Latent Variables (LV) was selected according to the first lowest RMSEcv. After choosing the LV, a verification was carried out to confirm the models were not overfitted by doing both ANOVA of the cross-validated predictive residuals (p-value) and permutation tests where 20 different models were developed and compared with the original model. Finally, a 10%-out cross-validation was carried out as the internal validation, obtaining a Root Mean Squared Error of Cross Validation (RMSEcv) and misclassification results (expressed as mean of three iterations ±standard deviation), which were used to evaluate the suitability of the three type of models (three iteration each).

2.5.3 External validation (EV)

The external validation was conducted by using each training model to predict the class of the corresponding validation samples. The prediction efficiency of each model was evaluated by calculating the mean percentage and standard deviation of correct classification across three iterations.

2.5.4 Evaluation of PLS-DA regression coefficients

The regression coefficients of PLS-DA model built using the full sample set were evaluated to assess the contribution of variables from each EIC. A regression coefficient was considered significant when its value exceeded the standard error of cross validation. For the variables contributing the most to prediction, the spectrum of the corresponding chromatographic peak was obtained in the full scan mode, in order to tentatively identify the compounds that characterize each class of pine nut and discriminate it from others.

3. Results and discussions

3.1 Data obtaining

In this study, chromatographic data was obtained by applying headspace solid-phase microextraction gas chromatography (HS-SPME-GC-MS) on the whole pine nut kernels. Preliminary tests comparing the analysis of whole and ground pine nuts showed no significant difference in the response of the VST (data not shown). Consequently, whole pine nut kernels were used for the analysis to reduce sample manipulation, process time and cost, which is favourable for its application in routine analyses.

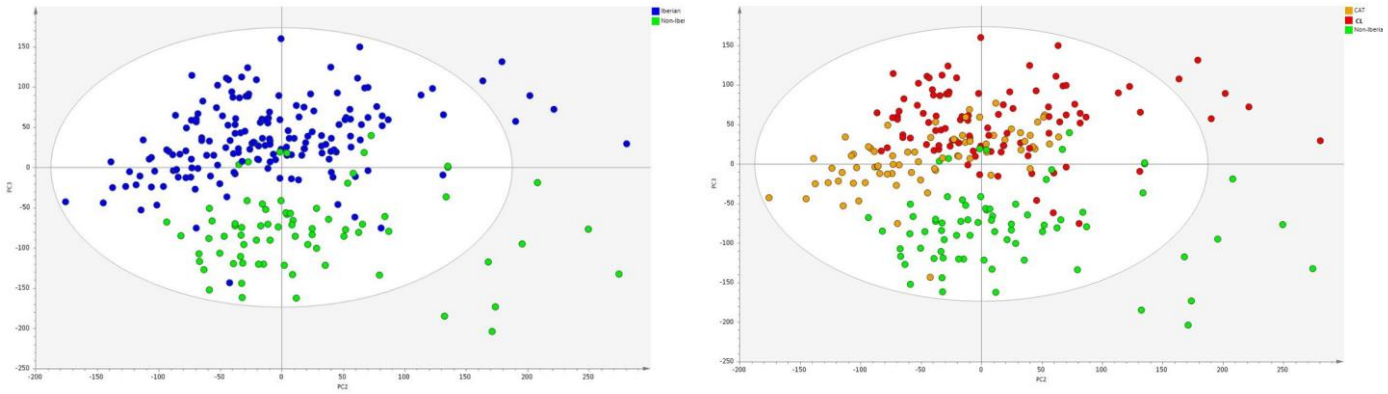
3.2. Data pre-processing and exploratory analysis

Once the EICs of the seven ions specific for VST (*m*/*z* 93, 95, 119, 159, 161, 189, and 204) were obtained, they were normalized and subsequently aligned using the COW algorithm, specific for chromatographic data. Alignment was performed in order to correct the retention shifting between samples caused by instrumental factors inherent in chromatographic techniques. Normalization was performed to correct magnitude changes that can occur when analysing a large sample set by GC-MS over an extended period due to variations in instrumental response (35).

When all data was aligned and normalized, the seven ion matrices were concatenated conforming a two-way unfolded matrix (253 samples x 46347 variables). Two different treatments prior to multivariate analysis were tested: centering and autoscaling. On the one hand, centering reduces differences between high and low abundant metabolites of the same sample by subtracting the mean of each variable from the data. On the other hand, autoscaling makes samples comparable by removing the scale differences among them; it involves centering the data and dividing it by the standard deviation (36). Autoscaling is recommended for chromatographic data when comparing minor and major compounds with different intensities (37). In this study, autoscaling proved to be the best pretreatment.

On exploratory analysis, no potential outliers were detected according to Hotelling's T^2 range and model residuals. Preliminary examination of the PCA score plots indicated that, even in a non-supervised analysis, pine nuts clustered successfully based on their geographical and botanical origin. Figures 1a and 1b depict the same PCA score plot obtained using the entire sample set, but evidencing the samples according to their belonging to different classes and sub-classes. In Figure 1a, samples are coloured as "Iberian" and "Non-Iberian" pine nuts. Although there was a very slight overlap between both groups, the clustering of the Iberian (*Pinus pinea*) and non-Iberian (other pine nut species) samples was remarkable. On both categories, some samples stood out from the central circle. They were not considered outliers because they aligned with the variability observed in the other samples, indicating they represented natural variability within the same group.

Observing the same PCA score plot but dividing "Iberian" *Pinus pinea* samples into "CAT" and "CL" origin (Figure 1b), it is notable to mention that, even between samples of the same species, there was a discrete clustering based on geographical origin. Even so, "CAT" and "CL" samples overlapped significantly, indicating that these samples, although cultivated in different regions, could share some similar characteristics.



(**a**) (**b**)

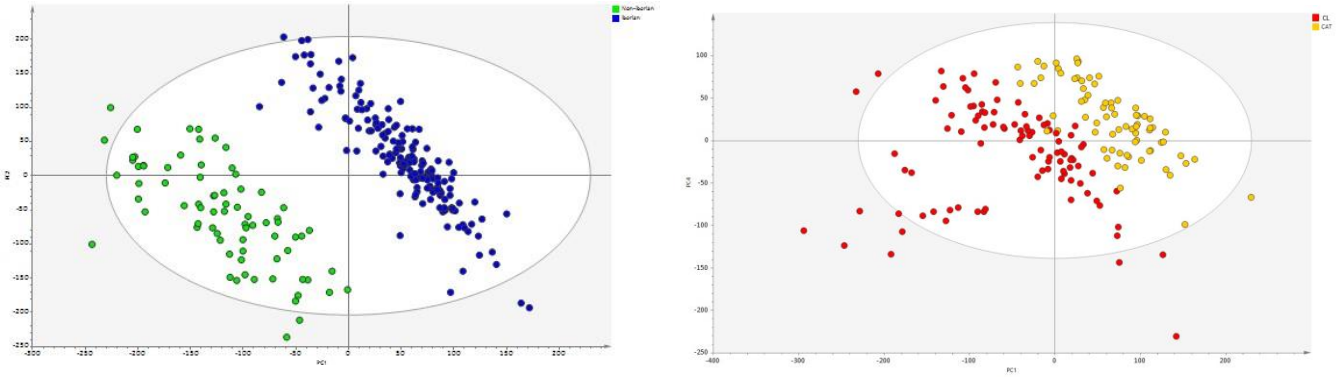
Figure 1. Score plots of PCA with third and second principal components based on pine nut VST data coloured by **a)** "Iberian"/"non-Iberian" categories (n=245, 6 PC, Q^2=0.613); **b)** "CAT"/"CL"/"non-Iberian" categories (n= 245, 6 PC, Q^2=0.613).

3.3. PLS-DA authentication models development and internal validation

As PCA is an unsupervised analysis, it can be significantly influenced by instrumental noise and variables unrelated to sample classification. Conversely, the supervised technique PLS-DA identifies the most distinctive features between categories while minimizing the influence of unrelated variables. As this is expected to enhance discrimination, PLS-DA was applied for the development of subsequent classification models.

Two different types of binary PLS-DA models were built to assess the efficiency of VST fingerprinting for pine nuts authentication: (i) a model to classify samples from different species and origins (n=245) as "Iberian" *Pinus pinea* and "non-Iberian" samples from distinct species; (ii) a model to classify only Iberian *Pinus pinea* samples (n=170) into "CAT" and "CL" categories. No potential outliers were found in any of these models.

As expected, since PLS-DA is specifically designed to discriminate between classes, PLS-DA score plot of "Iberian" / "non-Iberian" PLS-DA model (Figure 2a) showed better separation between classes with respect to the corresponding PCA score plot (Figure 1a). Moreover, the Iberian samples presented lower dispersion compared to the non-Iberian ones. The lower variability among Iberian samples is likely because they were from the same species and a more confined geographical area.



(**a**) (**b**)

Figure 2. Score plots of PLS-DA models based on pine nuts VST fingerprinting (mean and standard deviation of three iterations): **a)** "Iberian"/"non-Iberian" model (n=245, 4 latent variables or LVs, RMSEcv=0.095, $p < 0.05$); **b)** "CAT"/"CL"- model (n=170, 5 LVs, RMSEcv=0.157, $p < 0.05$).

The score plot of the "CAT" / "CL" model (Figure 2b), revealed appreciable differences between "CAT" and "CL". Despite some slight overlap, consistent with the patterns observed in the PCA score plot, the PLS-DA was effective in distinguishing between the two classes. This suggests that while "CAT" and "CL" pine nuts shared characteristics due to being from the same species, they could still be differentiated based on their regional cultivation differences. Regarding the dispersion of the samples, "CAT" samples exhibited tighter clustering than "CL" samples.

To assess the discriminant capacity of these models, an internal validation was conducted through a leave 10% out cross-validation (Tables 2 and 3). The cross validation results for both the "Iberian" / "non-Iberian" and the "CAT" / "CL" PLS-DA models demonstrated a classification accuracy of 100% in all cases.

Table 2. Results of the leave 10%-out cross-validation of the "Iberian" vs "non-Iberian" PLS-DA model (mean ± standard deviation of three iterations).

"Iberian" vs "non-Iberian" model[1]				
	Members (n)	Iberian (n)	Non-Iberian (n)	Correctly classified (%)
Iberian	170	170 ± 0	0 ± 0	100% ± 0
Non-Iberian	75	0 ± 0	75 ± 0	100% ± 0
Total	245	170 ± 0	75 ± 0	100%± 0

[1] N = 245, 4 LVs, RMSEcv=0.095, ANOVA p-value <0.05

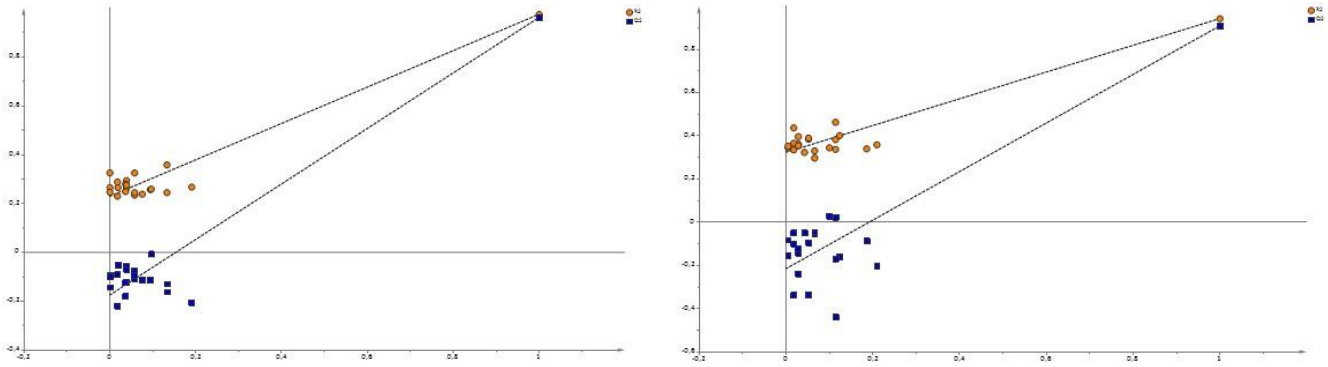
Table 3. Results of the leave 10%-out cross-validation of the "CAT" vs "CL" PLS-DA model (mean ± standard deviation of three iterations).

"CAT" vs "CL" model[1]				
	Members (n)	CAT (n)	CL (n)	Correctly classified (%)
CAT	74	74 ± 0	0 ± 0	100% ± 0
CL	96	0 ± 0	96 ± 0	100% ± 0
Total	170	74 ± 0	96 ± 0	100% ± 0

[1]N = 170, 5 LVs, RMSEcv=0.157, ANOVA p-value <0.05

To exclude model overfitting, ANOVA results and permutation test were carried out. Overfitting occurs when a model is too finely tuned to the specific dataset, accounting for not only the relationship between predictors and response but also the noise and other extraneous factors, making the model less applicable to new data. This test shuffles class labels to create multiple random models. If the actual model outperforms these random models, it confirms that the observed group differences are real and not by chance. The performance of the original model compared to random models is assessed by the Prediction Coefficient (Q^2) and the Coefficient of Determination (R^2). Q^2 measures the model's ability to predict new samples, while R^2 indicates how well the model explains the variability in the training data. Figure 3 illustrates the results of permutation test conducted for the PLS-DA models for pine nuts authentication. The positive Q^2 of the model, in opposition to the negative Q^2 values of the random models, confirms the absence of model overfitting.

In resume, ANOVA results ($p < 0,05$) and permutation test showed that the models were not overfitted and had a high discriminant capacity.



(**a**) (**b**)

Figure 3. Permutation test assessed by the Prediction Coefficient (Q^2) and the Coefficient of Determination (R^2) of (**a**) "Iberian" vs "Non-Iberian" PLS-DA model (**b**) "CAT" vs "CL" PLS-DA model.

3.4. External validation

External validation is a crucial step in the development of PLS-DA models to ensure reproducibility of predictions, and validate the meaningfulness of results for practical implementation of the model. If misclassified samples in external validation are greater than those in internal validation, it may indicate that the model is overfitted, considering noise instead of the underlying pattern. To conduct external validation of each model ("Iberian/"non-Iberian" and "CAT/"CL"), the corresponding sample sets (n=245 and n=170, respectively) had been randomly divided into training set (80% of the samples: n=196 and n=136, respectively) and validation set (80% of the samples: n=49 and n=34, respectively). This process was conducted three times (three iterations) to ensure that the external validation was set-independent. After optimization and internal validation described in 3.3, training models were applied to predict the class of the respective validation samples. The prediction efficiency of each model was evaluated by calculating the percentage of correct classification, expressed as mean and standard deviation across the three iteration sets. In line with the results obtained in the internal validation, excellent results (Tables 4 and 5) were obtained in the external validation for all models. All the pine nut samples were correctly classified as "Iberian" and "non-Iberian", and *Pinus pinea* Iberian samples were classified by their region of origin with correct classification rates higher than 96% in all categories, and an overall accuracy of 98%. These results demonstrated the high efficiency of PLS-DA models based on VST fingerprinting for pine nut authentication.

Table 4. Results of external validation of the "Iberian" vs "non-Iberian" PLS-DA model. Mean and standard deviation of the three sample sets (3 iterations), for each category.

"Iberian" vs "non-Iberian" model[1]				
	Members (n)	Non-Iberian (n)	Iberian (n)	Correctly classified (%)
Non-Iberian	15	15 ± 0	0 ± 0	100% ± 0
Iberian	34	0 ± 0	34 ± 0	100% ± 0
Total	49	15 ± 0	34 ± 0	100% ± 0

[1] N = 196, 4 LVs, ANOVA p-value <0.05

Table 5. Results of external validation of the "CAT" vs "CL" PLS-DA model. Mean and standard deviation of the three sample sets (3 iterations), for each category.

"CAT" vs "CL" model[1]				
	Members (n)	CAT (n)	CL (n)	Correctly classified (%)
CAT	15	15 ± 0	10 ± 0	100% ± 0
CL	19	0,7± 0,6	18,3 ± 0,6	96% ± 0,03
Total	34	15,7 ± 0,6	18,3 ± 0,6	98% ± 0,02

[1]N = 136, 5 LVs, ANOVA p-value <0.05

3.5. Exploration of PLS-DA regression coefficients

Regression coefficients of PLS-DA models built using the full sample set were evaluated to determine which variables from each EIC contributed most significantly to the model's prediction, ensuring that the models rely on meaningful chemical information. This exploration was carried out for both the "Iberian" vs "non-Iberian" model and the "CAT" vs "CL" model.

For the variables contributing the most to prediction, the spectrum of the corresponding chromatographic peak was obtained in the full scan mode, in order to tentatively identify the compounds that distinguish each class of pine nut. All EICs provided relevant information, as indicated by the regression coefficient plots (Figures 4a and 4b). Particularly, *m*/*z* 93, 95, 119 and 204 provided the most influential contributors to discrimination. Total Ion Chromatogram (TIC) highlighting the variables associated to the most significant regression coefficients (Figures 3c and 3d) revealed that not only major but also very minor compounds significantly contributed to the discrimination of both models. This underscores that minor VST , typically overlooked in a targeted approach, played a crucial role in these discrimination models, remarking why fingerprinting approach could be a better option for pine nuts authentication.

The most relevant compounds for each model, highlighted in Figures 4c and 4d were tentatively identified on the basis of their mass spectra and elution order as mono and sesquiterpene compounds, confirming that the models were based on meaningful data. Monoterpene compounds mainly distinguished non-Iberian pine nuts, while Iberian ones were mainly distinguished by their sesquiterpene pattern. For the Iberian samples, the most relevant compounds presented a mass spectrum that could be tentatively attributed to limonene (a cyclic monoterpene), amorphene, cubebene or junipene (all three sesquiterpene hydrocarbons), among others. For the non-Iberian samples, the most relevant peaks could be tentatively identified as monoterpene compounds such as α-pinene, β-pinene, cymene, or myrcene, among others. Several chromatographic peaks were only present in samples of one of the "Iberian" or "non-Iberian" classes. This information may be considered for the authentication of pine nuts not only on the entire kernel or flour, but also as part of complex processed foods. It must be clarified that the goal of the study was not to conduct an exhaustive study of all discriminant variables or shift towards targeted analysis. Instead, we focused on verifying the terpene nature of the most relevant variables and gaining insights into their overall molecular structure. More detailed and focused studies would be required to study deeply the specific chemical structure of compounds that were relevant to the model prediction.

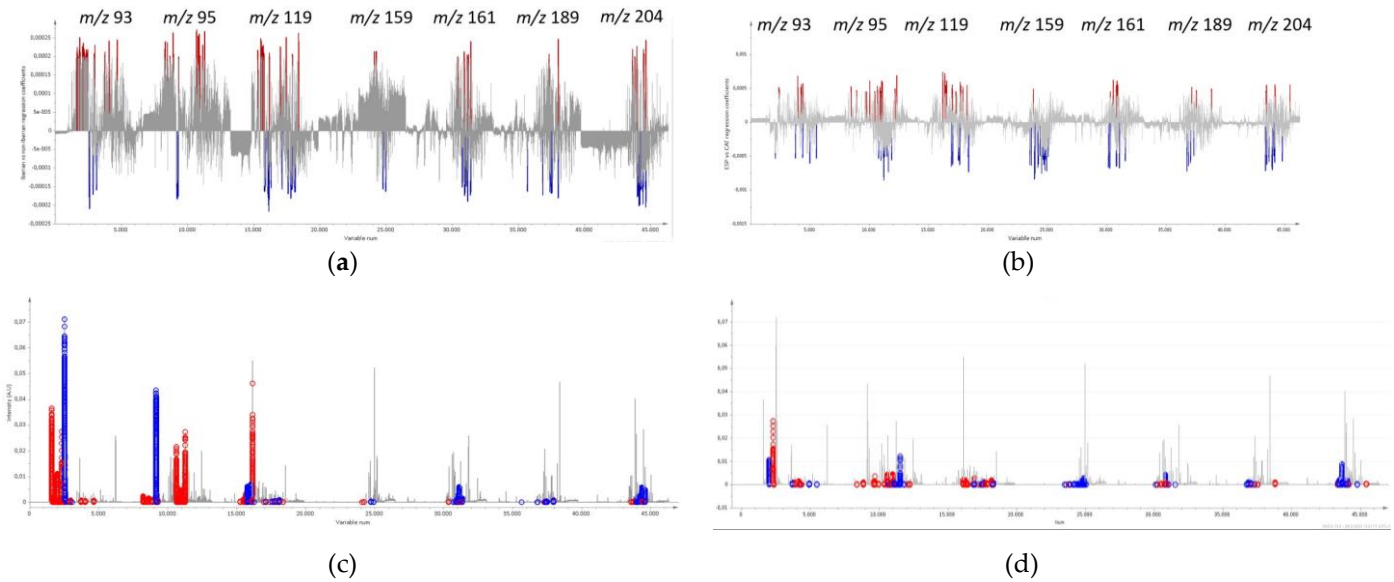


Figure 4. Exploration of regression coefficients of PLS-DA models. **a)** PLS-DA regression coefficients of the "Iberian" vs "non-Iberian" model, selected from the significant ones according to a threshold of 0.0002 and -0.00015 (Blue: relevant for "Iberian"; red: relevant for "non-Iberian"). **b)** PLS-DA regression coefficients of the "CAT" vs "CL" model, selected from the significant ones according to a threshold of 0.00045 and -0.0005 (Blue: relevant for "CAT"; red: relevant for "CL"). **c)** Total Ion Chromatogram (TIC) highlighting the acquisition points corresponding to the most relevant regression coefficients of (a) (Blue for "Iberian" coefficients; red for "non-Iberian" coefficients). **d**) Total Ion Chromatogram (TIC) highlighting the acquisition points corresponding to the most relevant coefficients (Blue for "CAT" coefficients; red for "CL" coefficients).

4. Conclusions

In conclusion, VST fingerprinting obtained through HS-SPME-GC-MS proved to be a suitable method for geographical and botanical authentication of pine nuts. VST, previously studied for the authentication of other food matrices and abundantly produced by conifers, have proven to be effective markers for pine nut authentication. In addition, the use of a solvent-free and automatable data acquisition technique applied on the whole pine nut kernels, could reduce both time and costs, making it suitable for routine analyses on official controls. Moreover, the applied chemometric approach (PLS-DA) has allowed the discrimination of samples according to the characteristic patterns of each class. Successful discrimination results have been obtained on the models discriminating between Iberian and non-Iberian samples and between Iberian samples cultivated in different geographical areas (Catalonia and Castile-Leon) with correct classification values of 100% for internal validation and values above 96% of correct classification for external validation, ensuring that model predictions are reliable. Finally, the study of the regression coefficients has demonstrated that the model's predictions are based on significant chemical information, with both major and minor VST contributing individually to the method's discrimination. This highlights why the fingerprinting approach could be a better option for pine nut authentication.

References

1. Mutke S, Piqué M, Calama R. Mediterranean stone pine for agroforestry: proceedings of the Agropine 2011 International Meeting : Valladolid (Spain), 17-19 November 2011. Zaragoza, Spain: CIHEAM; 2013.
2. Ciesla WM. Non-wood forest products from conifers. Rome: Food and Agriculture Organization of the United Nations; 1998. 124 p. (Non-wood forest products).
3. Bonari G, Chytrý K, Çoban S, Chytrý M. Natural forests of Pinus pinea in western Turkey: a priority for conservation. Biodivers Conserv. 2020 Dec 1;29(14):3877–98.
4. Reyes JB, Pérez SFO. Aproximación al sector del piñón en España.
5. Mutke S, Calama R, González-Martínez SC, Montero G, Gordo FJ, Bono D, et al. Mediterranean Stone Pine: Botany and Horticulture. In: Janick J, editor. Horticultural Reviews [Internet]. 1st ed. Wiley; 2011 [cited 2024 May 23]. p. 153–201. Available from: https://onlinelibrary.wiley.com/doi/10.1002/9781118100592.ch4
6. Nuts & dried fruits statistical yearbook 2022/2023. INC International Nut and Dried Fruit Council, 2023;
7. Calama R, Gordo J, Madrigal G, Mutke S, Conde M, Montero G, et al. Enhanced tools for predicting annual stone pine (Pinus pinea L.) cone production at tree and forest scale in Inner Spain. For Syst. 2016 Dec 2;25(3):e079–e079.
8. January 2024 report on EU Agri-Food Fraud Suspicions. European Commission; 2024.
9. February 2024 report on EU Agri-Food Fraud Suspicions. European Commission; 2024.
10. March 2024 report on EU Agri-Food Fraud Suspicions. European Comission; 2024.
11. Awan HUM, Pettenella D. Pine nuts: a review of recent sanitary conditions and market development. 2017;
12. Munk MD. "Pine Mouth" Syndrome: Cacogeusia Following Ingestion of Pine Nuts (Genus: Pinus). An Emerging Problem? J Med Toxicol. 2010 Jun;6(2):158–9.
13. Munk MD. Pine Mouth (Pine Nut) Syndrome: Description of the Toxidrome, Preliminary Case Definition, and Best Evidence Regarding an Apparent Etiology. Semin Neurol. 2013 May 15;32(05):525–7.
14. Valdés A, Beltrán A, Mellinas C, Jiménez A, Garrigós MC. Analytical methods combined with multivariate analysis for authentication of animal and vegetable food products with high fat content. Trends Food Sci Technol. 2018 Jul;77:120–30.
15. Loewe V, Navarro-Cerrillo RM, García-Olmo J, Riccioli C, Sánchez-Cuesta R. Discriminant analysis of Mediterranean pine nuts (*Pinus pinea* L.) from Chilean plantations by near infrared spectroscopy (NIRS). Food Control. 2017 Mar 1;73:634–43.
16. Moscetti R, Radicetti E, Monarca D, Cecchini M, Massantini R. Near infrared spectroscopy is suitable for the classification of hazelnuts according to Protected Designation of Origin. J Sci Food Agric. 2015 Oct;95(13):2619–25.

17. Torres-Cobos B, Rosell M, Soler A, Rovira M, Romero A, Guardiola F, et al. Investigating isotopic markers for hazelnut geographical authentication: Promising variables and potential applications. Food Chem. 2024 Aug;449:139083.

18. Felbinger C, Kutzsche F, Mönkediek S, Fischer M. Genetic profiling: Differentiation and identification of hazelnut cultivars (Corylus avellana L.) using RAPD-PCR. Food Control. 2020 Jan;107:106791.

19. Lang C, Weber N, Möller M, Schramm L, Schelm S, Kohlbacher O, et al. Genetic authentication: Differentiation of hazelnut cultivars using polymorphic sites of the chloroplast genome. Food Control. 2021 Dec;130:108344.

20. Destaillats F, Cruz-Hernandez C, Giuffrida F, Dionisi F. Identification of the Botanical Origin of Pine Nuts Found in Food Products by Gas−Liquid Chromatography Analysis of Fatty Acid Profile. J Agric Food Chem. 2010 Feb 24;58(4):2082–7.

21. Torres-Cobos B, Quintanilla-Casas B, Rovira M, Romero A, Guardiola F, Vichi S, et al. Prospective exploration of hazelnut's unsaponifiable fraction for geographical and varietal authentication: A comparative study of advanced fingerprinting and untargeted profiling techniques. Food Chem. 2024 May;441:138294.

22. Amaral JS. Target and Non-Target Approaches for Food Authenticity and Traceability. Foods. 2021 Jan 16;10(1):172.

23. 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 Apr;86:537–43.

24. Exploratory Data Analysis - ScienceDirect [Internet]. [cited 2024 Jun 7]. Available from: https://www.sciencedirect.com/science/article/abs/pii/B978044459528700003X

25. Callao MP, Ruisánchez I. An overview of multivariate qualitative methods for food fraud detection. Food Control. 2018 Apr;86:283–93.

26. Cubero-Leon E, Peñalver R, Maquet A. Review on metabolomics for food authentication. Food Res Int. 2014 Jun;60:95–107.

27. Damascelli A, Palmisano F. Sesquiterpene Fingerprinting by Headspace SPME–GC–MS: Preliminary Study for a Simple and Powerful Analytical Tool for Traceability of Olive Oils. Food Anal Methods. 2013 Jun;6(3):900–5.

28. Bortolomeazzi R, Berno P, Pizzale L, Conte LS. Sesquiterpene, Alkene, and Alkane Hydrocarbons in Virgin Olive Oils of Different Varieties and Geographical Origins. J Agric Food Chem. 2001 Jul 1;49(7):3278–83.

29. Sesquiterpene - an overview | ScienceDirect Topics [Internet]. [cited 2024 Jun 7]. Available from: https://www.sciencedirect.com/topics/neuroscience/sesquiterpene

30. 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 Control. 2022 Sep;139:109055.

31. Vichi S, Guadayol JM, Caixach J, López-Tamames E, Buxaderas S. Monoterpene and sesquiterpene hydrocarbons of virgin olive oil by headspace solid-phase microextraction coupled to gas chromatography/mass spectrometry. J Chromatogr A. 2006 Aug 25;1125(1):117–23.

32. Kim E, Yang S, Jeon BB, Song E, Lee H. 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. 2024 Mar;15(3):566.

33. Kadri N, Khettal B, Aid Y, Kherfellah S, Sobhi W, Barragan-Montero V. Some physicochemical characteristics of pinus (*Pinus halepensis* Mill., *Pinus pinea* L., *Pinus pinaster* and *Pinus canariensis*) seeds from North Algeria, their lipid profiles and volatile contents. Food Chem. 2015 Dec 1;188:184–92.

34. Vichi S, Lazzez A, Kamoun NG, López-Tamames E, Buxaderas S. Evolution of Sesquiterpene Hydrocarbons in Virgin Olive Oil during Fruit Ripening. J Agric Food Chem. 2010 Jun 9;58(11):6972–6.

35. Sun J, Xia Y. Pretreating and normalizing metabolomics data for statistical analysis. Genes Dis. 2024 May;11(3):100979.

36. Centering, scaling, and transformations: improving the biological information content of metabolomics data | BMC Genomics | Full Text [Internet]. [cited 2024 Jun 14]. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-7-142

37. R. Chrétien J. Chemometrics in chromatography. TrAC Trends Anal Chem. 1987 Nov 1;6(10):275–8.